Epistemic Diversity as Distribution of Paper Dissimilarities

Jochen Gläser[1], Michael Heinz[2] and Frank Havemann[2]

[1] *Jochen.Glaser@ztg.tu-berlin.de*
Center for Technology and Society, TU Berlin, Hardenbergstr. 16-18, Berlin, 10623 (Germany)

[2] *{Michael.Heinz, Frank.Havemann}@ibi.hu-berlin.de*
Berlin School of Library and Information Science, Humboldt-University of Berlin, Dorotheenstraße 26, 10099 Berlin (Germany)

Abstract
We continue our quest for measures of epistemic diversity that fit the inherent properties of thematic structures in science. Starting from theoretical considerations, we argue that currently available measures of diversity are not applicable to the epistemic diversity of published scientific knowledge because topics are fluid and overlap. Consequently, we abandon attempts to assign papers to topics and instead explore opportunities to measure diversity based on paper dissimilarities. Considerations of the exploitation of information and signal-to-noise ratios in networks of papers let us dismiss an earlier attempt to base a dissimilarity measure on the resistance distance between papers in the network of papers and their cited sources. In this paper, we explore a dissimilarity measure based on papers' 'views' on the whole network, with the 'view' of a paper consisting of all other papers in the network ranked according to the length of their shortest paths to the paper. We present test results on the diversity of topics, journals and country outputs for information science (2008) as well as on the diversity of country outputs in astronomy and astrophysics (2010).

Conference Topics
Methods and techniques; Indicators

Introduction
The epistemic diversity of research – the diversity of empirical objects, methods, problems, or approaches to solving them – has become a matter of concern for science policy. Attempts by science policy to increase the selectivity of research funding and the growth in strength and homogeneity of incentives for universities have led to concerns about an undue reduction of the diversity of research. Several specific warnings refer to the UK's research assessment exercise (Gläser et al., 2002, Molas-Gallart & Salter, 2002, Rafols et al., 2012). A similar concern has been raised in Germany, where profile-building activities at all universities may make the small subjects disappear (HRK, 2007). Laudel & Weyer (2014) observed in the Netherlands that universities' uniform responses to political signals contributed to the disappearance of one field and the stagnation of another.

Discussions about dangers to the epistemic diversity of research have in common that they lack both theoretical backing and empirical evidence. Epistemic diversity is an ill-understood topic in science studies. It is rarely clear what the concept is intended to refer to, how epistemic diversity might affect research, and how it can be operationalized. Theoretical reasoning drawing on analogies to biodiversity assumes diversity is good for science (e.g. Rafols et al., 2012). However, arguments lack empirical grounding, and no specific arguments about necessary and sufficient levels of diversity or about dangers of too much diversity can be made. The empirical studies of interdisciplinarity (e.g. Bordons et al., 2004; Rafols & Meyer, 2007; Rafols et al., 2012) were forced to use rather coarse indicators such as the journal classification of the web of science, and could not theoretically justify the measures they applied.

The aim of our paper is to present a systematic approach to the measurement of diversity that derives possible bibliometric measures of diversity from properties of the system whose diversity is to be measured, namely scientific knowledge.

We start from a theoretical definition of 'topics' in science and demonstrate that the properties of topics do not match the built-in assumptions of current indicators. While this does not necessarily invalidate the indicators, the assumptions underlying the measurement of diversity in science must be made explicit, and their applicability be argued. We suggest two additional strategies that may alleviate the problems resulting from the mismatch between properties of topics and prerequisites of indicators. The first strategy abandons the explicit identification of topics and measures the diversity of paper networks rather than scientific knowledge. We propose a measure of paper similarity that takes some of the properties of scientific knowledge into account, and demonstrate our approach by applying the measure to two data sets. The second strategy, which is outlined in this paper but not applied, uses the same similarity measure for determining the disparity of topics, thereby enabling the application of existing diversity measures.

Theoretical background

In the most general sense, 'diversity' is the property of a system, namely its heterogeneity, which is caused by the disparity of its elements. Among the many aspects of a science system to which the concept diversity can be applied, we are interested in the diversity of published scientific knowledge. Other aspects of a field's diversity such as the diversity of informal knowledge, instrumentation, empirical objects, or scientific training of researchers, will not be considered here. The *epistemic diversity of a research field* is thus defined here as the diversity of published knowledge claims about scientific problems, solutions, empirical objects, approaches and methods, which are communicated by the field's researchers in publications.

The definition of epistemic diversity as a property of published knowledge suggests using bibliometric methods for its measurement. These methods must support the reconstruction of knowledge structures from publications in a way that is both valid (i.e. returns knowledge structures researchers work with) and supports the measurement of diversity. Fulfilling both requirements is made difficult by inherent properties of knowledge structures in science. In the following, we first discuss the built-in assumptions of current measures of diversity. We then argue that properties of scientific knowledge and of its representation in publications do not meet these assumptions, and discuss opportunities to reconstruct knowledge structures from publications and to measure the epistemic diversity of research.

Built-in assumptions of current approaches to the measurement of diversity

Diversity has been an important topic of biological and environmental research for some time. These fields are mainly concerned with the impact of diversity on the stability and development of biotopes and species. Two approaches to the measurement of biodiversity can be distinguished:

a) The diversity of biotopes[1] composed of several species is measured with a three-level hierarchical approach. Biotopes are considered as consisting of species, which in turn consist of individuals. Three factors contribute to the diversity of such a system, namely

- variety (the number of species in the biotope),
- disparity (the extent to which the species differ from each other), and
- evenness (the distribution of individuals across the different species).

Depending on the research question, these factors can be assessed separately (e.g. if only the number of species is measured) or be combined in synthetic measures such as Shannon's Entropy (combining variety and evenness) or the Rao-Index (combining all three measures).

[1] A biotope is a physical environment (habitat) with a distinctive assemblage of conspicuous species (Olenin & Ducrotoy, 2006: 22).

This approach to diversity is applied in fields outside the biosciences as well (see Rafols et al., 2012, Stirling, 2007). It requires that

- the system whose diversity is to be measured can be analytically decomposed in three levels (system, categories, and elements),
- the contribution of differences between individuals of the same species to the biotope's diversity can be neglected,
- the categories can be constructed as disjunct by assigning each element to exactly one category or by fractional assignments of elements to categories, and that
- all categories share a property that can be used to calculate disparity.

b) The diversity of species composed of individuals is measured on the basis of a two-level approach. In this approach, variety and evenness become meaningless because there is no intermediate level of categories to which elements can belong. The only remaining basis for measuring the diversity of the system is the disparity of individuals. While this approach is used less frequently, it can be considered to be more fundamental because it conceptualizes diversity as the degree to which the elements of a system (here: a species) differ from each other. This approach is applicable as long as a system can be delineated and elements share a property that can be used to calculate disparity.

Both approaches share a premise concerning the disparity of categories and elements. Categories and elements are conceptualized as stable, and their pairwise disparities as independent, i.e. not affected by other categories respectively elements. New elements entering the system (i.e. individuals of a species being born or migrating to a biotope) do not affect the disparity between existing elements or between the categories, and new categories (i.e. species migrating to a biotope) do not affect the disparity between the categories or between the elements that are already present. The same applies to the disappearance of elements or categories.

Properties of topics in scientific knowledge

If the approaches to the measurement of diversity are to be applied to scientific knowledge, the system, categories and elements must be determined. For the three-level approach, the system would be the knowledge of a field, topics in this field would serve as categories, and knowledge claims (the claim for some empirical, theoretical or methodological statement to be true) would constitute the elements of the system. For the diversity measures discussed above to be applicable, these knowledge structures would need to fulfil the built-in assumptions of the measures. We therefore begin by briefly discussing the properties of scientific knowledge in its structures.

Scientific knowledge is produced by scientific communities whose members

- observe the community's shared body of knowledge,
- interpret this knowledge in the light of their own research experience,
- identify gaps in that knowledge and design research processes for producing the knowledge that closes the observed gap, and
- offer their interpretation and the new knowledge to their community.

The interpretation of the community's knowledge and claims about new knowledge are fully or partially shared by some members of the community. We define a topic as *a focus on theoretical, methodological or empirical knowledge that is shared by a number of researchers and thereby provides these researchers with a joint frame of reference for the formulation of problems, the selection of methods or objects, the organization of empirical data, or the interpretation of data* (on the social ordering of research by knowledge see Gläser, 2006). This definition resonates with Whitley's (1974) description of research areas but abandons the assumption that topics form a hierarchy. The only demand the definition makes is that some scientific knowledge is perceived similarly by researchers and influences their decisions.

Due to this nature as shared and collective perspectives, topics have structural and dynamic properties that affect the opportunities for measurement. *Structural properties* include the following:

1) All topics are *emergent meso- or macro-structures*, i.e. they are collective-level products of autonomous interpretations and uses of knowledge by individual researchers.

2) From this follows that topics are *local* in the sense that they are primarily topics to the researchers whose decisions are influenced and who contribute to them, and only secondarily topics to those colleagues who are outside observers.

3) Given the multiple objects of knowledge that can serve as common reference for researchers, it is inevitable that topics *overlap*. Overlaps are ubiquitous because any research is likely to address several topics at once, e.g. by including theories about an object, methodologies for investigating it, and empirical information about an object. They also occur when a knowledge claim belongs to several topics at once (e.g. formulae used in bibliometrics belonging to mathematics but also expressing bibliometric relationships).

4) Knowledge has a *fractal structure* (e.g. van Raan, 2000), and topics can have any size between small (emerging topics that in the beginning may concern just two or three researchers) and very large thematic structures such as bibliometrics. The 'size' of a topic can be defined in various ways – as scope (range of phenomena covered), level of abstraction (which is again linked to the range of phenomena covered), or number of research processes or researchers influenced by it. In all these dimensions there is a continuum from very small to very large topics.

5) The degree to which knowledge influences researchers' actions, and the strength of links between new findings and existing knowledge that are constructed by researchers, also vary between 'very weak' and 'very strong'. As a result, the '*distinctiveness'* of topics varies. Some topics are unambiguously seen as being different from other knowledge by most researchers of a field and are thus well separated from surrounding knowledge, while others are much less pronounced.

These structural properties of topics let them form an inconsistent poly-hierarchy for which not even meaningful levels can be determined. This also implies that no field or collection of papers has exactly one definite thematic structure. Different perspectives can be applied to fields and collections of papers and will return different topical structures. Topics may overlap in their boundaries or pervasively. They vary considerably in their size and 'distinctness', i.e. the extent to which they actually constitute a shared concern of researchers.

Dynamic properties of topics are shaped by their role in the knowledge production process. As coinciding perspectives of researchers, topics are perpetually changing. Researchers constantly revise their perspectives on the existing knowledge and thus the relationships of their perspectives to those of their colleagues. They also utilize and contribute to more than one topic (e.g. theoretical, methodological and empirical ones). Hence, their production of new knowledge may instigate at least one and in many cases all of the following changes:

* Enrichment: Since new knowledge is added to the system, the community's knowledge on a topic is likely to grow.

* Restructuring: The new knowledge is linked to existing knowledge and thereby links existing knowledge, i.e. the density of connections in the system of knowledge increases.

* Reduction: The new knowledge may devalue existing knowledge by proving it to be wrong or may reduce it by subordinating it to a generalisation.

Through these processes, the size of topics, their distinctness and relations between them are constantly changed. New topics may emerge at any time, and existing topics may disappear or radically change.

Representation of knowledge in publications and reconstructions of topics

Since bibliometric methods reconstruct knowledge structures from publications, the representation of knowledge in publications provides the opportunities and constraints for a bibliometric measurement of diversity, which we now discuss in more detail. In the sociology of science, knowledge claims are treated as the basic unit through which new knowledge is communicated (e.g. Cozzens, 1985, Pinch, 1985). Knowledge claims are claims that some new knowledge produced by the author is true; a publication usually contains several such claims.

For the new knowledge claims to be added to the community's body of knowledge, they must be used by other community members in their subsequent knowledge production. This requires the new knowledge to be available to all potential users, which is achieved by publication. With each publication, researchers construct

- an account of the state of the current knowledge on a topic,
- the claim that there is a specific gap in that knowledge,
- the claim to have developed an approach whose application can close that gap,
- the new knowledge produced with this approach, which is claimed to close the gap, and
- in many cases conclusions concerning implications of the new knowledge including the necessity of further specific research (Gläser, 2006: 125-126, Swales, 1986: 45).

These claims embed the new knowledge that is offered to the community in the existing knowledge. However, they do so selectively and *ad hoc*. The claims in a publication are organised in a way that maximises the chances of the new knowledge's further use by emphasizing originality, relevance, validity and reliability of the new knowledge. Links to the existing knowledge are crafted to further this impression.

The new knowledge claims shape subsequent knowledge production processes if they inform the formulation of problems, choice of methods or interpretation of results by readers of the publication. If they do so, the researchers using them are likely to indicate the link of the new knowledge they offer to these knowledge claims, thereby treating them as part of the community's knowledge. This 'elementary process' of adding knowledge causes the dynamic properties described in the previous section. If a new knowledge claim is added, the community's knowledge becomes enriched, and its structure changes because the claim creates new links between, reinforces or remove existing links. New knowledge claims may also invalidate existing claims or subsume them to more general statements if they are used by other community members in this way.

Consequences for the measurement of diversity

The properties of knowledge claims and topics affect the opportunities to reconstruct topics from publications with bibliometric methods, i.e. by using properties of publications such as authors, journals, references, or terms. To begin with, no method for the bibliometric reconstruction of individual knowledge claims has been proposed so far. Knowledge claims are represented in series of sentences and clauses that are distributed across a publication. Reconstructing them would be a task for linguistics but is still impossible for that field, too.

Bibliometric methods are better suited for the reconstruction of topics because the latter are larger and span many publications. However, from the properties of topics described earlier follows that none of the bibliometrically usable properties of a paper can be assumed to be thematically homogeneous in the sense of representing only one topic. Since research processes are influenced by and address more than one topic, topics overlap in research processes, publications (and thus references), terms, journals, and authors. Furthermore, researchers apply their individual perspectives on the scientific knowledge when constructing

and linking topics, which is why links to topics may occur unpredictably in a variety of scientific fields. Consequently, any finite sub-set of papers is unlikely to include all publications addressing a specific topic, which means that any hierarchy of topics is also only partially covered by the paper set.

Owing to the mismatch between properties of publications that can be used for the reconstruction of topics and the representation of topics in publications, bibliometric methods inevitably reduce the complexity of the underlying knowledge structures. This is not a problem in itself because all models reduce complexity. The question is not how the reduction of complexity can be avoided but whether a specific reduction of complexity is appropriate to the purpose. Answers to this question should be part of a bibliometric methodology that links specific purposes of topic reconstruction to specific strategies that are applied. The absence of such a methodology is one of the major obstacles for bibliometrics.

When we apply these methodological considerations to the measurement of epistemic diversity, we can distinguish three strategies for solving the problems posed by properties of scientific topics. The first strategy, which has been applied in all attempts to measure epistemic diversity so far, constructs distinct topics to which papers are assigned. The three-level approach is then used for the measurement of diversity.

A second possible strategy would be to construct overlapping topics to which papers belong partially. In order to apply three level-diversity measures, the topics would have to be made disjunct by fractionalising the papers. The disparity of topics would need to be measured based on the difference in paper membership. While this strategy still has some problems in the case of pervasive overlaps of topics, it would create a more precise representation of topics and still enable the application of three-level diversity measures.

The third strategy, which we apply in the remainder of the paper, circumvents the problem of topic reconstruction by applying the two-level approach. Since knowledge claims cannot be reconstructed from publications, the strategy measures paper diversity as a proxy for knowledge diversity. This strategy requires a similarity measure for published papers, which should reflect the properties of thematic structures in science discussed above.

Methods and Data

Network-based measures of paper similarity

Diversity measures for the two-level approach aggregate the pairwise similarities of all elements. Among the many ways in which the similarity of two papers in a network can be determined, we need to find those that strike a balance between utilizing as much information as possible and avoiding the inclusion of irrelevant information that contaminates the measure.

Bibliographic coupling is well-established, and is commonly considered as one of the best bibliometric measures of paper similarity (Ahlgren & Jarneving, 2008: 274-275). The strength of bibliographic coupling between two papers can be used directly as a measure of their similarity. However, bibliographic coupling is not a useful measure for the similarity of papers that are not coupled. All these papers must be considered equally dissimilar, which they are certainly not. Thus, bibliographic coupling is unsatisfactory as a measure of paper similarity in networks.

An alternative to using bibliographic coupling is the utilization of all connections in a network, e.g. by measuring similarity as resistance distance in networks of papers and their cited sources or in bibliographic coupling networks. In this approach, indirect links between papers are taken into account, i.e. information about the whole network is utilized for the calculation of all pairwise paper similarities (see Gläser et al., 2013 for an example). However, this approach inevitably uses information about detours through a network – i.e.

about connections that exist and can technically be made but are not meaningful in terms of paper similarities. In other words, the measure is distorted by paths that do not reflect thematic similarity. Furthermore, our own experiments showed the measure to favour papers with a high degree. Finally, using all paths in a paper network for the measurement of its diversity makes the measure particularly sensitive to changes in the network structure. If measures of paper similarity are based on the resistance distance, each paper that is added to the network changes the resistance distance and thus the similarities of all papers in the network. This is an extremely unrealistic assumption about the impact of new publications on the epistemic diversity of a field.

Between the use of only information about direct coupling and the use of information about all possible connections between papers lie measures such as length of the shortest path between two nodes. This measure makes little sense in networks of papers and their cited sources because each reference two papers have in common creates a path of the length two between them. For networks in which links reflect the relative strength of bibliographic coupling, the length of shortest paths captures more information.

By determining the length of the shortest path between two papers in a network, other connections are taken into account indirectly by dismissing them as longer paths. Still, the environment of a paper is largely neglected by such a measure. However, the length of shortest path can be used to construct an indirect measure of paper similarity that takes the environment of papers into account. We can construct the 'view' of a paper on its environment by ranking all other papers in the network according to their distance to that paper. The 'view' describes how dissimilar other papers in the network are in terms of their shortest paths. The similarity between two papers can be defined as the similarity of the two papers' 'views' on the network, which is measured by calculating the rank correlation of the two lists.

Thus, we measure the similarity of two papers by:

- determining the shortest paths between all pairs of papers in a bibliographically coupled network (weighted with the arccosine of Salton's Cosine),
- creating a 'view' of each paper by ranking all other papers according to increasing lengths of their shortest paths,
- calculating the similarity of two papers as the rank correlation (Spearman) between the two lists, and
- transforming the rank correlation in a similarity measure.

This measure, which can be interpreted as the similarity of the 'views' of the two papers on their scientific environment, avoids the influence of degrees. It is similar to the use of "preferences" in an "affinity" system by Balcan et al. (2012) in their construction of overlapping endogenous communities.

Data

To test our measure, we used two data sets. The first data set is the main component of publications (articles, letters and proceedings papers) in six information science journals, which consists of 492 papers (see Havemann et al., 2012 for a description of this data set). The second data set is the main component of 14,770 publications (articles, letters, and proceedings papers) published 2010 in 53 astronomy and astrophysics journals (see Havemann et al., 2015 for a description of this data set). For each data set, we constructed and analysed the bibliographic coupling network.

Methods

For each data set, we calculated pairwise paper similarities as transformed Spearman's rank correlation of the papers' 'views' on the network. The 'view' of a paper p_i on the network is the vector of shortest paths between p_i and the papers p_1 to p_n of the network. Thus, the dissimilarity of two papers – their distance – is calculated as

$$dist\ \left(view(p_i i), view\left(p_j j\right)\right) = 1 - \frac{r_{sp}\left(view(p_i), view\left(p_j\right)\right) + 1}{2}$$

Where r_{sp} is the Spearman's rank correlation coefficient of the two views.

We tested this similarity measure on our information science data set by using it for a Ward clustering and comparing the best matching Ward clusters to three topics we had previously identified by inspecting titles and keywords of the articles.

We then calculated the distributions of paper similarities for country subsets and journal subsets of papers in both data sets, and used the median of the distributions as single-number value of the subset's diversity.

Our diversity measure also enables the construction of 'collective views', i.e. of 'views' of paper sets on each other. We exploited this opportunity in a third step and constructed similarities between countries and journals in information science.

Results

Information science

Our Ward clustering with the similarity measure led to results that compare well to previous experiments with other algorithms (Table 1).

Table 1. Salton's Cosine of precision and recall of pre-defined information science topics by five algorithms.[2]

Table	*MONC*	*HLC*	*FHC*	*RDDC*	*SPBC*
h-index	0.71	0.93	0.59	0.92	0.95
Bibliometrics	0.79	0.82	0.83	0.87	0.86
Webometrics	0.58	0.60	0.46	0.65	0.53

The three best performing algorithms – HLC, RDDC and SPBC – perform best for the h-index, good for bibliometrics including the h-index, and worst for Webometrics. These differences may be linked to the topics' internal diversity (Figure 1). Internal diversity is lowest for the h-index (all papers are very similar) and highest for webometrics (a high proportion of webometrics papers is not very similar). The differences in internal diversity may explain the differential success of algorithms in recapturing the topics.

[2] MONC= Merging overlapping natural communities, HLC=Hierarchical link clustering, FHC=Fuzzification of hard clusters (see Havemann et al., 2012). RDDC= Ward clustering with a similarity measure using the rank correlation of 'views' based on the resistance distance in direct citation networks (Gläser et al., 2013). SPPC= Ward clustering with a similarity measure using the rank correlation of 'views' based on the length of shortest paths in bibliographic coupling networks (algorithm presented in this paper). Among the three topics, bibliometrics also includes the h-index papers.

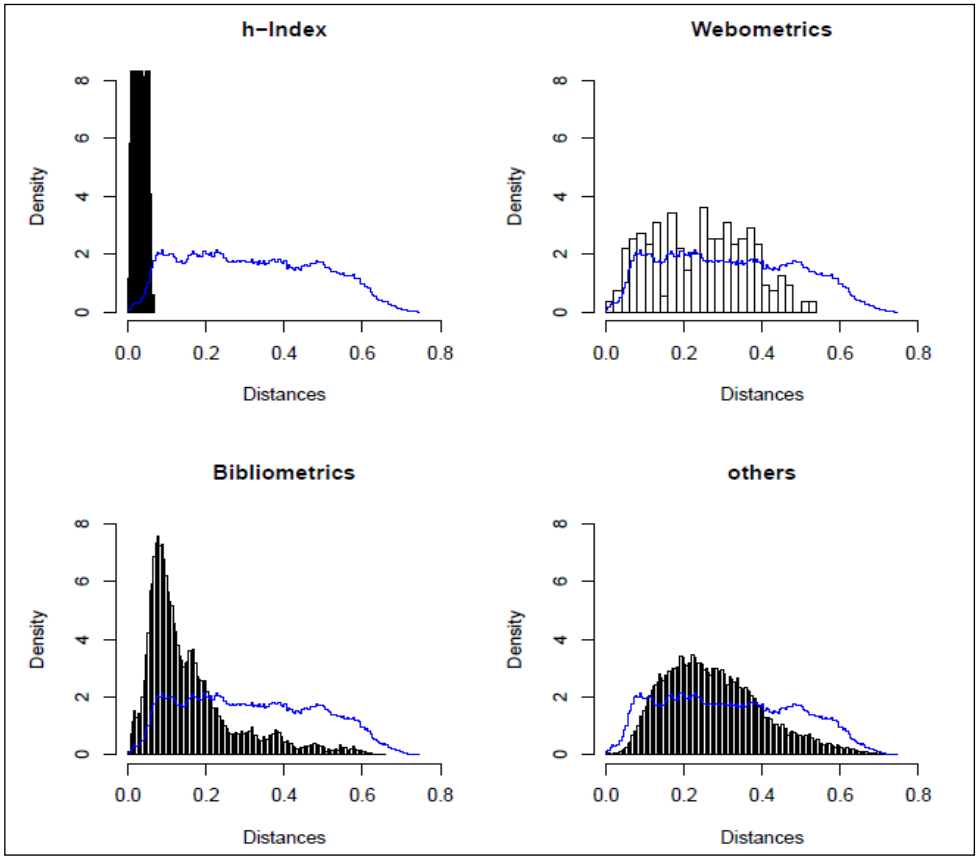


Figure 1. Internal diversity of three topics in the information science network (the blue lines represent the distribution for the whole network, the areas always equal one).

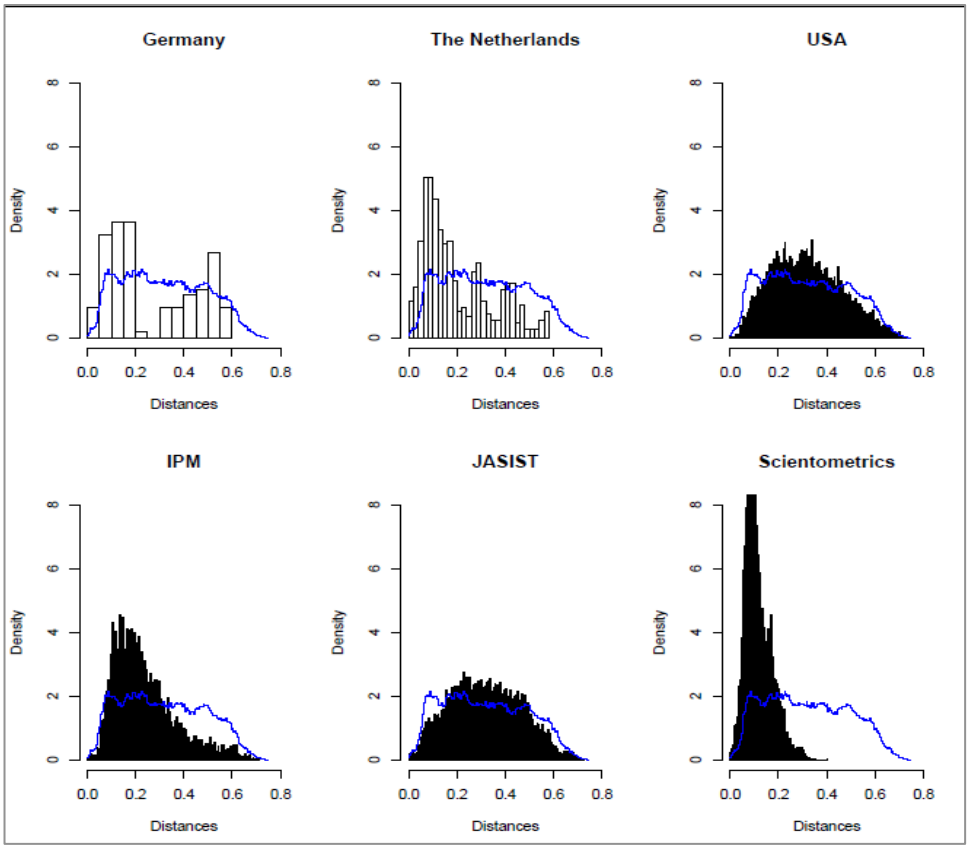


Figure 2. Diversity of information science publications from three countries and three journals.

Figure 2 shows the diversity of information science publications in three journals and of three countries. According to these distributions of distances,

a) Dutch information science publications are less diverse than the few German publications and the publications from the USA; and

b) *Scientometrics* was the least diverse (most focused) journal, followed by *JASIST* and *Information Processing and Management*.

Astronomy and astrophysics

The astronomy and astrophysics publication network is less diverse than the information science network. Taking the median as a single-number measure of diversity, the information science network (median = 0.32) is much more diverse than the astronomy and astrophysics network (median = 0.27). Owing to space limitations, we can provide only one comparison. Figure 3 compares the distribution of paper similarities for Chilean and US-American publications. Astronomy and astrophysics publications from Chile appear to be much less diverse (much more concentrated on one or few topics) than those from the USA.

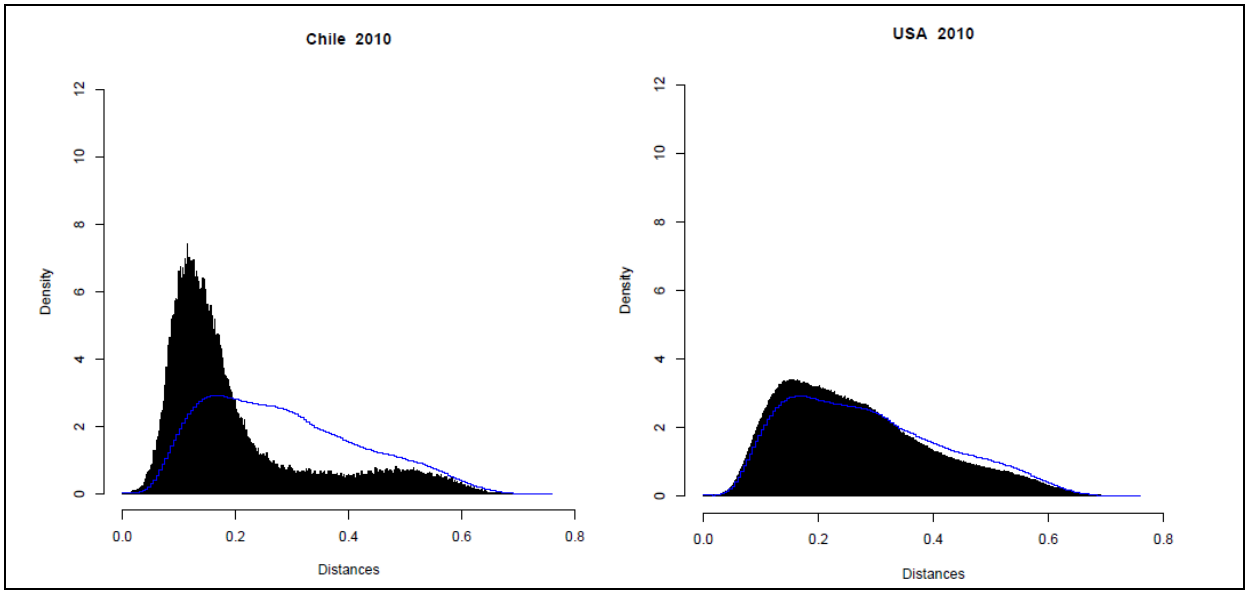


Figure 3. Diversity of astronomy and astrophysics publications from Chile and the USA (the blue lines represent the distribution for the whole network).

Discussion

A small but noxious problem for the application of our diversity measure is the occurrence of direct citations between publications from the same year. Direct citations can be considered a strong indicator of thematic similarity. However, it is not known how strong an indicator a direct citation is, and how it should be treated in comparison with bibliographic coupling of two publications. Our current solution is to add the citing and cited publication to each other's reference lists, i.e. integrating direct citation into bibliographic coupling. This solution is, however, as arbitrary as any other solution would be.

A more consequential limitation stems from our use of networks of papers as models of published knowledge. Adding a node with at least two links to a network indirectly changes connections between all nodes. This is not true for added knowledge, which can induce changes in similarities that remain local in that they affect only the knowledge to which it links directly. Although the length of the shortest path between two papers is not as sensitive to changes in networks as the measure we tried before (resistance distance), it remains to be seen whether time series of diversity constructed with our distance measure can be

interpreted. Since the literature in most fields keeps growing, time series of diversity have to cope with ever-growing paper networks.

Finally, a third limitation is inherent to our measure. Measuring the diversity of any set of papers with the approach suggested in this paper requires the set of papers to be embedded in a connected subgraph. If a research organisation has publications in many unrelated fields (as most universities do, providing an aggregate measure of the diversity of this organisations published output would be impossible. However, such an aggregate measure is likely to be meaningless in any case.

Conclusion

While further tests are of course necessary, the diversity measure proposed in this article appears to enable comparisons of paper sets from topics, journals, specialised organisations, or countries. The measure appears to use enough information to provide meaningful results without being sensitive to the noise created by network connections that have no bearing on the similarity of two papers. It is also compatible with sociological findings that ground the publication process in an author's personal experience and perspective. The 'view' of a paper on the network can easily be interpreted as the scientific perspective of its author.

Our discussion of diversity measures and their applicability to the epistemic diversity of published knowledge suggests two lines of further work. First, the problem of time series must be solved, i.e. the diversity of a field must be measured for networks of different sizes. This requires assessing the sensitivity of our diversity measure for changes in networks that are unrelated to epistemic diversity.

Second, a solution must be found for the measurement of diversity with a three-level approach. This is both theoretically and practically important because changes in the diversity of research are caused by the selective growth and shrinking of topics. Understanding the role of epistemic diversity for research requires causally attributing changes in the epistemic diversity to such processes of growth and decline, which in turn requires linking publications to topics. The obvious solution is making topics disjoint by fractionally assigning papers to overlapping topics. However, this does not solve all problems posed by thematic structures in science. Consider the following simple example: A paper on the h-index is simultaneously a paper in bibliometrics because the topic h-index is fully included in bibliometrics. How would one assign such a paper to the two topics?

Developing three-level measures for the diversity of overlapping topics might mean abandoning all established measures, and might prove a very challenging task.

References

Ahlgren, P. & Jarneving, B. (2008). Bibliographic coupling, common abstract stems and clustering: A comparison of two document–document similarity approaches in the context of science mapping, *Scientometrics, 76*, 273-290.

Balcan, M.-F., C. Borgs, M. Braverman, J. Chayes & S.-H. Teng. (2012). Finding endogenously formed communities. arXiv:1201.4899v2.

Bordons, M., F. Morillo & I. Gómez (2004). Analysis of cross-disciplinary research through bibliometric tools. In: H. F. Moed , W. Glänzel & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research*, Dordrecht: Kluwer, 437-456.

Cozzens, S. E. (1985). Comparing the Sciences: Citation Context Analysis of Papers from Neuropharmacology and the Sociology of Science. *Social Studies of Science 15*(1), 127-153.

Gläser, J. (2006). *Wissenschaftliche Produktionsgemeinschaften: Die soziale Ordnung der Forschung*. Frankfurt am Main: Campus.

Gläser, J. G. Laudel, S. Hinze & L. Butler (2002). Impact of Evaluation-based Funding on the Production of Scientific Knowledge: What to Worry About, and how to Find Out. Report to the BMBF. Retrieved January 20, 2015, from http://www.laudel.info/wp-content/uploads/2013/pdf/research%20papers/02expertiseglaelauhinbut.pdf.

Gläser, J., F. Havemann, M. Heinz & A. Struck. (2013). Measuring the diversity of research using paper similarities. In: S. Hinze & A.Lottmann (Eds.), *Translational twists and turns: Science as a socio-economic endeavor. Proceedings of the 18th International Conference on Science and Technology Indicators, Berlin, Germany, September 4 — 6, 2013*, Berlin, 130-139.

Havemann, F., J. Gläser, M. Heinz & A. Struck (2012). Identifying overlapping and hierarchical thematic structures in networks of scholarly papers: A comparison of three approaches. *PLoS ONE, 7*(3), e33255.

Havemann, F., Gläser, J. & Heinz, M. (2015). A link-based memetic algorithm for reconstructing overlapping topics from networks of papers and their cited sources. *15th International Conference on Scientometrics and Informetrics*, Istanbul, 29 June -3 July 2015.

Hochschulrektorenkonferenz (HRK) (2007). Die Zukunft der kleinen Fächer: Potenziale – Herausforderungen – Perspektiven, Bonn: Hochschulrektorenkonferenz. http://ipts.jrc.ec.europa.eu/home/report/english/articles/vol66/ITP1E666.html.

Laudel, G. & E. Weyer. (2014). Where have all the Scientists Gone? Building Research Profiles at Dutch Universities and its Consequences for Research. In: Richard Whitley and Jochen Gläser (Eds.), *Organizational Transformation And Scientific Change: The Impact Of Institutional Restructuring On Universities And Intellectual Innovation*, Bingley, UK: Emerald Group Publishing Limited, 111-140.

Molas-Gallart, J. & A. Salter (2002). Diversity and Excellence: Considerations on Research Policy. IPTS Report.

Olenin, S. & J.-P. Ducrotoy (2006). The concept of biotope in marine ecology and coastal management. *Marine Pollution Bulletin 53*, 20–29.

Pinch, T. (1985). Towards an Analysis of Scientific Observation: The Externality and Evidential Significance of Observational Reports in Physics. *Social Studies of Science 15,* 3-36.

Rafols, I., L. Leydesdorff, A. O'Hare, P. Nightingale & A. Stirling (2012). How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management. *Research Policy, 41*, 1262-1282.

Rafols, I. & M. Meyer (2007). How cross-disciplinary is bionanotechnology? Explorations in the specialty of molecular motors. *Scientometrics, 70*(3), 633-650.

Stirling, A. (2007). A general framework for analyzing diversity in science, technology and society. *Journal of The Royal Society Interface, 4*, 707–719.

Swales, J. (1986). Citation analysis and discourse analysis. *Applied Linguistics, 7*, 39-56.

Van Raan, A. F. J. (2000). On growth, ageing, and fractal differentiation of science. *Scientometrics,* 47, 347-362.

Whitley, R. (1974). Cognitive and social institutionalization of scientific specialties and research areas. In: R. Whitley (ed), *Social Processes of Scientific Development* (pp.69–95). London.

Using Bibliometrics-aided Retrieval to Delineate the Field of Cardiovascular Research

Diane Gal[1], Karin Sipido[1] and Wolfgang Glänzel[2]

{diane.gal, karin.sipido}@med.kuleuven.be , wolfgang.glanzel@kuleuven.be
[1]Department of Cardiovascular Sciences, KULeuven, 3000 Leuven (Belgium)
[2]ECOOM and Dept. MSI, KU Leuven, 3000 Leuven (Belgium) & Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

Abstract
A hybrid search strategy, using lexical and citation based methods, is presented in this paper as a robust method to delineate the broad field of cardiovascular research. Overall, this study aims to provide scientifically reliable and accurate data driven evidence about cardiovascular research by establishing a dataset of published research in this field. A workflow is presented that outlines the methods carried out to establish a core dataset based on a core set of journals, to identify and use search terms to detect a broader dataset, and then to apply measures of similarities between the citations of these two datasets to ensure relevance of the final dataset. The final core set of journals established comprises of 120 unique journals covered in Thomson Reuters *Web of Science Core Collection* (WoS) database including a total of 320,647 documents from 1991 to 2013. The search terms utilised include 107 cardio-specific terms that initially identify 1.8 million unique documents when searching the title, abstract and keywords. Upon application of the citation-based similarity measures the final combined dataset consists of 845,071 publications. Overall, establishing a relevant dataset of cardiovascular research means placing a greater emphasis on having a precise dataset, reducing recall in the process.

Conference Topic
Methods and techniques

Introduction
Experts in the cardiovascular field are concerned that there is a decline in quality and innovation in cardiovascular research and that fragmentation of this broad field is leading to loss of cross-pollination and missed opportunities for translation of research from bench to bedside. In this context we have launched a project to examine cardiovascular research output over a 23 year period to provide rigorous and reliable scientific information about cardiovascular research activities. The findings of this project are expected to serve as a complement to expert opinion and previously published studies (Huffman et al., 2013; Jones, Cambrosio, & Mogoutov, 2011; Sipido et al., 2009; van Eck, Waltman, van Raan, Klautz, & Peul, 2013; Yu, Shao, He, & Duan, 2013), to provide scientifically reliable and accurate data driven evidence about cardiovascular research.
The objectives of the project are to:

- Characterise the size, growth, topics and visibility of research outputs over 23 years;
- Analyse the geographical distribution of research outputs and its evolution;
- Visualise and analyse research collaboration; and
- Identify emerging topics in cardiovascular research.

To gain a comprehensive view of research in this field a broad scope and definition has been applied to include papers published in scientific journals from basic, clinical and epidemiological studies related to the cardiovascular system, including the heart, the blood vessels and/or the pericardium. The main source of data is the *Web of Science Core Collection*. The purpose of this paper is to describe the methods utilised, and the roadmap set, to establish a dataset of published research undertaken in the cardiovascular field.

Methods

Hybrid search strategies for subject delineation, previously described and published (Bolaños-Pizarro, Thijs, & Glänzel, 2010; Glänzel, Janssens, & Thijs, 2009; Zitt & Bassecoulard, 2006), have been adapted to establish a dataset of cardiovascular research. This includes (1) establishing a core dataset based on a core set of journals and core search terms, (2) identifying a broader dataset of publications through the use of search terms, and then, (3) applying measures of similarities by citations between the documents in these datasets to select a final dataset with acceptable precision and recall. A workflow/roadmap was developed to outline the main steps taken to establish the dataset, as can be seen in Figure 1.

Core Journal Dataset

All data have been retrieved from Thomson Reuters Web of Science Core Collection. The core set of journals was selected through expert review of the scope/aim of all 183 journals included in the 'Cardiac & Cardiovascular Systems' and the 'Peripheral Vascular Disease' Web of Science Categories. The scope/aim for each journal was obtained through online web-based searches. Using an online survey tool, two experts reviewed the title and scope/aim of each journal to assess the relevance of the journal and indicate whether they had experience with each journal (e.g. reading, editing, reviewing, submitting a document for publication). Journals that were assessed by at least 1 expert as being a core cardiovascular journal – defined as a journal publishing greater than 90% of its articles, reviews, letters and notes on the cardiovascular domain – were included in the core journal dataset. Disagreements between the experts were reviewed by the project team. Journals were excluded from the core dataset only when the expert excluding the journal was the only one that had previous experience with the journal. The final dataset was obtained by identifying all articles, letters, notes and reviews published journals that are covered in the 1991–2013 volumes of the WoS database.

Search Terms Datasets

A number of sources were reviewed to identify relevant cardiovascular-specific search terms, including:

- Medical Subject Headings (MeSH)
- International Classification of Diseases (ICD)-10
- Cochrane Hypertension/Heart/Peripheral Vascular Disease Groups/Systematic Reviews
- Cardioscape project taxonomy (European Society of Cardiology, 2014)
- Recent published research (Bolaños-Pizarro et al., 2010; Huffman et al., 2013; Jones et al., 2011; van Eck et al., 2013)

Subsequently, a group of eight topic experts representing a mix of clinical scientists, basic scientists and epidemiologists were invited to review the combined list of 105 search terms to assess their relevance in identifying as broad a range of cardiovascular research publications as possible. All search terms were included where at least half of the reviewers agreed that they were relevant search terms to include in the search strategy.

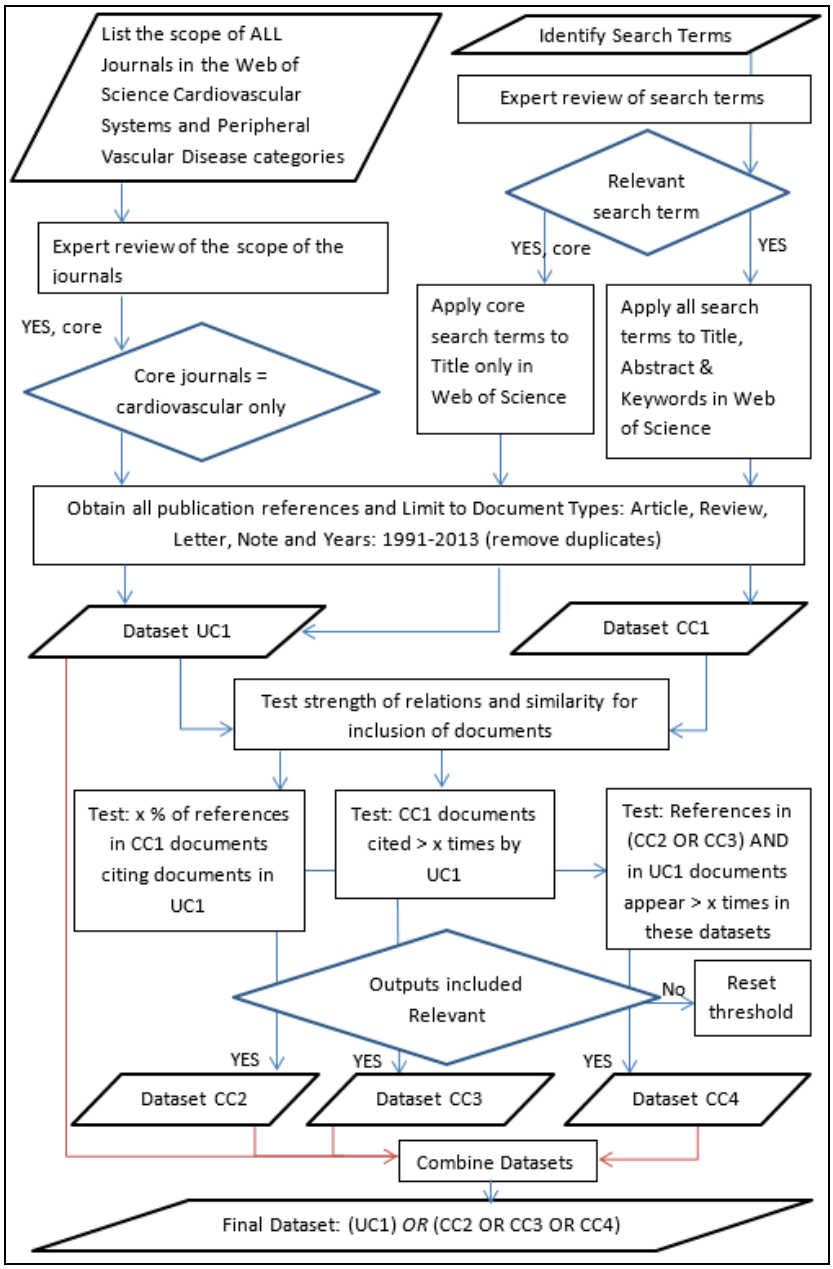


Figure 1. Workflow of field delineation of Cardiovascular Research

In addition, experts were asked to suggest any potentially missing search terms. New search terms suggested and disagreements were reviewed by the project team. The broad search terms dataset was obtained by applying the full search strategy to the complete Web of Science database, to identify all articles, letters, notes and reviews published between 1991 and 2013. To add to the core journal dataset, highly cardiovascular specific or core search terms were selected that when searched in the title would identify core cardiovascular publications.

Similarity Measures and Thresholds

For the extension of the core dataset, i.e., the seed of relevant literature, we followed an algorithm using a logical combination of *unconditional* and *conditional* criteria (Glänzel, 2014). In the present project we have linked literature retrieved based on conditional criteria (the broad search terms set) to the set of surely relevant documents (the core journals and core search terms set), using citation-based similarities. In particular, three measures of similarity

between the core dataset and the broad search terms dataset were utilised: a) the share of references of broad search terms documents that cite the core documents, b) the number of references of the core documents that cite the broad search terms documents and c) the number of shared references between the core dataset and the restricted search terms dataset. The thresholds for each measure were set following iterative testing, whereby a low threshold was first applied and a random sample of the titles and abstracts of 500 documents was reviewed for relevance to the cardiovascular field. The threshold was altered until the sample contained a high precision and the level of noise (peripheral and irrelevant documents) was reduced to an acceptable level, defined as a 5% level of noise. To confirm the relevance of the documents identified, the random samples considered to have acceptable thresholds were reviewed by one topic expert.

Findings

Core Dataset

After expert review, 120 journals were included as core journals. The two expert reviewers agreed on the exclusion of 61 journals and disagreed on the inclusion of 39 journals (21% of all 183 journals), of these only two journals were excluded as the expert who had experience with the journal was the one that excluded it. For the remaining 37 journals, they were included since both experts had previous experience for three journals and neither expert had experience for 34 journals. The final core journal documents therefore consist of 320,647 articles, letters, notes and reviews from 1991 to 2013. Thirteen of the search terms, identified below, were considered to be highly cardiovascular specific. The core search terms when searched only in the title, added 141,676 documents to the core journal documents, resulting in a core dataset of 462,323 documents. Review of this dataset confirmed that it provides a precise sample of cardiovascular-specific documents for this study.

Broad Search Terms Dataset

After expert review by 6 topic experts and the project team, 107 search terms were included in the final search strategy. Of the original 105 terms reviewed, three search terms were removed since more than half of the experts suggesting to remove them. A total of 22 unique terms were also suggested by three of the topic experts. The project team assessed and included four of these new terms. Then one additional term was added to the search strategy to include this term with and without its common prefix. The final broad search terms dataset consists of 1,656,278 unique articles, letters, notes and reviews from 1991 to 2013 where the search terms could be identified in the abstract, keywords or title. All documents in the core dataset were removed from this broad search term dataset.

A comparison of all documents obtained by searching the abstract, keywords and title is presented in Figure 2.

As a validation of the search strategy and selection of core journals, when the search strategy was applied to the 120 core journals, 95% of all core journal dataset documents were identified by the search terms.

Similarity Measures and Thresholds

An initial test was undertaken to limit the search terms dataset by removing all documents that had no links with the core journal documents. A total of 228,000 documents had no links meaning they did not cite the core journal set, they were not cited by the core journal set *and* they did not have any common references with the core journal set. This reduced the search terms set to less than 1.6 million documents, however upon review of random samples it was

clear that stronger measures of similarity would be needed to further restrict the search terms dataset to include the most relevant documents in the final dataset.

Iterative testing and review of random samples led to the selection of a combined dataset where at least 12% of the references in the broad search documents cited documents in the core dataset or where the broad search documents where cited greater than 4 times by the core documents. For this chosen dataset, no more than 10% of the random samples were considered not relevant or peripheral to the cardiovascular field. Documents from the third measure of similarity using bibliographic coupling was not included in the final dataset since it was not possible to achieve less than a 10% noise level through iterative testing and review of random samples. The final restricted broad search terms dataset consists of 382,748 unique articles, letters, notes and reviews from 1991 to 2013.

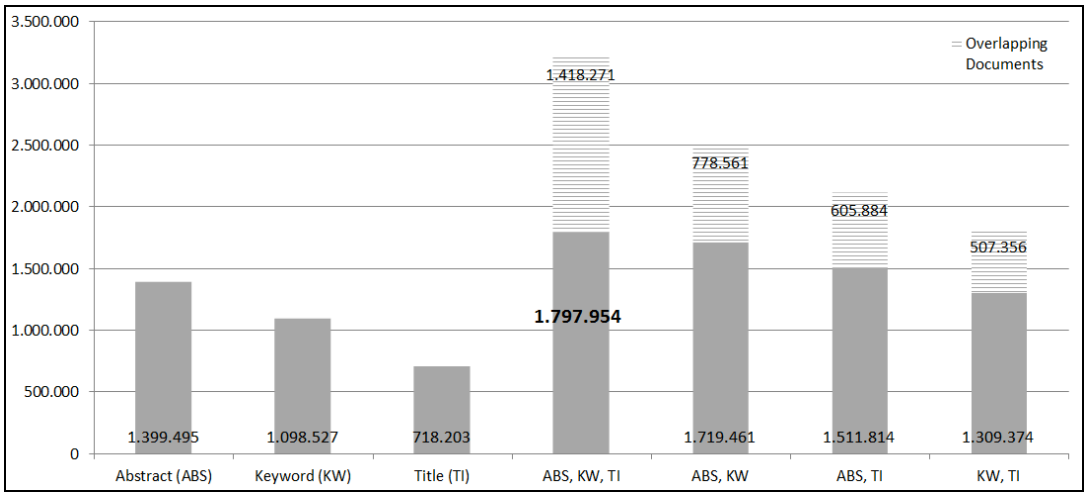


Figure 2. Number of documents identified when searching 107 search terms in Abstracts, Keywords and Titles [Data sourced from Thomson Reuters Web of Science Core Collection].

Final Combined Dataset

Combined, the core and restricted datasets create a final dataset of 845,071 unique documents from the cardiovascular field. Overall, the combined dataset has a 4.5% noise level (estimated).

Discussion

Only one previously published bibliometric study of cardiovascular research used a hybrid search strategy to establish its dataset (Bolaños-Pizarro et al., 2010). However, due to the broad scope of this study, which aims to include all types of research – from basic to clinical research, a broader list of cardio-specific search terms was created. Attention was also placed on ensuring that the search terms selected could identify cardiovascular research over the long time period of the study, as well as, enable the identification of new and emerging fields in cardiovascular research. The 107 search terms greatly increases the recall of documents, though this also means that a greater amount of noise was present in the broad search terms dataset. Hence, the importance of utilising measures of similarity between the two datasets to restrict the broad search terms dataset to include only the most relevant documents. This was done through testing various thresholds of citation-based similarities, as the final step of this robust method to delineate complex fields of research. Including both directions of citation-based similarities (ie. documents from core journals dataset citing documents in search terms dataset and vice versa) also ensures that the distribution of documents sampled is representative over time. The initial threshold of 5% noise was re-evaluated through testing

and due to the broad nature of the cardiovascular field a higher level of noise (10%) was considered acceptable as this includes peripheral research that has a component linked to cardiovascular research. The broad search terms dataset has been reduced to less than a quarter of initial documents identified to ensure the final dataset is as precise as possible and can be considered a representative sample of cardiovascular research over the 23 year period.

Conclusions

Bibliometrics-aided retrieval is a robust method to delineate the field of cardiovascular research. Through using this method, a representative dataset of cardiovascular research was established irrespective of changes in the field, such as vocabulary used, over the time-frame of this study. Overall, establishing a relevant dataset of cardiovascular research means placing a greater emphasis on having a precise dataset, reducing recall in the process.

Acknowledgments

Thank you to Bart Thijs for his input into the study methods.

References

Bolaños-Pizarro, M., Thijs, B., & Glänzel, W. (2010). Cardiovascular research in Spain. A comparative scientometric study. *Scientometrics*, *85*(2), 509–526. doi:10.1007/s11192-009-0155-2

European Society of Cardiology. (2014). *CardioScape: A survey of the European cardiovascular research landscape* (p. 52). Retrieved June 2, 2015 from: http://www.cardioscape.eu/static_file/CardioScape/PNO%20report/CardioScape_Summary%20Report_30092014.pdf

Glänzel, W. (2014). Bibliometrics-aided retrieval – where information retrieval meets scientometrics. *Scientometrics*. doi:10.1007/s11192-014-1480-7

Glänzel, W., Janssens, F., & Thijs, B. (2009). A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics. *Scientometrics*, *79*(1), 109–129. doi:10.1007/s11192-009-0407-1

Huffman, M. D., Baldridge, A., Bloomfield, G. S., Colantonio, L. D., Prabhakaran, P., Ajay, V.S., Lewison, G., & Prabhakaran, D. (2013). Global cardiovascular research output, citations, and collaborations: A time-trend, bibliometric analysis (1999–2008). *PLoS ONE*, *8*(12), e83440. doi:10.1371/journal.pone.0083440

Jones, D. S., Cambrosio, A., & Mogoutov, A. (2011). Detection and characterization of translational research in cancer and cardiovascular medicine. *Journal of Translational Medicine*, *9*(1), 57. doi:10.1186/1479-5876-9-57

Sipido, K. R., Tedgui, A., Kristensen, S. D., Pasterkamp, G., Schunkert, H., Wehling, M., Dambrauskaite, V. (2009). Identifying needs and opportunities for advancing translational research in cardiovascular disease. *Cardiovascular Research*, *83*(3), 425–435. doi:10.1093/cvr/cvp165

Van Eck, N. J., Waltman, L., van Raan, A. F. J., Klautz, R. J. M., & Peul, W. C. (2013). Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research. *PLOS ONE*, *8*(4). doi:10.1371/journal.pone.0062395

Yu, Q., Shao, H., He, P., & Duan, Z. (2013). World scientific collaboration in coronary heart disease research. *International Journal of Cardiology*, *167*(3), 631–639. doi:10.1016/j.ijcard.2012.09.134

Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing & Management*, *42*(6), 1513–1531. doi:10.1016/j.ipm.2006.03.016

Locating an Astronomy and Astrophysics Publication Set in a Map of the Full Scopus Database

Kevin W. Boyack[1]

[1] *kboyack@mapofscience.com*
SciTech Strategies, Inc., 8421 Manuel Cia Pl NE, Albuquerque, NM 87122 (USA)

Abstract
A dataset containing 111,616 documents in astronomy and astrophysics has been created and is being partitioned by several research groups using different algorithms. In this paper, rather than partitioning the dataset directly, we locate the data in a previously created model in which the full Scopus database was partitioned. Given that the other research groups are partitioning the data directly, use of this method will allow comparisons between using local and global data for community detection. In other words, use of this method will allow us to start to answer the question of how much the rest of a large database affects the partitioning of a journal-based set of documents. We find that the astronomy document set, while spread across hundreds of partitions in the Scopus map, is located in only a few regions of the map. Thus, the use of a global map to partition astronomy documents is likely to give very similar results to partitioning using local approaches because of the insularity of the field of astronomy. However, we do not expect local and global data to give as similar results for other fields, because most other fields are less insular than astronomy.

Conference Topic
Methods and techniques

Introduction
Partitioning of a dataset into groups of similar objects – alternatively known as clustering, community detection or topic detection – is a current research topic in a number of fields, including scientometrics and network science. A number of different algorithms are used to partition scientific literature into topics or clusters. While many of these are based on the property of modularity (cf., Blondel, Guillaume, Lambiotte, & Lefebvre, 2008; Newman & Girvan, 2004; Waltman & van Eck, 2013), others are based on graph layout and pruning (Martin, Brown, Klavans, & Boyack, 2011) or on complex network flows (Rosvall & Bergstrom, 2008). Despite the obvious differences between these algorithms, they are all based on a common principle – that of dividing a literature set into partitions where the within-partition signals are much stronger or denser than the between-partition signals.

It is well known that different topic detection algorithms give somewhat different results for the same data set. What is not known is the specifics of why particular algorithms give particular results, or exactly what operations of a particular algorithm lead it to give different results than those obtained by another algorithm. In general, we know very little about what types of features result from different algorithms, and how these affect the output structures. This can make it difficult to interpret the partitions and maps that are produced by an algorithm. Are the partitions produced by an algorithm representative of actual structures in science, are they merely artifacts resulting from the algorithm and its parameters, or are they something in between? This is a difficult question to which, we suspect, the answer is far beyond the scope of even a large study. Nevertheless, we are hopeful that a comparison of partitioning methods and their results using a single dataset might lead to some general understanding of the types of features that result from different methods and algorithms. This type of understanding has the potential to enable both researchers and decision makers to more clearly understand and interpret the results of a particular partitioning.

To this end, a number of researchers (see papers from this special session) have come together to explore this question. As a first step, each research group has created a partitioning of a

single dataset using their own algorithms. The work-in-progress papers in this session describe the partitioning algorithms and results from each group. The multiple results will be combined and compared in a next phase of the project to determine similarities and differences in the features resulting from the different methods and algorithms. Beyond that, we collectively hope to learn more about both common and unique structural features that result from the different algorithms.

This paper details the method used by SciTech Strategies to partition an "astronomy and astrophysics" literature dataset. It differs from the other methods in one significant aspect – the other groups have all created local solutions (partitioning the dataset directly), while we have created a global model (partitioning the entire Scopus database) and have located the astronomy dataset within those partitions (Klavans & Boyack, 2011). Use of this method enables us to start to answer the question of how much the rest of the database affects the partitioning process.

Global Model

Our global model of science consists of 48,533,301 documents from Scopus. Of these, 24,615,844 documents are indexed source documents from Scopus 1996-2012, while the remaining 23,917,457 are non-source documents that were each cited at least twice by the set of source documents. The method used to generate the document set and citing-cited pairs list is very similar to that used for the recent "non-source" map of Boyack and Klavans (2014).

The model was created by taking the over 582 million citing-cited pairs within this set of 48.5 million documents, calculating similarity values between pairs of documents based on direct citation, and then partitioning the documents using the new CWTS smart local moving algorithm (Waltman & van Eck, 2013). The citing-cited pairs were provided by SciTech Strategies (STS) to Ludo Waltman at CWTS, who ran the similarity calculation and partitioning steps. The CWTS smart local moving algorithm was used to create a four-level hierarchical solution, with resolution values chosen to result in a solution with roughly 100k, 10k, 1000, and 100 clusters. Details of the partitioning are given in Table 1.

Table 1. Multi-level partitioning of the Scopus database using the CWTS smart local moving algorithm.

Level	*Partitions Desired*	*Resolution*	*Partition Min Size*	*# Partitions*	*Partitions > Min Size*	*# Pubs*	*% Pubs Lost*
1	100000	3e-5	50	114679	91726	48399235	0.28%
2	10000	3e-6	500	13157	10059	47323189	2.49%
3	1000	3e-7	5000	1048	849	46929303	3.30%
4	100	5e-8	50000	122	114	46705047	3.77%

Visual maps of the partition solutions at level 1 and level 2 were created using the following process. At each level, 1) pairwise similarity between partitions was calculated from the titles and abstracts of the documents in each partition using the BM25 textual similarity measure, 2) each resulting similarity list was filtered to retain the top-n (5-15) similarities per partition, and 3) layout of the partitions on the x,y plane was done using the DrL algorithm. These steps are ones we commonly use to create science maps, and are described in more detail in Boyack & Klavans (2014). In each case, only those partitions that were of the minimum size desired (91,726 for level 1, and 10,059 for level 2) were included in the map. Field counts for each cluster in each map were calculated using UCSD map of science journal-to-field assignments (Börner et al., 2012), and each cluster was assigned to its dominant field and correspondingly colored in the map. The two maps are similar in that they show that the 12 large fields (see

legend at the bottom of Figure 1) occupy similar positions in both maps. The change in granularity of the partitions does not change the overall look and feel of the map.

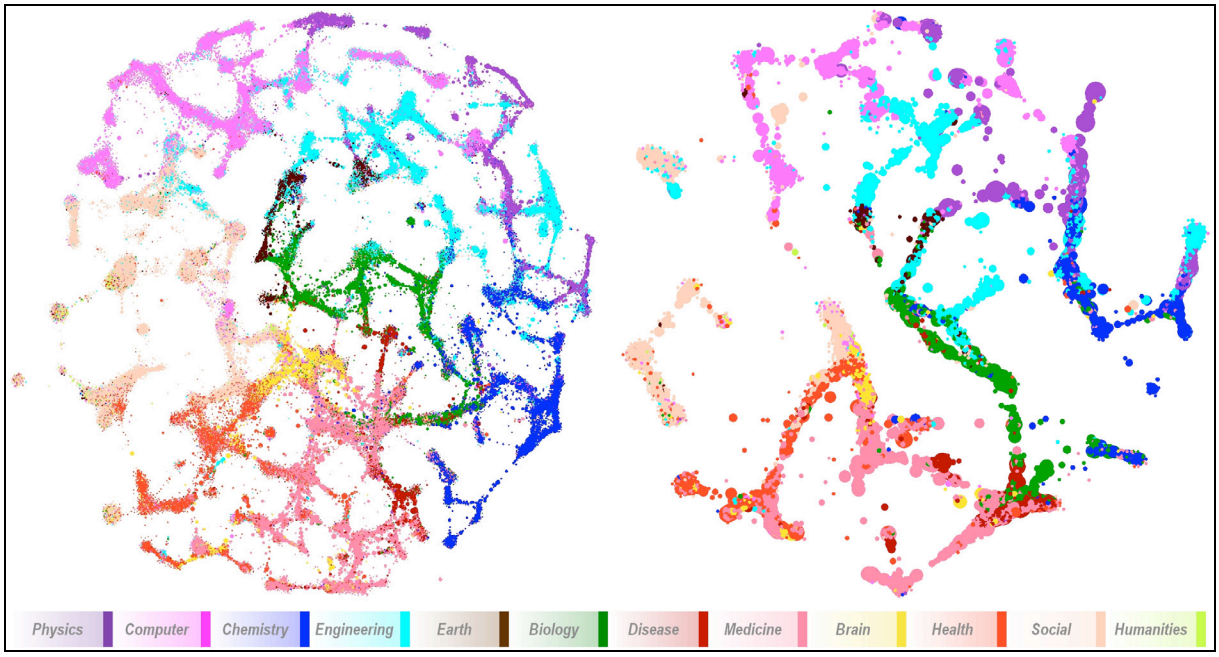


Figure 1. Visual maps of the Scopus database using level 1 (left) and level 2 (right) partitions.

Astronomy Dataset

The astronomy dataset used by each research group consists of 111,616 document records with accompanying data from the Web of Science. This dataset was created by researchers at Humboldt University for use by project participants, and is comprised of documents published from 2003-2010 in a set of 59 astronomy and astrophysical journals, limited to articles, letters, and proceedings papers. Over half of the documents were from four journals, as shown in Table 2.

Table 2. Dominant journals in the astronomy and astrophysics dataset.

Journal	Count
Astrophysical Journal	19582
Physical Review D	19061
Astronomy & Astrophysics	14668
Monthly Notices of the Royal Astronomical Society	11599

In order to use the Scopus-based global model and map, Scopus identifiers for the WoS records were identified to the extent possible by matching source data (journal, title, volume, page, year). Definitive matches were obtained for 107,888 (96.66%) of the documents. Of the 3,728 documents that were not matched, roughly half were in source titles that are not covered by Scopus (such as the IAU Symposium), and thus could only be matched if they were cited non-source materials. The remaining unmatched records were in source titles that are covered by Scopus, but that we could not match. This lack of uniformity between databases is primarily due to differences in the way titles are listed (particularly for non-ASCII characters) and missing records. Despite the unmatched records, we consider a match rate of nearly 96.7% to be very good, and certainly sufficient for reasonable comparison with the partitions from other groups. Once the matching was done, documents from the astronomy dataset were located in global map at three levels (1, 2, and 3 from Table 1).

Astronomy is known to be a relatively insular discipline, with fewer links (percentage basis) to and from other disciplines than for most other disciplines. Thus, we expected the effect of including an additional 48 million documents in a cluster solution to have only a modest effect on the partitioning of the astronomy document set. We did not expect the astronomy documents to be scattered throughout the map. As expected, the astronomy documents are heavily concentrated in the global model. At level 1, 50% of the astronomy documents are in partitions where the astronomy set documents comprise at least 66.5% of the partition contents (limited to the years of study, 2003-2010). In other words, when sorting partitions by concentration of the astronomy document set within the partition, 50% of the total papers are accounted for in partitions with a concentration of over 66.5%. Using an alternative measure, when partitions are sorted by the number of papers from the astronomy document set, the number of non-set papers equals the number of set papers only when 90,000 of the 111,616 papers are accounted for, as shown in Figure 2.

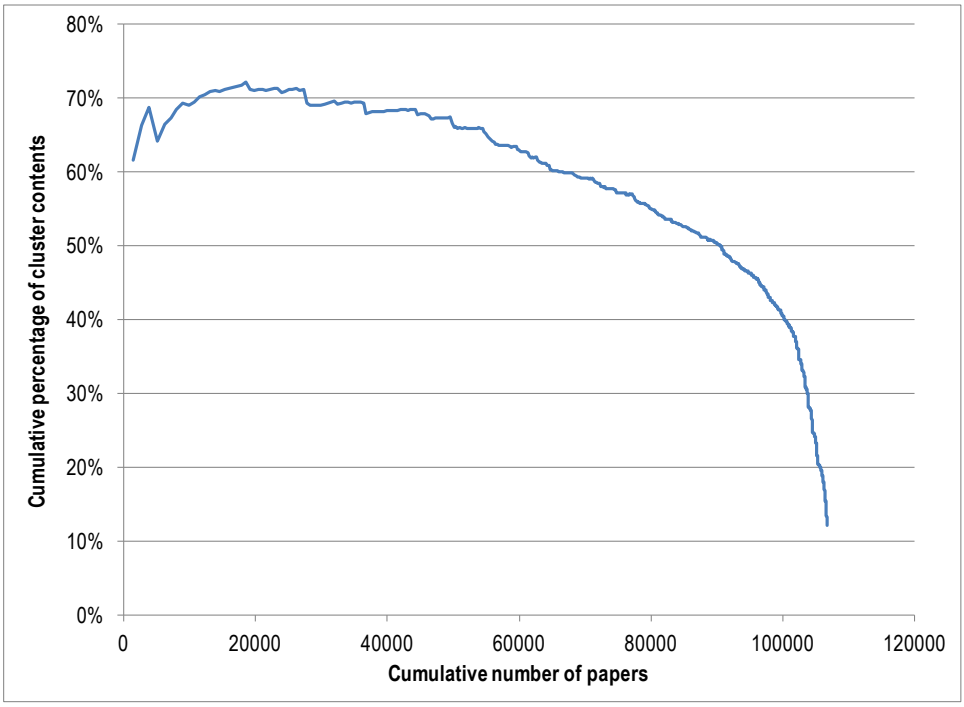


Figure 2. Distribution of the astronomy dataset across partitions in the level 1 solution.

Overlays showing the positions of the partitions with at least 50 documents from the astronomy set are shown for both the level 1 and level 2 maps in Figure 3. For level 1, this comprises 408 partitions and 90,763 documents (84.1% of the matched documents), while for level 2 it comprises 119 partitions and 101,895 documents (94.4% of the matched documents). Both maps make it clear that while the documents are parsed out into hundreds of partitions, each representing distinct topics, these topics are concentrated in only a few areas in the map.

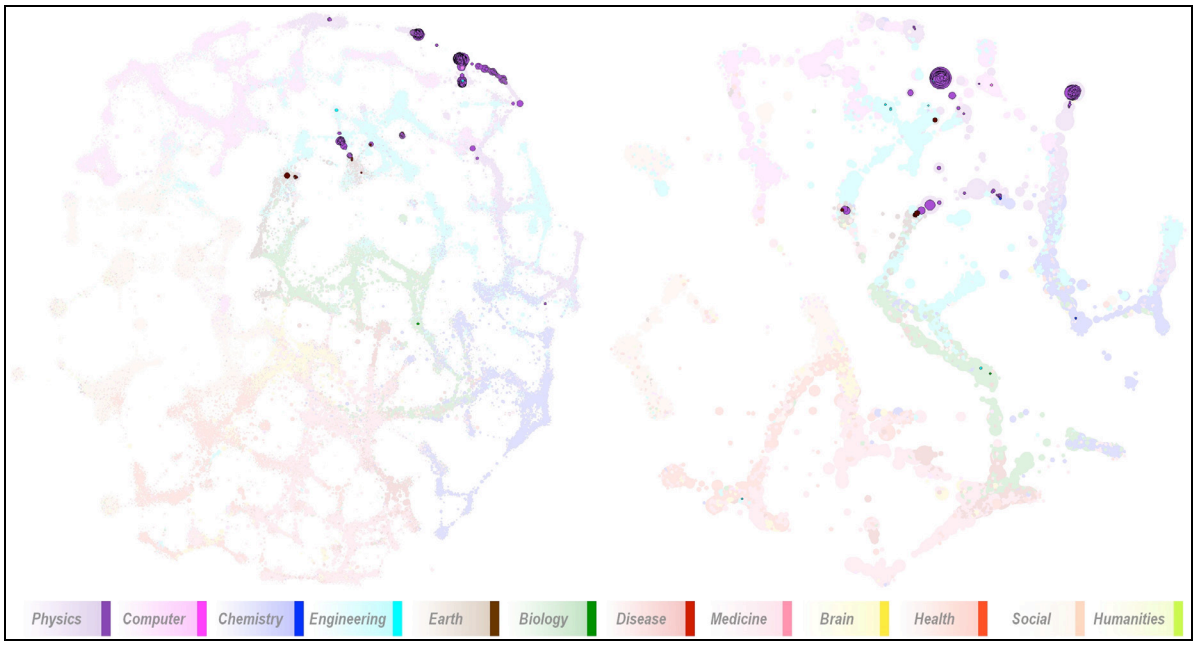


Figure 3. Overlays of the positions of the astronomy set documents on the Scopus level 1 (left) and level 2 (right) maps of Figure 1.

Discussion

Recalling that the astronomy document set was based on a set of journals, the high level of concentration of the overlays shown in Figure 3 suggests that use of journals is a very reasonable strategy for building a dataset in the field of astronomy. Astronomy journals have a very tight profile on a document-based map. By contrast, high profile journals in other fields, such as JACS, Physical Review Letters, and New England Journal of Medicine, have very broad profiles, and overlays for these journals (not shown here) spread across large regions of the map. Thus, while a dataset based on journals is useful to characterize astronomy, journals may be far less useful for characterizing other fields. Correspondingly, the use of a global map to partition astronomy documents is likely to give very similar results to partitioning using local approaches because of the insularity of the field of astronomy. We would not expect the use of a global map to partition a local document set from another field to work as well. Or, rather, we would expect the journal-based approach to fall short in other fields because it would leave out so much of the relevant contextual literature.

References

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 10*, P10008.

Boyack, K.W., & Klavans, R. (2014). Including non-source items in a large-scale map of science: What difference does it make? *Journal of Informetrics, 8*, 569-580.

Börner, K., Klavans, R., Patek, M., Zoss, A.M., Biberstine, J.R., Light, R.P., Larivière, V., & Boyack, K.W. (2012). Design and update of a classification system: The UCSD map of science. *PLoS ONE, 7*(7), e39464.

Klavans, R., & Boyack, K.W. (2011). Using global mapping to create more accurate document-level maps of research fields. *Journal of the American Society for Information Science and Technology, 62*(1), 1-18.

Martin, S., Brown, W.M., Klavans, R., & Boyack, K.W. (2011). OpenOrd: An open-source toolbox for large graph layout. *Proceedings of SPIE - The International Society for Optical Engineering, 7868*, 786806.

Newman, M.E.J. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*, 026113.

Rosvall, M. & Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the USA, 105*(4), 1118-1123.

Waltman, L. & van Eck, N.J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B, 86*, 471.

Scientific Workflows for Bibliometrics

Arzu Tugce Guler[1], Cathelijn J. F. Waaijer[2] and Magnus Palmblad[1]

[1]*a.t.guler@lumc.nl, n.m.palmblad@lumc.nl*
Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden (The Netherlands)

[2]*c.j.f.waaijer@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Faculty of Social and Behavioural Sciences, Leiden University, Leiden (The Netherlands)

Abstract
Scientific workflows organize the assembly of specialized software into an overall data flow and are particularly well suited for multi-step analyses using different types of software tools. They are also favourable in terms of reusability, as previously designed workflows could be made publicly available through the myExperiment community and then used in other workflows. We here illustrate how scientific workflows and the Taverna workbench in particular can be used in bibliometrics. We discuss the specific capabilities of Taverna that makes this software a powerful tool in this field, such as automated data import via communication with Web services, smooth data extraction from XML by XPath and various data analyses and visualizations with the statistical language R. The support of the latter allows integration of a number of recently developed R packages for bibliometric analysis. A number of simple examples illustrate the possibilities of Taverna in the field of bibliometrics and scientometrics.

Conference Topic
Methods and techniques

Introduction
Information processing permeates the scientific enterprise, generating and organizing knowledge about nature and the universe. In the modern era, computational technology enables us to automate data handling, reducing the need for human labor in information processing. Often information is processed in several discrete steps, each building on previous ones and utilizing different tools. Manual orchestration is then frequently required to connect the processing steps and enable a continuous data flow. An alternative solution would be to define interfaces for the transition between processing layers. However, these interfaces then need to be designed specifically for each pair of steps, depending on the software tools they use; which compromises reusability. Whether the data flow is automated or done by the researcher manually, the latter still has to deal with many low-level aspects of the execution process (Gil, 2008).

Scientific workflow managers connect processing units through data and control connections and simplify the assembly of specialized software tools into an overall data flow. They smoothly render stepwise analysis protocols in a computational environment designed for the purpose. Moreover, the implemented protocols are reusable. Existing workflows can be shared and used by other workflows, or they can be modified to solve different problems. Several general purpose scientific workflow managers are freely available, and a few more optimized for specific scientific fields (De Bruin, Deelder, & Palmblad, 2012). Most of these managers provide visualization tools and have a graphical user interface, e.g. KNIME (Berthold et al., 2007), Galaxy (Goecks, Nekrutenko, & Taylor, 2010) and Taverna (Oinn et al., 2004). Not surprisingly, scientific workflows are now becoming increasingly popular in data intensive fields such as astronomy and biology.

In this paper, we describe the use of scientific workflows in bibliometrics using the *Taverna Workbench*. Taverna Workbench is an open source scientific workflow manager, created by

the myGrid (Stevens, Robinson, & Goble, 2003) project, and now being used in different fields of science. Taverna provides integration of many types of components such as communication with Web Services (WSDL, SOAP, etc.), data import and extraction (XPath for XML, spreadsheet import from tabular data), and data processing with Java-like Beanshell scripts or the statistical language R (Wolstencroft et al. 2013). Beanshell services allow the user to either program a small utility from scratch and towards a specific goal, or to integrate already existing software in the workflow. The R support is a particularly powerful feature of Taverna. Although R was initially developed as a language for statistical analysis, its widespread use has seen it adopted for many tasks not originally envisioned—a fate not unlike its commercial cousin, MATLAB. One such task is text mining. The R package *tm* (Feinerer, Hornik, & Meyer, 2008) provides basic text mining functionality and is used by a rapidly growing number of higher-level packages, such as *RTextTools* (Jurka, Collingwood, Boydstun, Grossman & van Atteveldt, 2014), *topicmodels* (Grün & Hornik, 2011) and *wordcloud* (Fellows, 2013). Similarly, there are many toolkits and frameworks for text mining in Java that could also be called from within a Taverna workflow.

An Example Workflow

We designed a simple workflow, *compare_two_authors* (see below), to generate a histogram for the number of publications over time and a co-word map for the titles of the two authors' publications. The workflow takes as inputs PubMed results in XML, the names of two authors, a list of excluded words and a minimum number of occurrences.

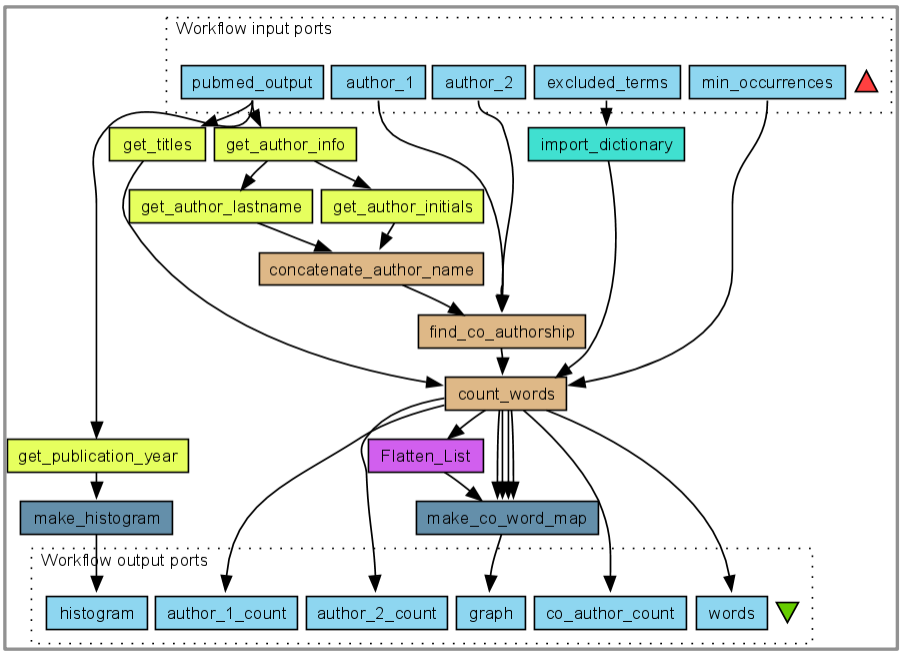


Figure 1. A workflow designed in Taverna for analyzing scientific output over time and comparing word usages of two authors.

The excluded terms are contained in a text file, so the *spreadsheet import* service in Taverna is used to extract each word in the file, line by line. The PubMed results are in XML format, and the extraction of publication years, titles and author names are done by *XPath* services. XPath is a query language for selecting elements and attributes in an XML document. The XPath service in Taverna eases this process by providing a configuration pane to render an XML file of interest as a tree and automatically generate an XPath expression as the user

selects a specific fragment from the XML (Fig. 2). The results of the query can either be passed as text or as XML to other workflow components.

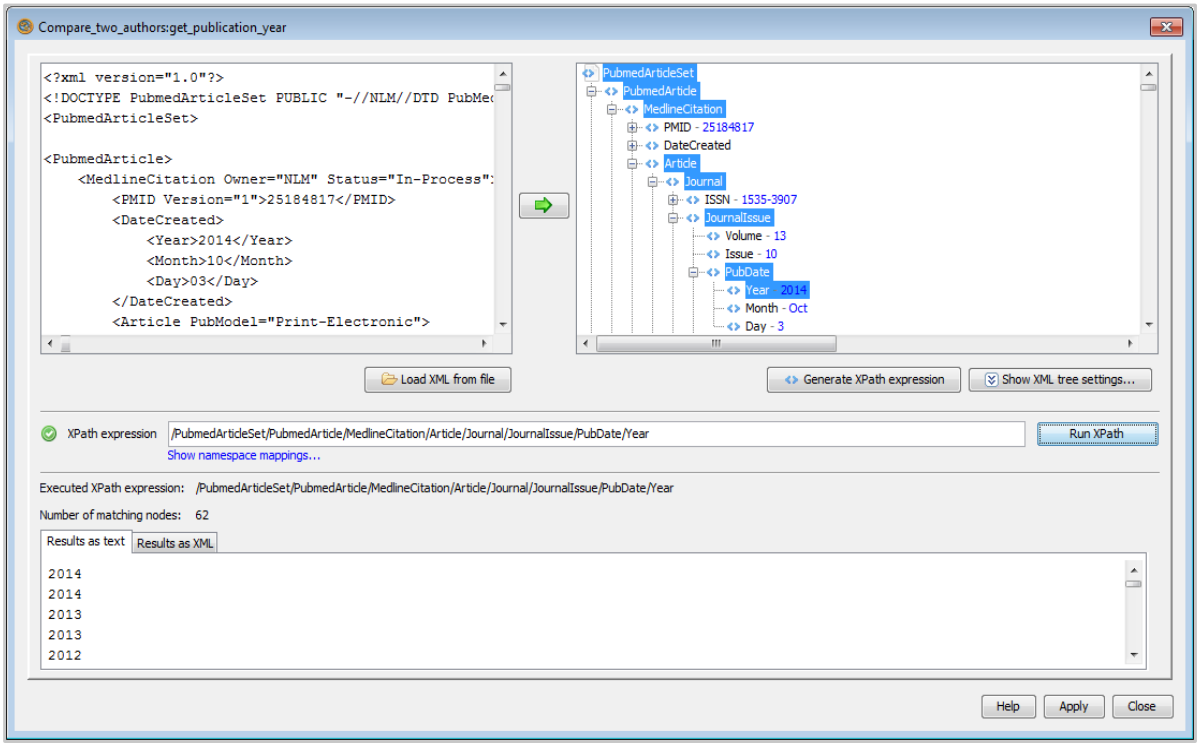


Figure 2. XPath configuration pane for extracting publication year from PubMed XML.

The data extracted by the spreadsheet import and XPath services is fed to a series of Beanshell components that find co-authorships and count co-occurrence of words in the extracted titles. Beanshell is a light-weight scripting language that interprets Java. In our workflow, the Beanshell services do simple operations on strings, such as concatenation of surnames and initials that are extracted separately using XPath (*concatenate_author_names*), matching strings to find co-authorships (*find_co_authorship*) and counting the number of words occurring in each title authored by one or both authors (*count_words*). The two authors' usage of the words, excluding *excluded_terms*, that appear at least *min_occurrences* times in total, are then used to draw a co-word map using the *igraph* (Csárdi & Nepusz, 2006) R package. It is generally up to the workflow designer what part of the workflow to code in Java (Beanshell), in R, or in third language called via the *Tool* command-line interface. More types are available for data connectors between R components (logical, numeric, integer, string, R-expression, text file and vectors of the first four types) than between Beanshell components, where everything is passed as strings. When dealing with purely numerical data, we recommend R over Beanshells within Taverna.

After all the necessary inputs are provided, the workflow is ready to be executed. In the Taverna Workbench *Results* perspective (Fig. 3), each completed process is grayed out to show the progress of the workflow run. The execution times, errors and results are also visible in this perspective.

We ran the workflow for two scientists active in our own field, mass spectrometry, Gary L. Glish and Scott A. McLuckey, whom we knew to have worked on similar topics and also co-authored a number of papers. However, the workflow will work on any two authors with publications indexed by PubMed. The co-word map in Figure 4 visualizes the co-occurrence of words in titles by the location and thickness of the connecting edge, while the relative frequency of usage by the two authors is indicated by the color (from white to gray).

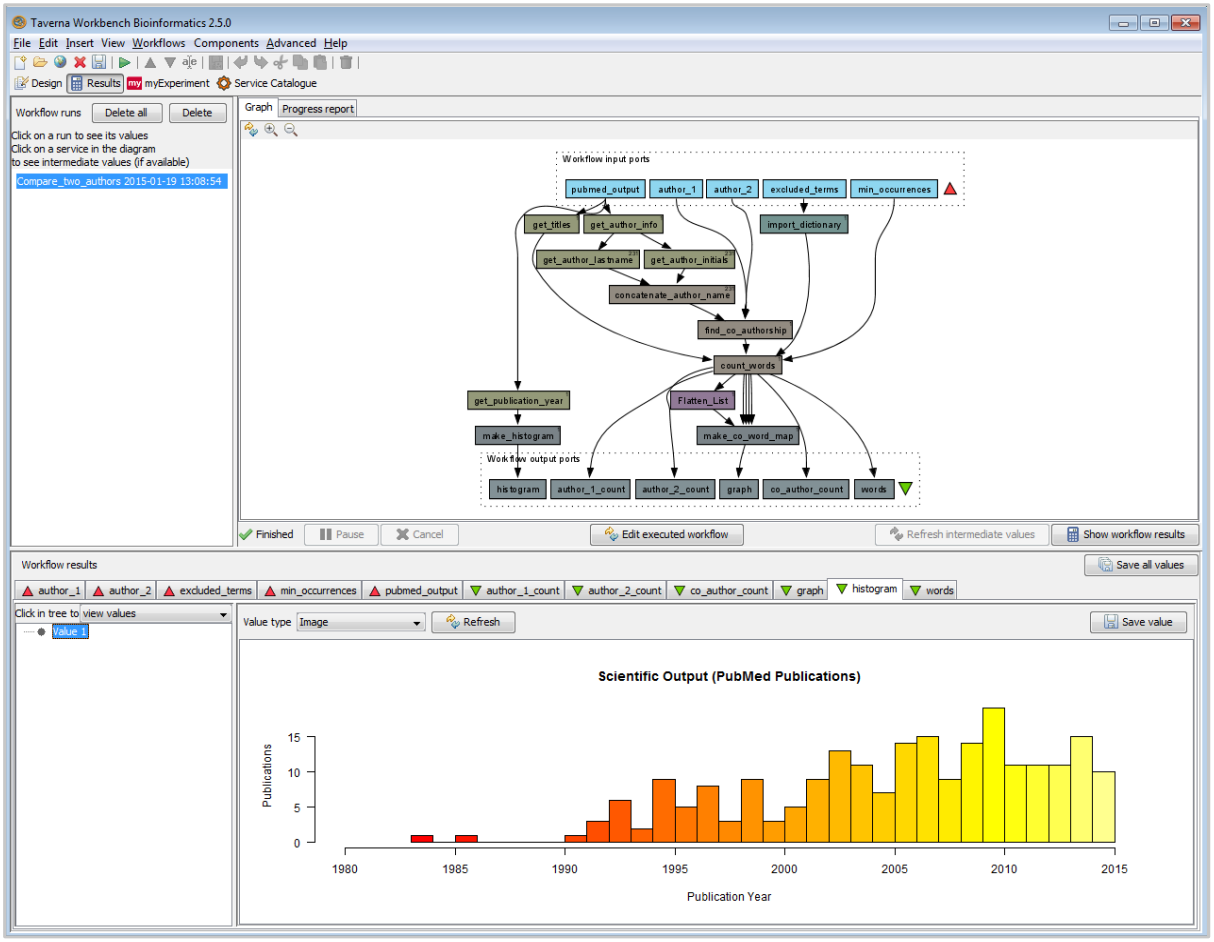


Figure 3. Workflow progress and output in the Taverna workbench Results perspective.

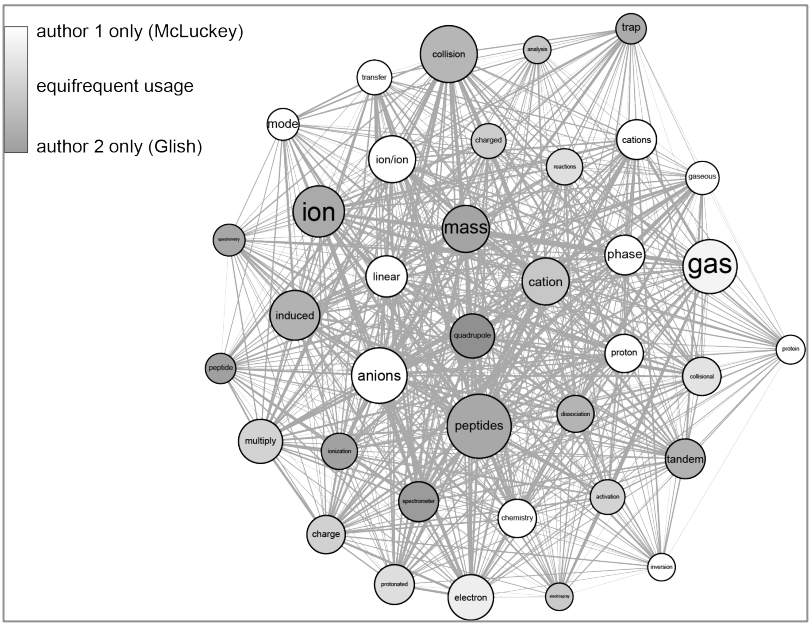


Figure 4. Co-word map output from the *compare_two_authors* workflow.

Connecting to Web Services and External Databases

Automatically generating networks directly from online data is also possible in Taverna workbench. Taverna can invoke WSDL (Web Services Description Language) style Web services given the URL of the service's WSDL document. The WSDL is an XML-based interface description language often used together with a SOAP (Simple Object Access protocol) to access the functions and parameters of a service. Many bibliographic resources are available through Web services, such as Web of Science (WoS). Some services, including the WoS, require authentication. An entire bibliometric study can be contained inside a single Taverna workflow that takes the user queries, or questions of the study, generate the Web service requests, execute these, retrieve the data and proceed with further (local) bibliometric and statistical analysis, and visualization.

A Taverna workflow that invokes WSDL services from WoS to automatically execute a query may look like in the figure below. This Taverna workflow takes as input common search parameters and a generic WoS query string, and passes these to the Web service via the WoS WSDL interface. Values that have only one possible value, such as the language (English, "en") are here hard-coded in the workflow as *Text constants*.

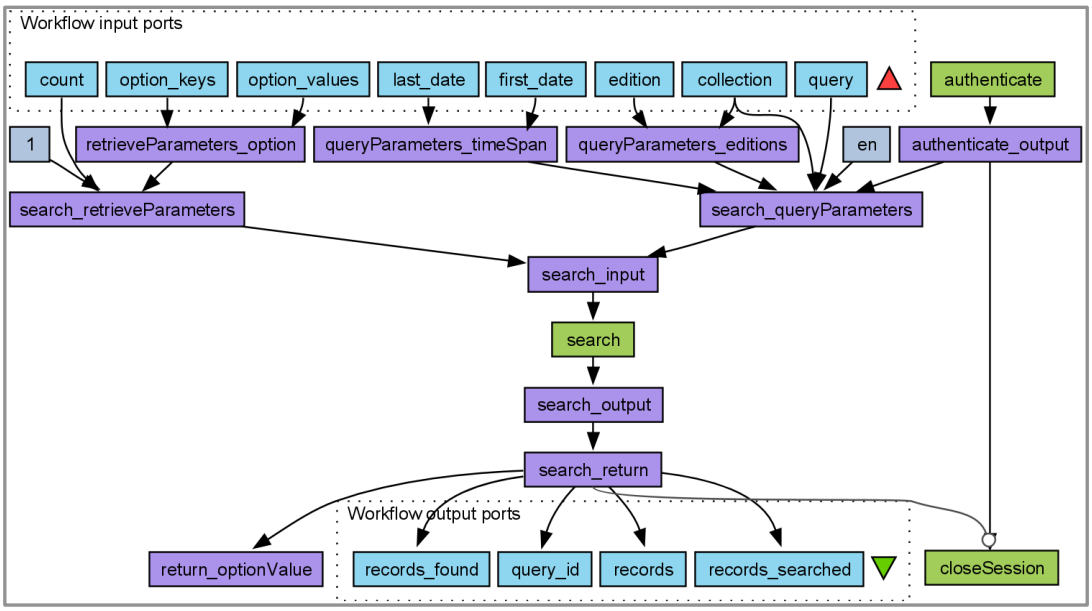


Figure 5. A simple workflow for retrieving bibliometric data using Web services.

Future Work

The use of scientific workflows in bibliometrics is still in its infancy. Modules that accomplish basic bibliometric tasks could be designed and combined in various ways for different studies, thus benefiting from modularity and reusability of scientific workflows. As mentioned above, the direct support of R inside Taverna workflows is particularly useful for bibliometrics. A number of R packages for bibliometric analysis have recently been released, ranging from simple data parsers such as the *bibtex* package (Francois, 2014) for reading BibTeX files to libraries or collections of functions for scientometrics, such as the *CITAN* package (Gagolewski, 2011). The latter package contains tools to pre-process data from several sources, including Elsevier's Scopus, and a range of methods for advanced statistical analysis. The *igraph* package itself comes with some functions specifically for bibliometric analysis, e.g. *cocitation* and *bibcoupling*. Clustering or rearranging the graph spatially so that strongly connected words appear closer together is possible with *igraph*, but may also be assisted by other packages. More crucially, the example workflow here does not yet

implement any advanced text mining functionality for homonym disambiguation or natural language processing. The *openNLP* R package provides an interface to openNLP (Hornik, 2014) and may be used to extract noun phrases and clean up the co-word maps.

Several of our Taverna workflows for bibliometrics and scientometrics, including the two workflows in Figure 1 and Figure 5, can be found in the myExperiment (Goble et al., 2010) group for Bibliometrics and Scientometrics (http://www.myexperiment.org/groups/1278.html). As always, we are grateful for any feedback on these workflows.

Acknowledgements

The authors would like to thank Dr. Yassene Mohammed for technical assistance and Thomson Reuters for granting access to the Web of Science Web services lite.

References

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., & Wiswedel, B. (2007). KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)* (pp. 319-326). Heidelberg: Springer.

Csárdi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, *1695*, 1695.

De Bruin, J. S., Deelder, A. M., & Palmblad, M. (2012). Scientific Workflow Management in Proteomics. *Molecular & Cellular Proteomics*, *11*, M111.010595–M111.010595.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal Of Statistical Software*, *25*(5), 1–54.

Francois, R. (2014). bibtex: bibtex parser. R package version 0.4.0. Retrieved from http://CRAN.R-project.org/package=bibtex.

Fellows, I. (2013). wordcloud: Word Clouds. R package version 2.4. Retrieved from http://CRAN.R-project.org/package=wordcloud.

Gagolewski, M. (2011). Bibliometric impact assessment with R and the CITAN package. *Journal of Informetrics*, *5*(4), 678–692.

Gil, Y. (2008). From Data to Knowledge to Discoveries: Scientific Workflows and Artificial Intelligence. *To Appear in Scientific Programming*, *16*(4), 1–25.

Goble, C.A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., Borkum, M., Bechhofer, S., Roos, M., Li, P., & De Roure, D. (2010). myExperiment: A repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research*, *38*(May), 677–682.

Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, *11*, R86.

Grün, B., & Hornik, K. (2011). topicmodels : An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*, 1–30.

Hornik, K. (2014). openNLP: Apache OpenNLP Tools Interface. R package version 0.2-3. Retrieved from http://CRAN.R-project.org/package=openNLP.

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E,. & van Atteveldt, W. (2014). RTextTools: Automatic Text Classification via Supervised Learning. R package version 1.4.2. Retrieved from http://CRAN.R-project.org/package=RTextTools.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A., & Li, P. (2004). Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, *20*(17), 3045–3054.

Stevens, R. D., Robinson, A. J., & Goble, C. a. (2003). myGrid: personalised bioinformatics on the information grid. *Bioinformatics (Oxford, England)*, *19 Suppl 1*(1), i302–i304.

Wolstencroft, K. et al. (2013). The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Research* 41(W1), W557-W561.

Expertise Overlap between an Expert Panel and Research Groups in Global Journal Maps

A.I.M. Jakaria Rahman[1], Raf Guns[2], Ronald Rousseau[3] and Tim C.E. Engels[4]

[1] *jakaria.rahman@uantwerpen.be*
Centre for R&D Monitoring (ECOOM), Faculty of Political and Social Sciences, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp (Belgium)

[2] *raf.guns@uantwerpen.be*
Institute for Education and Information Sciences, University of Antwerp, Venusstraat 35, B-2000 Antwerp (Belgium)

[3] *ronald.rousseau@uantwerpen.be*
ronald.rousseau@kuleuven.be
Institute for Education and Information Sciences, University of Antwerp, Venusstraat 35, B-2000 Antwerp (Belgium); and KU Leuven, Dept. of Mathematics, B-3000 Leuven (Belgium)

[4] *tim.engels@uantwerpen.be*
Centre for R&D Monitoring (ECOOM), Faculty of Political and Social Sciences, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, and
Antwerp Maritime Academy, Noordkasteel Oost 6, B-2030 Antwerp (Belgium)

Abstract
There are no available methods to measure overlap in expertise between a panel of experts and evaluated research groups in discipline-specific research evaluation. This paper explores a bibliometric approach to determining the overlap of expertise, using the 2009 and 2011 research evaluations of ten Pharmaceutical Sciences and nine Biology research groups of the University of Antwerp. We study this overlap at the journal level. Specifically, journal overlay maps are applied to visualize to what extent the research groups and panel members publish in the same journals. Pharmaceutical Sciences panel members published more diversely than the corresponding research groups, whereas the Biology research groups published more diversely than the panel. Numbers of publications in the same journals vary over a large scale. A different range of coverage was found for different research groups; there is also a significant difference between maximum and minimum coverage based on discipline. Future research will focus on similarity testing, and a comparison with other disciplines.

Conference Topic
Methods and techniques

Introduction
Expert panel review is considered the standard for determining research quality of individuals and groups (Nedeva et al., 1996; Rons, et al., 2008; Butler & McAllister, 2011; Lawrenz et al., 2012), but also, for instance, for research proposals submitted to research funding organizations. The principal objective of such evaluations is to improve the quality of scientific research. Currently, there are no available methods that can measure overlap in expertise between a panel and the units of assessment in discipline-specific research evaluation (Engels et al., 2013). Rahman et al. (2014) explored expertise overlap between panel and research groups through publishing in the same Web of Science subject categories. Since one category may comprise a wide array of different subfields and topics (Bornmann, et al., 2011), it is up for discussion how relevant it is to have panel members and research group members publishing in the same subject categories. This paper presents a journal level analysis to explore this issue. Journals cover more closely related subfields and topics (Tseng & Tsay, 2013). This paper uses overlay maps at the journal level (Leydesdorff & Rafols,

2012), with special attention to the quantification of similarity between groups and panel for two disciplines.
In 2007, the University of Antwerp (Belgium) introduced site visits by expert panels that promise communication and participation between expert and research groups. It is expected that each research group's expertise is well covered by the expertise of the panel members.
We have used the data collected in the frame of research evaluation by the University of Antwerp. This research in progress paper explores the expertise overlap between expert panel and research groups of the department of Biology and Pharmaceutical Sciences. Hence, the research questions are:

1) To what extent is there overlap between the panel's expertise and the expertise of the groups as a whole?
2) To what extent is each individual research group's expertise covered by the panel's expertise?

Data and Method

In this paper, we present an analysis of the 2009 assessment of ten research groups (2001-2008) of the Department of Pharmaceutical Sciences, and the 2011 assessment of the nine research groups (2004-2010) belonging to the Department of Biology, University of Antwerp. The citable items from the Science Citation Index Expanded of the Web of Science (WoS) published by the research groups in the reference period were considered.
Both panels were composed of five members (including the chair). All the publications of the individual panel members up to the year of assessment were taken into account. The combined publication output of the Pharmaceutical Sciences panel members is 1,029 publications. In total, these publications appeared in 300 different journals. The number of publications per panel member ranges from 124 to 353, in 39 to 93 different journals. The Biology panel members' publication output amounts to 786 publications in 217 different journals. The number of publications per panel member ranges from 76 to 262, in 36 to 76 journals. There are no co-authored publications between panel members in both cases.

Table 1: Publication profile of the Pharmaceutical Sciences and Biology research groups

Pharmaceutical Sciences research groups (2001-2008)			**Biology research groups (2004-2010)**		
Group code	**Number of Publications**	**Number of Journals**	**Group code**	**Number of Publications**	**Number of Journals**
PSRG - A	40	22	BRG - A	168	53
PSRG - B	62	32	BRG - B	58	33
PSRG - C	61	35	BRG - C	212	212
PSRG - D	32	17	BRG - D	175	68
PSRG - E	64	42	BRG - E	168	69
PSRG - F	34	21	BRG - F	58	35
PSRG - G	67	31	BRG - G	280	139
PSRG - H	39	27	BRG - H	67	42
PSRG - I	29	10	BRG - I	86	52
PSRG - J	11	09	**----**	**----**	**----**
All groups together	**372**	**180**	All groups together	**1,153**	**372**

PSRG = Pharmaceutical Sciences Research Group; BRG = Biology Research Group.

Table 1 lists the number of publications of the research groups. The Pharmaceutical Sciences research groups published 372 publications in 180 journals, including 67 joint publications

between the groups, while the Biology research groups generated 1,153 publications in 372 journals, and there are 119 joint publications between the groups.

For this paper, we adopted the overlay mapping methods based on a global journal map from Web of Science data (Leydesdorff & Rafols, 2012). Journals overlay maps were created for the panels, all individual research groups, and the combined research groups of each department. To this end, all Source titles (Journal titles hereafter) pertaining to the entire citable journal output of the panel members and the groups were retrieved and entered into network software, and overlay information was added to the global journal map. The overlap of research group and panel publications was visualized on a global journal map based on the retrieved journal titles, using the visualization program VOSviewer (van Eck & Waltman, 2010).

Analysis and Results

Panel profiles versus Group profiles

Pharmaceutical sciences panel publications are found in 300 different journals, whereas those of the combined Pharmaceutical Sciences groups cover 180 journals. The journal overlay maps for the Pharmaceutical Sciences combined groups (Fig. 1) and the panel (Fig. 2) clearly show that the publication scope of the panel is wider than that of the combined groups. The panel publications are strong (11.86%) in 'Pharmaceutical Research', 'British Journal of Clinical Pharmacology', and 'Archiv der Pharmazie' journals, whereas the research group publications are clustered (8.6%) in 'Kidney International', 'Planta Medica', 'Environmental Science & Technology' journals.

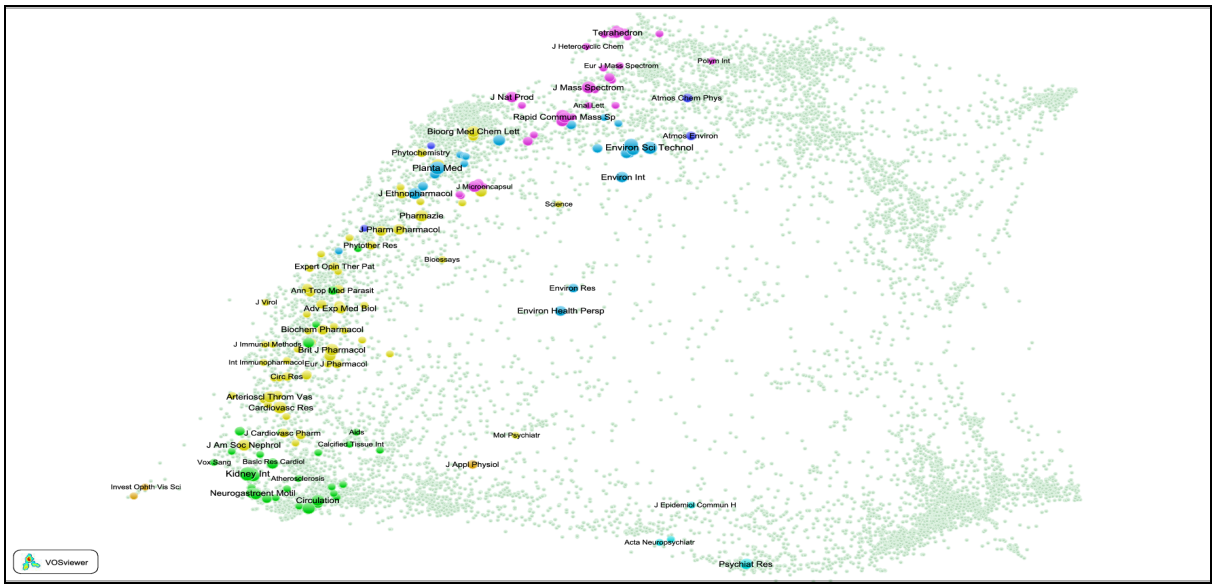


Figure 1. Pharmaceutical Sciences groups' publications overlay to the global journal maps.

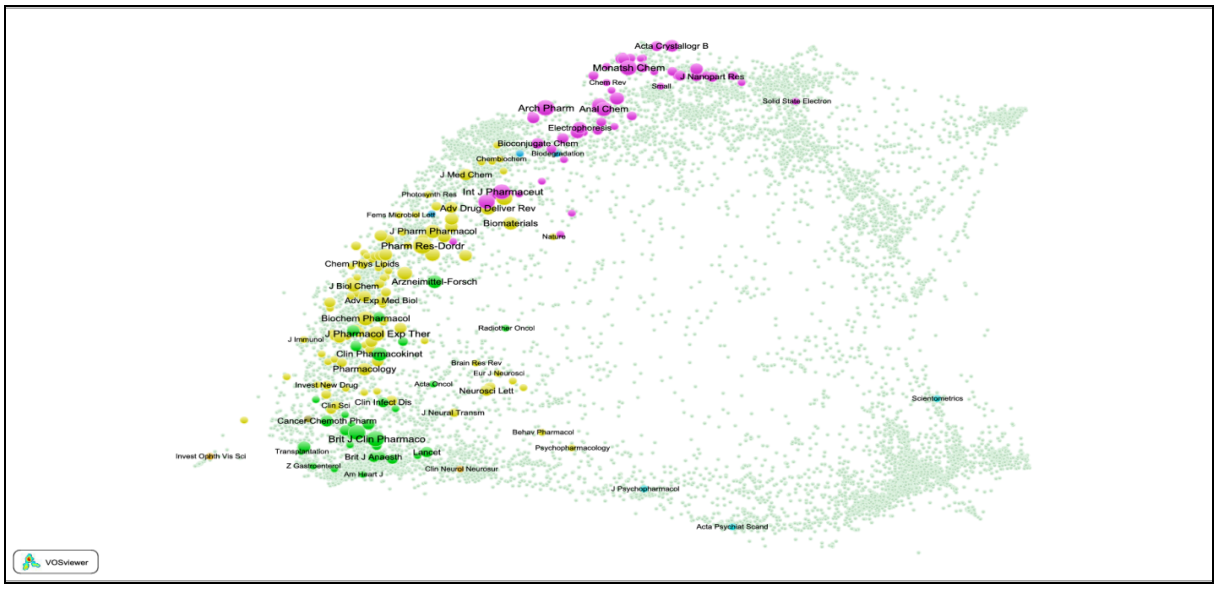


Figure 2. Pharmaceutical Sciences Panel publications overlay to the global journal maps.

Contrariwise, Biology panel publications appeared in 218 journals, while those of the combined Biology groups cover 372 journals. The overlay maps for the Biology department (Figs. 3 and 4) revealed a wider publication scope for the combined research groups compared to the Biology panel. The panel's publications are strong (8.58%) in 'Environmental Pollution', 'Biological Journal of the Linnean Society', and 'Journal of Experimental Biology', whereas the groups' publications tend to be mainly clustered (12.47%) in 'Experimental and Applied Acarology', 'General and Comparative Endocrinology', 'Journal of Experimental Biology'.

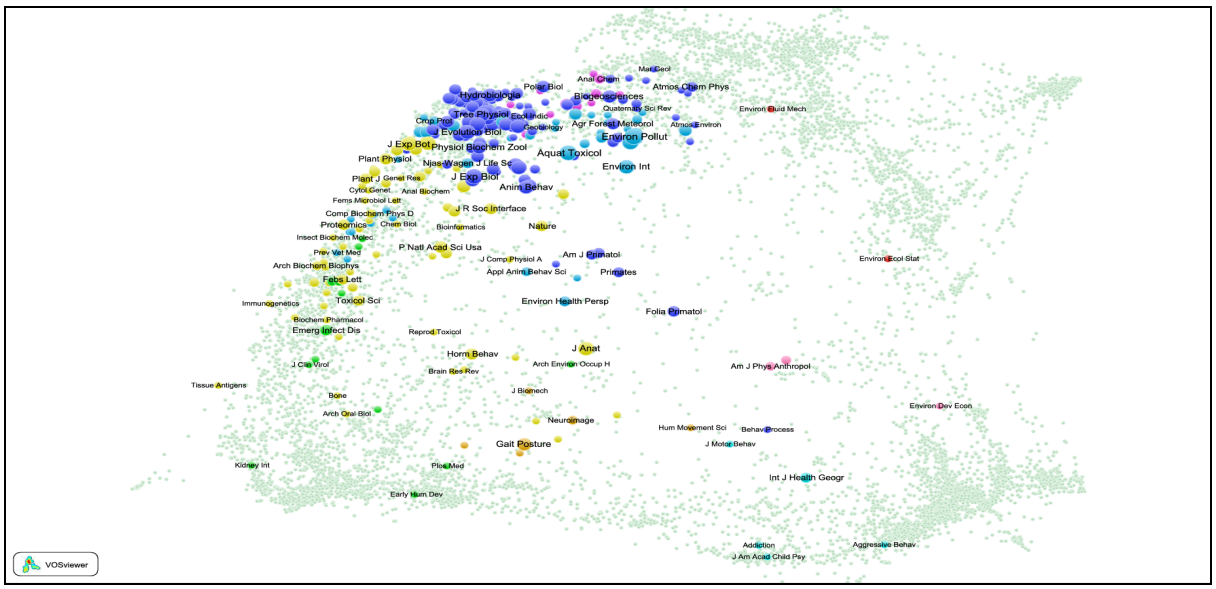


Figure 3. Biology groups' publications overlay to the global journal maps.

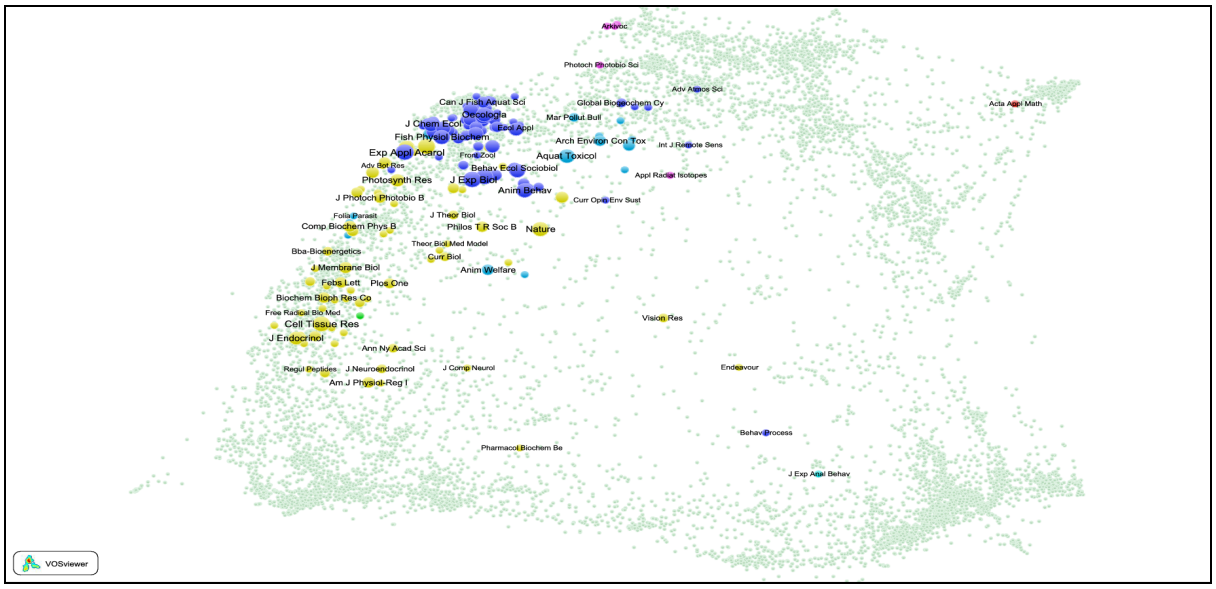


Figure 4. Biology Panel members' publications overlay to the global journal maps.

Table 2 shows that there is no common journal in the top five journals between the Pharmaceutical Sciences panel and groups. Table 2 further shows that there is only one common journal, *Journal of Experimental Biology*, (panel 3.82%, groups 2.26%) in the top five journals between Biology panel and groups.

Table 2: Top five Journals title for the panels and the groups

Panel publications			Group publications		
Pharmaceutical Sciences Department					
Journals Title	**Records**	**% of 1029**	**Journals Title**	**Records**	**% of 372**
Pharmaceutical Research	52	5.05	Kidney International	13	3.5
British Journal of Clinical Pharmacology	35	3.4	Planta Medica	11	2.96
Archiv der Pharmazie	35	3.4	Environmental Science Technology	8	2.15
Clinical Pharmacology Therapeutics	27	2.62	Journal of Mass Spectrometry	7	1.88
Monatshefte Fur Chemie	23	2.23	Chemosphere	7	1.88
Biology Department					
Journals Title	**Records**	**% of 786**	**Journals Title**	**Records**	**% of 1153**
Experimental and Applied Acarology	35	4.45	Environmental Pollution	40	3.47
General and Comparative Endocrinology	33	4.2	Biological Journal of the Linnean Society	33	2.86
Journal of Experimental Biology	30	3.82	Journal of Experimental Biology	26	2.26
Proceedings of the Royal Society B:Biological Sciences	22	2.8	Aquatic Toxicology	23	1.2
New Phytologist	22	2.8	Environmental Science Technology	22	1.91

Together, the Pharmaceutical Sciences panel and groups have 60 journals in common. In addition, 240 journals have panel publications but no group publications, while 120 journals contain group publications but no panel publications. Further, Biology panel and group publications were common in 93 journals. Moreover, 125 journals contained panel publications but no group publications and 279 journals have group publications but no panel publications.

These findings demonstrate that Pharmaceutical Sciences panel published more diversely than the groups, whereas the opposite is true for the Biology department. However, the Pharmaceutical Sciences panel overlaps in one third of the journals of groups' publications, whereas the Biology panel overlaps almost half the journals where biology groups have publications too.

Panel profile versus Individual group profile

Overlay maps of the publications of the individual groups were created, and subsequently compared with the two panel overlay maps. Most Pharmaceutical Sciences research groups have at least one journal in common with the panel; this is the case for PSRG-A (50%), PSRG-B (40.63%), PSRG-C (31.42%), PSRG-D (58.82%), PSRG-E (40.78%), PSRG-F (61.9%), PSRG-G (16.13%), PSRG- H (37.03%), and PSRG-J (20%). Only PSRG-I has none. All Biology research groups have one or more journals in common with the panel: BRG-A (41.51%), BRG-B (18.75%), BRG-C (33.33%), BRG-D (35.29%), BRG-E (42.65%), BRG-F (48.57%), BRG-G (35.97%), BRG-H (19.05%), BRG-I (25%).

These data show that the research outputs of three of the ten Pharmaceutical Sciences research groups (A, D, F) are 50–62 percent, four groups (B, C, E, H) are 30–40 percent, two groups (G, J) are 20 to 15 percent covered by the panels' expertise thematically, whereas one group (group I) is not covered at all. At the same time, three out of nine Biology research groups (A, E, F) are 40-50 percent, three research groups (C, D, G) are 30-40 percent, and another three research groups (B, H, I) are below 25 percent covered by the panel's expertise.

Conclusion

The results indicate that the Biology research groups published more diversely than the panel, which is similar to the findings in Rahman et al. (2014). However, the Pharmaceutical Sciences panel published more diversely than research groups, which is opposite to what was found in Rahman et al. (2014) where the research groups published more diversely in Web of Science subject categories than the panel did. The most likely reason is that all panel members' publications are taken into account (published over the course of over 20 years, often working in different countries and on different topics), whereas the research groups have a specific focus and choose the journals accordingly.

Pharmaceutical Sciences panel overlaps in one third of the journals of the corresponding group's publications, whereas the Biology panel overlaps in close to half the journals where Biology groups have publications. In addition, the number of publications in the same journals by the expert panel and research group varied, and a different range of coverage was found for different research groups. There is also a significant difference between maximum and minimum coverage based on discipline. To quantify which overlap leads to the best standard for evaluation, a considerable range of percentage of common journals between the panel and research group needs to be identified. The considerable range of percentage will express a well-covered, partially covered, and hardly covered expertise based on journal level matching. In subsequent analysis, we will compare results with corresponding results for other disciplines and explore other criteria for adequate relations between evaluation panels and groups.

Acknowledgments

This investigation has been made possible by the financial support of the Flemish Government to ECOOM, among others. The opinions in the paper are the authors' and not necessarily those of the government. The authors thank Nele Dexters for assistance.

References

Bornmann, L., Mutz, R., Marx, W., Schier, H., & Daniel, H.-D. (2011). A multilevel modelling approach to investigating the predictive validity of editorial decisions: do the editors of a high profile journal select manuscripts that are highly cited after publication? Journal of the Royal Statistical Society: Series A (Statistics in Society), *174*(4), 857–879.

Butler, L., & McAllister, I. (2011). Evaluating University research performance using metrics. *European Political Science*, *10*(1), 44–58.

Engels, T. C. E., Goos, P., Dexters, N., & Spruyt, E. H. J. (2013). Group size, h-index, and efficiency in publishing in top journals explain expert panel assessments of research group quality and productivity. *Research Evaluation*, *22*(4), 224–236.

Lawrenz, F., Thao, M., & Johnson, K. (2012). Expert panel reviews of research centers: The site visit process. E*valuation and Program Planning*, *35*(3), 390–397.

Leydesdorff, L., & Rafols, I. (2012). Interactive overlays: A new method for generating global journal maps from web-of-science data. *Journal of Informetrics*, 6(2), 318–332.

Nedeva, M., Georghiou, L., Loveridge, D., & Cameron, H. (1996). The use of co-nomination to identify expert participants for Technology Foresight. *R&D Management*, *26*(2), 155–168.

Rahman, A. I. M. J., Guns, R., Rousseau, R., & Engels, T. C. E. (2014). Assessment of expertise overlap between an expert panel and research groups. In Ed Noyons (Ed.), Context Counts: Pathways to Master Big and Little Data. Proceedings of the Science and Technology Indicators Conference 2014 Leiden (pp. 295–301). Leiden: Universiteit Leiden.

Rons, N., De Bruyn, A., & Cornelis, J. (2008). Research evaluation per discipline: a peer-review method and its outcomes. *Research Evaluation*, *17*(1), 45–57.

Tseng, Y.-H., & Tsay, M.-Y. (2013). Journal clustering of library and information science for subfield delineation using the bibliometric analysis toolkit: CATAR. *Scientometrics*, *95*(2), 503–528.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.

Contextualization of Topics - Browsing through Terms, Authors, Journals and Cluster Allocations[1]

Rob Koopman[1], Shenghui Wang[1] and Andrea Scharnhorst[2]

[1] *rob.koopman@oclc.org*, [1] *shenghui.wang@oclc.org*
OCLC Research, Schipholweg 99, Leiden (The Netherlands)

[2] *andrea.scharnhorst@dans.knaw.nl*
DANS-KNAW, Anna van Saksenlaan 51, The Hague (The Netherlands)

Abstract
This paper builds on an innovative Information Retrieval tool, *Ariadne*. The tool has been developed as an interactive network visualization and browsing tool for large-scale bibliographic databases. It basically allows to gain insights into a topic by contextualizing a search query (Koopman et al., 2015). In this paper, we apply the Ariadne tool to a far smaller dataset of 111,616 documents in astronomy and astrophysics. Labeled as the *Berlin dataset*, this data have been used by several research teams to apply and later compare different clustering algorithms. The quest for this team effort is how to delineate topics. This paper contributes to this challenge in two different ways. First, we produce one of the different cluster solutions and second, we use *Ariadne* (the method behind it, and the interface - called *LittleAriadne*) to display cluster solutions of the different group members. By providing a tool that allows the visual inspection of the similarity of article clusters produced by different algorithms, we present a complementary approach to other possible means of comparison. More particularly, we discuss how we can - with *LittleAriadne* - browse through the network of topical terms, authors, journals and cluster solutions in the *Berlin dataset* and compare cluster solutions as well as see their context.

Conference Topic
Methods and techniques; Mapping and Visualization

Introduction
What are essence and boundary of a scientific field? How can a topic be defined? Those are questions that are core to bibliometrics. Rigour and stability in defining boundaries of a field are important for research evaluation and funding distribution. However, if you as a researcher would seek for information about a certain topic of which you are not an expert yet, your information needs are quite different. Among the many possible hits for a search query you might want to know which are core works (articles, books) and which are rather peripheral. You might want to use different rankings (Mutschke & Mayr, 2014) or get some context. On the whole you would have less need to define a topic and a field in a bijective, univocal way. The same holds if you want to compare different clustering algorithms. Here again, you are in need to illustrate similarities and differences between different allocations of documents to clusters. Ways to contextualize them and browse through these contexts would be desirable. This is our starting point.
Decades of bibliometrics research have produced many different algorithms to cluster bibliographic records. They often focus on one entity of the bibliographic record. For example, articles and terms those articles contain (in title, abstract and/or full text) form a

[1] This paper is submitted as part of the Special Session at the ISSI conference 2015 "Same data – different results? The performative nature of algorithms for topic detection in science".

bipartite network from which we can either build a network of related terms (co-word analysis) or a network of related articles (based on shared words). The first method, sometimes also called lexical, has been often applied in scientometrics to produce so-called topical or semantic maps. The same exercise can be applied to authors and articles, articles and journals, in effect each element of the bibliographic record for an article (Havemann & Scharnhorst, 2012). If we extend the bibliographic record with the list of references, we enter the area of citation analysis. Here two methods are widely used: direct citations (known as delivering often sparse matrices) and co-citation maps (known as a good method to identify research fronts). Hybrid methods combine citation and lexical analysis (e.g., Zitt & Bassecoulard, 2006; Janssens et al., 2009). The majority of studies applies one technique. But, sometimes analysis and visualization of multi-partite networks can be found (cf. Van Heur, Leydesdorff, & Wyatt 2013).

Each of the possible different network representations of articles stands for another aspect of connectivity between published scientific works. Co-authorship networks shed light on the social dimension - the invisible colleges - of knowledge production (Mali et al., 2012; Glänzel & Schubert, 2004). Citation relations are interpreted as traces of flows of knowledge (Price, 1965; Radicchi, Fortunato, & Vespignani, 2012). Depending on which element of the bibliographic record is used, we obtain different perspectives on how a field or a topic is to be conceived - as conceptional, cognitive unit; as a community of practice; or as institutionalized in journals. We can call this a *measurement effect*. Another source of variety next to differences resulting from what to analyze is how to analyze it. Finding clusters is part of network analysis. But, clusters can be defined in different ways, and aside of different possible definitions of cluster to determine them for a large-scale network can be algorithmically challenging. Consequently, we find different solutions for one algorithm (if parameters in the algorithm are changed) and different solutions for different algorithms. One could call this an effect of the *choice of instrument for the measurement*. Last but not least, we can ask ourselves, if topics clearly delineated from each other really exist. Often in science very different topics still are related to each other. There exist unsharp boundaries and almost invisible long threads in the *fabric of science* (Boyack & Klavans, 2010), which might inhibit to find a contradiction-free solution. There is a seeming paradox between the fact that experts often can rather clearly identify what belongs to their field or a certain topic, and that it is so hard to quantitatively represent this with bibliometrics methods. However, a closer look into science history and science and technology studies reveals that what belongs to a field or a topic can still differ substantially also in the opinions of different experts; it changes over time; and even a defined canon or body of knowledge determining the essence of a field or a topic might be still subject to controversies and changes.

In the quest to define a topic two things collide. The principal, methodological and data-based ambiguity of what a topic is and the necessity to define a topic for purposes of education, knowledge acquisition and evaluation. This makes it such an intriguing problem to be solved. Because different perspectives can be valid, there is also a need to preserve the above sketched diversity or ambiguity. Having said this, for the sake of scientific reasoning it is also necessary to be able to further specify the validity and appropriateness of different methods to define topics and fields. This paper contributes to the development of methods to compare algorithms and to visualize their different results.

We contribute to this sorting out process in two different ways. First, we apply standard clustering techniques to a specific article matrix built in a specific way from what we call a semantic matrix, in which rows are formed by entities from the bibliographic records of the articles (author names, journal ISSNs, topical terms, subjects, and other characteristics), columns by reduced dimensions from co-occurrence of entities and topical terms (one subset of the entities) over the whole set of articles. While we explain this in detail later, let us note

here that the approach is conceptually more similar to classical information retrieval techniques based on Salton's vector space model than to usual bibliometrical mapping techniques (Salton & McGill, 1983).
In a second step, we present an interactive visual interface called *LittleAriadne* that allows to display the context around those extracted and networked entities. The interface responds to a search query with a network visualization of most related terms, authors, journals and (other) cluster numbers. The query entry can be words, authors, but also cluster solutions. The displayed nodes or entities around a query term represent to a certain extent the context of this query. Depending on the query entry, we will see more or less other terms, journals, or authors. The interface allows to foreground one of entity types by selecting them. The interface has been originally developed for a much larger bibliographic database. In this paper our research questions are:

- Q1: How does the *Ariadne* algorithm work on a much smaller, field specific dataset? What possibility do we have to relate the produced contexts to domain knowledge?
- Q2: Can we use *Ariadne* to label the clusters produced by the different methods?
- Q3: Can we use *Ariadne* to compare different cluster assignments of papers, by treating those cluster assignments as additional entities? What can we visually learn about the topical nature of these clusters?

Data

The dataset used in this paper – called *Berlin dataset* - entails papers published in the period 2003-2010 in 59 astrophysical journals. Those papers have been downloaded from the Web of Science in the context of a German-funded research project called "Measuring Diversity of Research," conducted at the Humboldt-University Berlin - hence the coined name *Berlin dataset*. It contains 120,007 records in total. Eventually, 111,616 records of the document types Article, Letter and Proceedings Paper have been treated with different clustering methods (see the other contributions for this special session).
Some of those cluster outcomes have been shared and are later displayed in the visual interactive interface. Table 1 shows the label of the different sets of clusters *x* we have included in *LittleAriadne*, whereby *x={a, b, ..., f}*. We have noted by which group cluster solutions were produced in the *Source* column. Each clustering method produced a set of clusters, whereby *y* stands for the number of clusters in a set. In our paper we used cluster solutions from CWTS (label: cwts 1.8), Cornell, Humboldt-University Berlin (hu), SciTech (sts-rg), KU Leuven (bc15) and one of our own (oclc_20). Except of cluster set *e*, they are all of the same order of magnitude. Because *Ariadne* relies on statistics across a corpus of articles as large as possible to produce semantic relatedness, we decided to discard clusters with less than 4 articles. But, from the solutions with many clusters (*d, e*) we decided not to display all. The last column in Table 1 gives the final numbers of the clusters from different clustering solutions.

Method

Ariadne - an interactive visualization to navigate entities from large bibliographic databases

The *Ariadne* algorithm has been developed on top of the article database, *ArticleFirst* of OCLC. The interface, accessible at http://thoth.pica.nl/relate, allows users to visually and interactively browse 35 thousand journals, 3 million authors, 1 million topical terms associated with 65 million articles (Koopman et al., 2015). For the purpose of this paper, we applied the same method on the Berlin dataset and built an instantiation, *LittleAriadne*, accessible at http://thoth.pica.nl/astro/relate.

Table 1. Statistics of clusters generated from different methods.

x	Source	*y=#Cluster*	*#Cluster in Ariadne*
a	cwts 1.8	23	23
b	cornell	23	23
c	oclc_20	20	20
d	hu	139	48
e	sts-rg	5664	229
f	bc15	15	15

Table 2. An article from the Berlin dataset.

Article ID	ISI:000276828000006
Title	On the Mass Transfer Rate in SS Cyg
Abstract	The mass transfer rate in SS Cyg at quiescence, estimated from the observed luminosity of the hot spot, is log M-tr = 16.8 +/- 0.3. This is safely below the critical mass transfer rates of log M-crit = 18.1 (corresponding to log T-crit(0) = 3.88) or log M-crit = 17.2 (corresponding to the ""revised"" value of log T-crit(0) = 3.65). The mass transfer rate during outbursts is strongly enhanced
Author	[author:smak j]
ISSN	[issn:0001-5237]
Subject	[subject:accretion, accretion disks] [subject:cataclysmic variables] [subject:disc instability model] [subject:dwarf novae] [subject:novae, cataclysmic variables] [subject:outbursts] [subject:parameters] [subject:stars] [subject:stars dwarf novae] [subject:stars individual ss cyg] [subject:state] [subject:superoutbursts]
Cluster label	[cluster:a 19] [cluster:b 16] [cluster:c 15] [cluster:d 51] [cluster:e 17] [cluster:f 1]

Table 2 shows for one example article from the *Berlin dataset* those fields of the bibliographic record that we used for *LittleAriadne*. It also shows which categories of entities we have. The ISI record ID has been used among the teams to compare solutions. For *Ariadne* as an interface, it does not matter. *Ariadne* is different from a usual Information Retrieval search engine because it does not primarily deliver lists of documents matching a query, but a network of those entities which profile in the whole corpus 'resonate' most with the query entry. We come back to this aspect later. We further define so-called topical terms. Topical terms are frequent single or two-word phrases extracted from all titles and abstracts, for example, "mass transfer" and "quiescence" in our example. Next to the topical term, each author name is treated as an entity. In Table 2 we display the author name (and other entities below) in a syntax that can be used in the search field of the interface to search for a specific author. The next type of entities is the ISSN number of a journal. One can search for a single journal using the ISSN number, in the visual interface the journal title is used as label for a node representing a journal. Further, we have so-called subjects as separate entity type. Those subjects origin from the fields "Author Keywords" and "Keywords Plus" of the original Web

of Science records. As last type of entities we add - and this is specific for *LittleAriadne* - to each of the articles cluster labels from their assignments to clusters produced by different teams. For example, the article in Table 2 has been assigned to cluster number 19 by source a (cwts 1.8) number 16 by source b (cornell), and so on. In other words, we treat the cluster assignments of articles as they would be classification numbers or additional subject headings.

With the above detailed parsing of the bibliographic records we then build the matrix C (see Figure 1). In C, frequent topical terms, subjects, author names, cluster labels and journals appearing in the *Berlin dataset* form the rows, and topical terms as well as subjects are listed in columns. The relatedness between all entities is computed based on the *context* they share, instead of direct co-occurrences in the data. The context of these entities is captured by their co-occurrences with topical terms and subjects, that is, we count how often an author, or a cluster label co-occurs with a certain topical term or subject in an article, summing up over all articles in the corpus. In the Berlin dataset, we have in total 90,343 entities, including 59 journals, 27,027 author names (single instances, no author disambiguation applied), 358 cluster IDs, 39,577 topical terms and 23,322 subjects. This would produce a sparse matrix of roughly 90K x 63K that is expensive for computation.

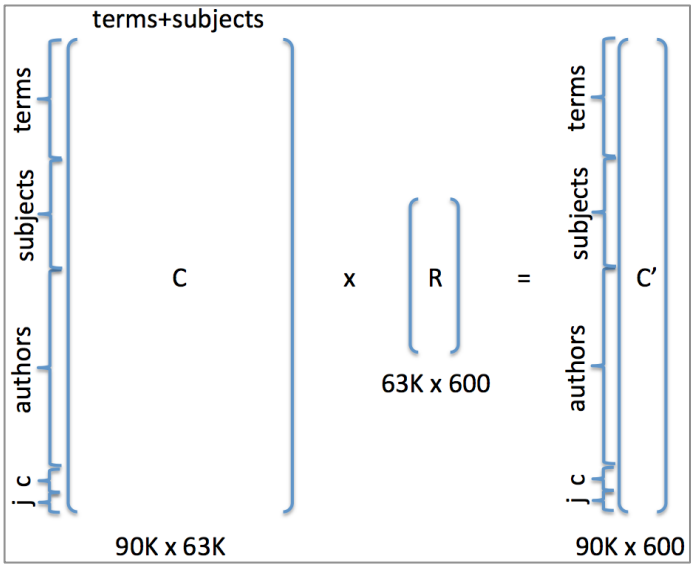


Figure 1. Dimension reduction using Random Projection.

To make the algorithm scale and produce a responsive visual interface, we applied *Random Projection* (Johnson & Lindenstrauss, 1984; Achlioptas, 2003) to reduce the dimensionality of the matrix. As shown in Figure 1, by multiplying C with a 63K x 600 matrix of randomly distributed -1 and 1, the original 90K x 63K matrix C is reduced to a *Semantic Matrix C'* of the size of 90K x 600, with each row vector representing the semantics of an entity. With this Semantic Matrix, the interactive visual interface dynamically computes the most related entities (e.g. ranked by cosine similarity) to a search query and presents a networked visualization of the context of a query term whereby entities are positioned closer to each other if they are more related to each other.

OCLC clusters production - Clustering the Berlin dataset using the Semantic Matrix

The Ariadne interface provides a networked view about entities associated with articles, but it does not produce article clusters straightaway. In order to cluster articles, we need to build a semantic representation of each article. We receive the semantic representation for an article by the following steps. For each article, we look up all entities related to this article in the

Semantic Matrix C'. For our example in Table 2 we have one vector representing the single author of that article in the whole Semantic Matrix, 12 vectors representing the subjects, one vector for the journal, 6 vectors representing the cluster labels and *n* vectors for all extracted topical words. In other words, each article is represented by a subset of vectors and the vector components correspond to the dimensions of the Semantic Matrix. We then take the average of those single entity vectors as the semantic representation of a specific article. All articles together form a matrix M with 111,616 rows and 600 columns. We applied a standard clustering technique - the MiniBatchKmeans method (Sculley 2010) - to M. We used the scikit-learn python library (http://scikit-learn.org/) for this. Applied to the *Berlin dataset* we receive a cluster solution with a comparable size of k=20 clusters, labeled as oclc_20, and a unique assignment of articles to this cluster.

Results - The Berlin dataset in *LittleAriadne*

We used the visual, interactive interface built for the *Berlin dataset* to the context around a specific cluster solution and the similarity between different ones. For this we performed different experiments, which correspond to the research questions Q1-Q3 of the introduction

- Experiment 1: We used *LittleAriadne* as information retrieval tool. We searched with query terms, inspected and navigated through the resulting network visualization. (Q1)
- Experiment 2: We used the semantic matrix to provide the most related topical terms for each cluster as an approximation of cluster labels. (Q2)
- Experiment 3: We used the query syntax to display two or more cluster solutions together in one overview. (Q3)

Experiment 1 - Information retrieval

In *LittleAriadne* we can now study the *Berlin dataset* as any other dataset. Figure 2 gives a snapshot of the context about "magnetic flux" used as query term.[2] The most related topical terms and subjects are shown, together with 3 most related clusters provided by CWTS, Cornell and SciTech (coded in different colors). Each node is clickable which leads to another visualization of the context of the selected node. When mousing over a node, one sees how often this entity occurs in the whole corpus. Given that different statistical methods are at the core of the Ariadne algorithm, this gives an indication of the reliability of the suggested position and links. In the interface one can further refine the display. For instance, one can choose the number of nodes to be shown or decide to limit the display to only authors, journals, topical terms or clusters. Within the interface, one can navigate the context of entities in the *Berlin dataset* by seamlessly travelling between authors, journals, topical terms and clusters in a visual and interactive way.

Experiment 2 -Labeling clusters

Please note, that in *LittleAriadne* we cannot see the position of articles in relations to the different entities. One could say that the articles produce the elements of the networked context, but they themselves are distributed over it. What we can do is to switch to a view that shows most related topical terms, subjects, journals, authors, and other clusters. The outcome of such a *click-through* action is shown in Figure 3.[3] In this example, the most related topical terms, subjects, one journal, and four other clusters are presented as the contextual information about the cluster "a 2".

[2] Figure 2 is accessible at http://thoth.pica.nl/astro/relate?input=magnetic+flux.

[3] Figure 3 is accessible at http://thoth.pica.nl/astro/relate?input=%5Bcluster%3Aa+2%5D.

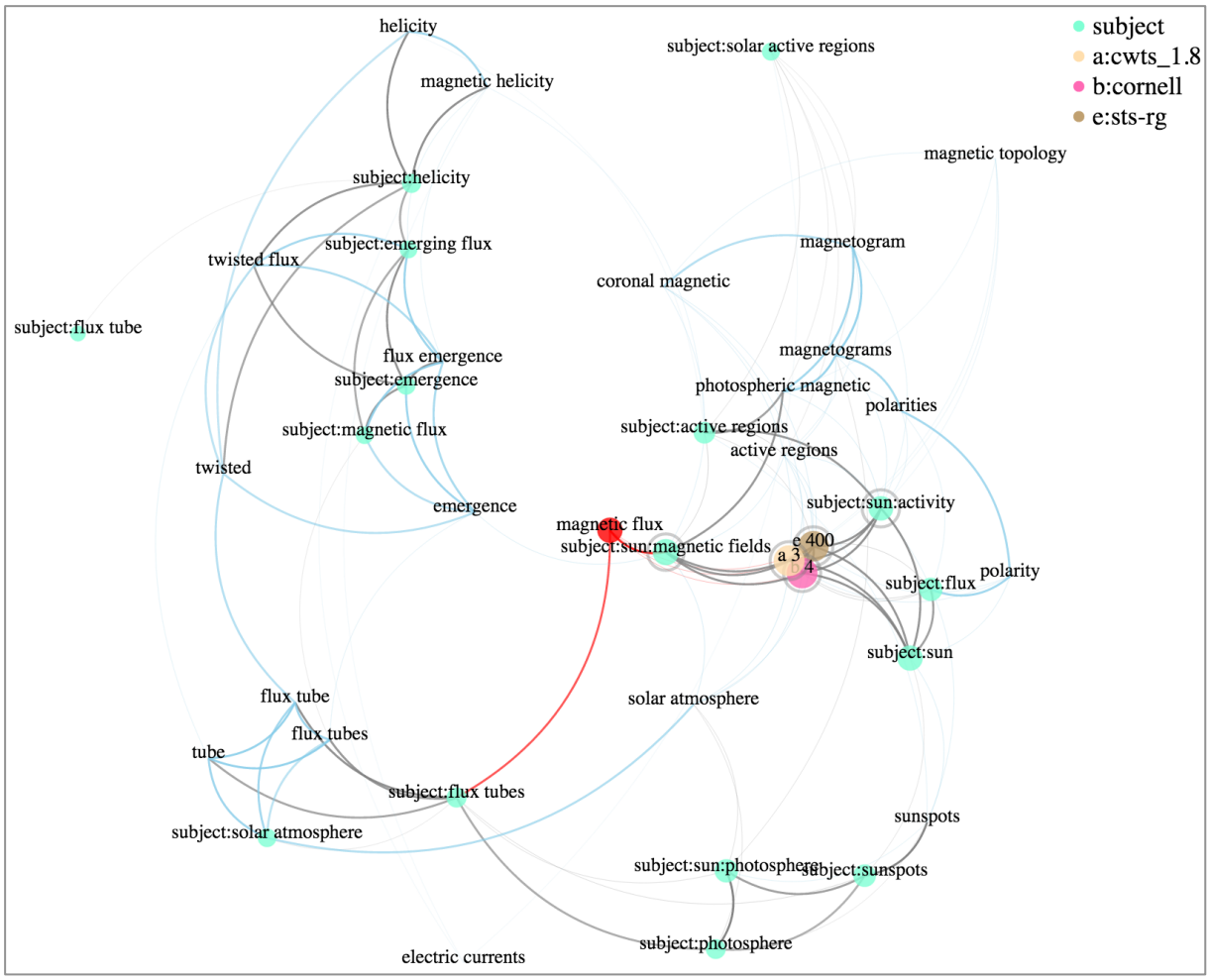


Figure 2. Context around "magnetic flux".

It is now possible to label each cluster using the most related topical terms. As shown in Table 3, the 9 topical terms most related to cluster "a 2" are "cosmology," "dark energy," "density perturbations," "cosmologies," "planck," "cosmological," "spatial curvature," "inflationary," and "inflation." Together they give a rough idea about what this cluster with 8,954 articles is about, but it requires domain expertise to evaluate and transform them into real cluster labels, meaning representing names of specialties, topics or fields used by the scientific community, a well-known problem of bibliometric mapping (Noyons, 2005).

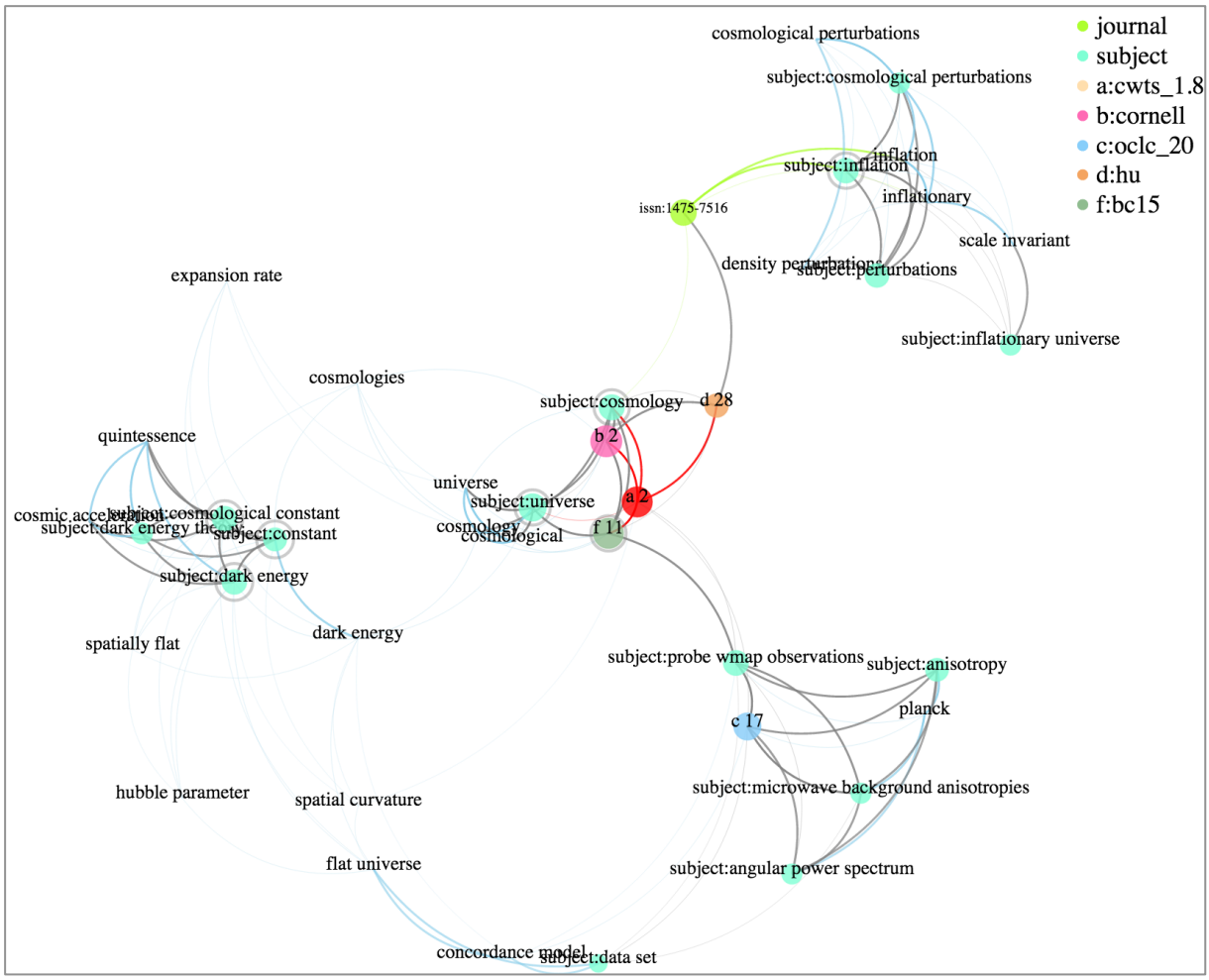


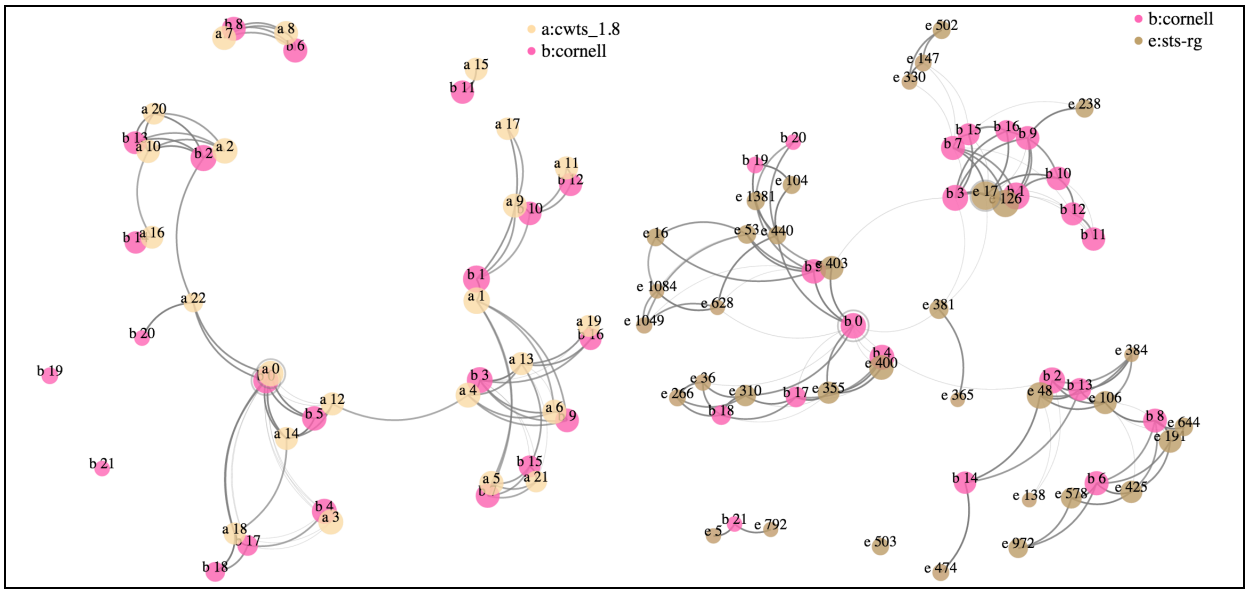
Figure 3. The contextual view of cluster “a 2”.

Table 3. Top related topical terms.

Cluster ID	Top 9 most related topical terms
a 2	"cosmology" "dark energy" "density perturbations" "cosmologies" "planck" "cosmological" "spatial curvature" "inflationary" "inflation"
b 2	"cosmology" "cosmological constant" "cosmologies" "cosmological" "universes" "dark energy" "quadratic" "tensor" "planck"
c 17	"power spectrum" "cosmological parameters" "cmb" "last scattering" "anisotropies" "microwave background" "power spectra" "planck" "cosmic microwave"
d 28	"density perturbations" "inflationary" "inflation" "dark energy" "scale invariant" "spatial curvature" "cosmological perturbations" "inflationary models" "cosmologies"
f 11	"cosmology" "cosmological" "dark energy" "universe" "planck" "density perturbations" "cosmologies" "spatial curvature" "flat universe"

Experiment 3 - Comparing cluster solutions

In *LittleAriadne* we extended the interface with a possibility to compare sets of clusters. In Figure 4 (a) we can visually see the high similarity between clusters from CWTS and those from Cornell.[4] Nearly each CWTS cluster is accompanied by a Cornell cluster. Figure 4 (b) shows two other sets of clusters which partially agree with each other but also clearly have different capacity in distinguishing different clusters.[5] Figure 5 shows all the cluster entities from all six clustering solutions. Given the amount of the clusters, it is difficult to grasp the detailed difference between solutions. However, this visualization does provide a general overview of all the clustering solutions, based on their similarities to each other.



(a) Highly similar (between CWTS 1.8 and Cornell)

(b) Partially agreeing (between Cornell and SciTech)

Figure 4. Comparison between sets of clusters.

Discussion and Conclusion

We present a method and an interface that allows browsing through the contexts of entities, such as topical terms, authors, journals and subjects associated with a set of documents. We have applied the method to the problem of topic delineation addressed in this special session. Because the tool shows (local) context and not the position of single documents in relation to clusters we think it has a potential to be complementary to any other method of cluster comparison. In particular, we have asked how the *Ariadne* algorithm works on a much smaller, field specific dataset. Not surprisingly, compared with our exploration in the ArticleFirst interface, we find more consistent representations. That means that specific vocabulary is displayed, which can be cross-checked in Wikipedia or Google Scholar, for which the interface offers a direct click through.

[4] Figure 4(a) is accessible at http://thoth.pica.nl/astro/relate?input=%5Bcluster%3Aa%5D%5Bcluster%3Ab%5D&type=S&show=50.

[5] Figure 4(b) is accessible at http://thoth.pica.nl/astro/relate?input=%5Bcluster%3Ae%5D%5Bcluster%3Ab%5D&type=S&show=300.

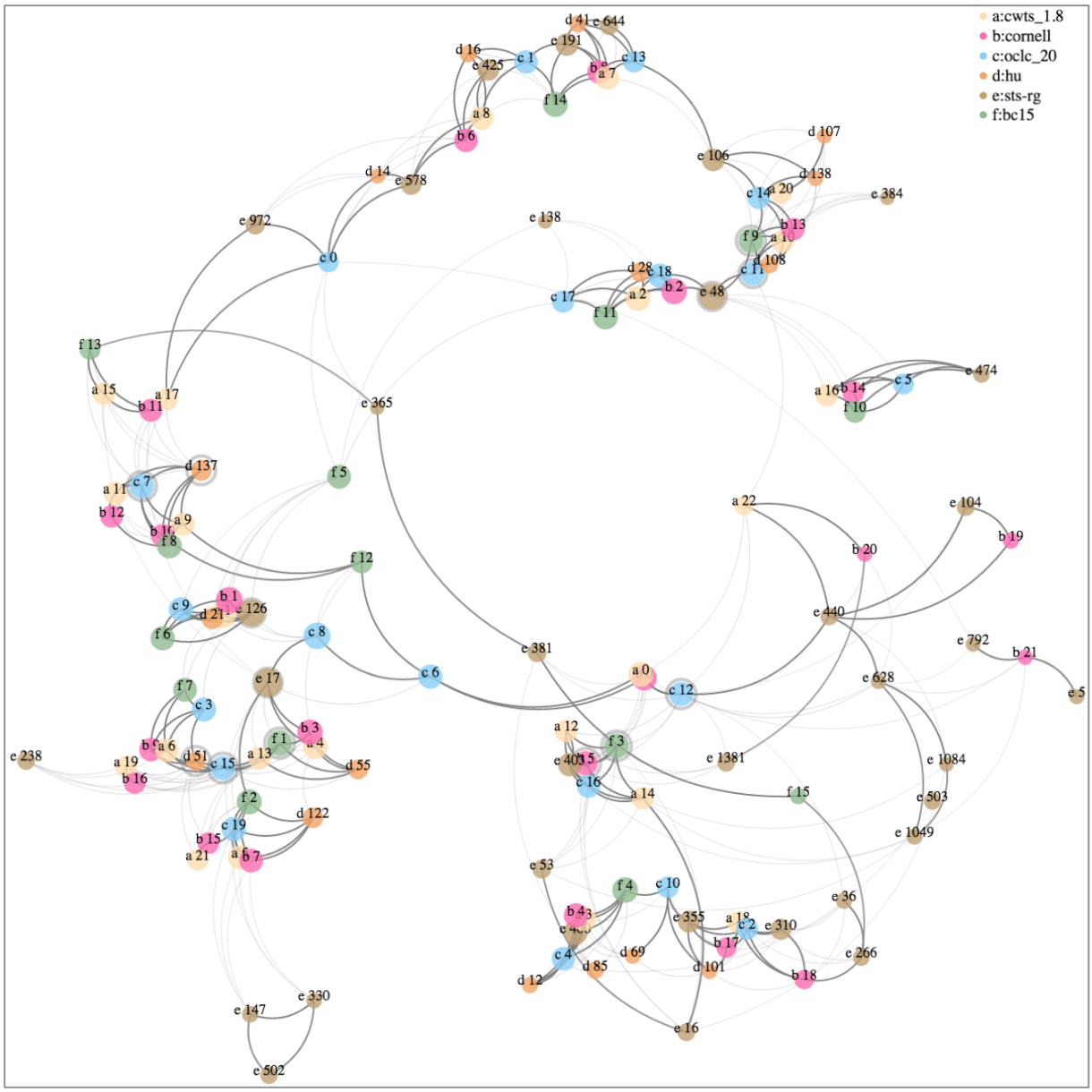


Figure 5. Comparing clusters from 6 clustering solutions.

On the other hand, the bigger number of topical terms in the larger database leads to a situation where almost every query term produces a response. In *LittleAriadne* searches for e.g., literary persons such as Jane Austen retrieve nothing - a blank screen. In preparation of this paper we surfed through the interface, and compared the most relevant topical terms around a cluster to other classifications used in Astrophysics, such as Physics and Astronomy Classification Scheme (PACS®[6]). In this punctual exploration we did find correlations between the names of PACS classes (subclasses, and related controlled vocabulary) and the selected topical terms in *LittleAriadne*. We will further compare the context around clusters and the suggested related topical terms with labels produced by other teams in this special session. Ultimately, the discussion with domain experts belongs to a proper evaluation of the interface. We demonstrated that we can use *LittleAriadne* to compare different cluster solutions mutually and even generate a wider overview. We will discuss in the special session how Ariadne can further be of use in the comparison of clustering and delineation of topics.

[6] http://www.aip.org/publishing/pacs/pacs-2010-regular-edition

At least, we hope that this interactive tool supports discussion about different clustering algorithms and helps to find the right meaning of clusters, and appropriate labels for them.
We also have plans to further develop the Ariadne algorithm. The Ariadne algorithm is general enough to accommodate additional types of entities to the semantic matrix. In the future, we plan to add citations, publishers, conferences, etc. with the aim to provide a richer contextualization of entities. We also plan to add links to articles that contribute to the contextual visualization, this way strengthening the usefulness of *Ariadne* not only for the associative exploration of contexts similar to scrolling through a systematic catalogue, but also as a direct tool for document retrieval. In this context we plan to further compare *LittleAriadne* and *Ariadne*. In a first attempt, we 'projected' the astrophysical documents into *ArticleFirst* by looking them up in the large semantic matrix built for Ariadne. We found the resulting representations less consistent when browsing through. That is not a surprise, because when merging them you see how field-specific content fits and miss-fits into many other contextualizations. The advantage of *LittleAriadne* is the confinement of the dataset to one scientific field and topics within. We hope by continuing such experiments also to learn more about the relationship between genericity and specificity of contexts, and how that can be best addressed in information retrieval.

Acknowledgments

Part of this work has been funded by the COST Action TD1210 KnoweScape.

References

Achlioptas, D. (2003) Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, *66*, 4, 671-687.

Boyack, K.W. & Klavans, R. (2010). Weaving the Fabric of Science. Courtesy of Kevin W. Boyack and Richard Klavans, SciTech Strategies, Inc. In K. Börner & E.F. Hardy (Eds), *6th Iteration (2009): Science Maps for Scholars*, Places & Spaces: Mapping Science. http://scimaps.org.

Glänzel, W., & Schubert, A. (2004). Analysing Scientific Networks through Co-authorship. In *Handbook of Quantitative Science and Technology Research* (pp. 257–276). doi:10.1007/1-4020-2755-9_12

Havemann, F., & Scharnhorst, A. (2012). Bibliometric Networks. *Arxiv Preprint: arXiv:1212.5211 [cs.DL]*, 20. Digital Libraries; Physics and Society. Retrieved from http://arxiv.org/abs/1212.5211

Janssens, F., Zhang, L., Moor, B. De, & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing and Management*, *45*(6), 683–702. doi:10.1016/j.ipm.2009.06.003

Johnson, W., & Lindenstrauss, J. (1984) Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics, 26*, 189–206.

Koopman, R., Wang, S., Scharnhorst, A., & Englebienne, G. (2015). Ariadne's Thread — Interactive Navigation in a World of Networked Information. In *CHI '15 Extended Abstracts on Human Factors in Computing Systems*. Seoul, South Korea, April 18-23, 2015 ACM 978-1-4503-3146-3/15/04, Preprint available at http://arxiv.org/abs/1503.04358

Mali, F., Kronegger, L., Doreian, P., & Ferligoj, A. (2012). Dynamic scientific co-authorship networks. In A. Scharnhorst, K. Börner, & P. van den Besselaar (Eds.), *Models of Science Dynamics* (pp. 195–232). Berlin, Heidelberg: Springer International Publishing. doi:10.1007/978-3-642-23068-4_6

Mutschke, P., & Mayr, P. (2014). Science models for search: a study on combining scholarly information retrieval and scientometrics. *Scientometrics*, 102(3): 2323-2345. doi:10.1007/s11192-014-1485-2

Noyons, C. (2005). Science Maps within a Science Policy Context. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 237–255). Springer International Publishing. doi:10.1007/1-4020-2755-9_11

Price, D. J. de Solla. (1965). Networks of Scientific Papers. *Science (New York, N.Y.)*, *149*, 510–515. doi:10.1126/science.149.3683.510

Radicchi, F., Fortunato, S., & Vespignani, A. (2012). Citation networks. In A. Scharnhorst, K. Börner, & P. van den Besselaar (Eds.), *Models of Science Dynamics* (pp. 233–257). Berlin, Heidelberg: Springer International Publishing. doi:10.1007/978-3-642-23068-4_7

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Sculley, D. (2010). Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 1177-1178. DOI=10.1145/1772690.1772862 http://doi.acm.org/10.1145/1772690.1772862

Van Heur, B., Leydesdorff, L., & Wyatt, S. (2013). Turning to ontology in STS? Turning to STS through "ontology." *Social Studies of Science*, *43*(3), 341–362. doi:10.1177/030631271245814

Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing and Management*, *42*(6), 1513–1531. doi:10.1016/j.ipm.2006.03.016

A Link-based Memetic Algorithm for Reconstructing Overlapping Topics from Networks of Papers and their Cited Sources

Frank Havemann[1], Jochen Gläser[2] and Michael Heinz[1]

[1] *frank.havemann@ibi.hu-berlin.de* [1] *michael.heinz@rz.hu-berlin.de*
Humboldt-Universität zu Berlin, Berlin School of Library and Information Science, Dorotheenstraße 26, 10099 Berlin (Germany)

[2] *jochen.glaser@ztg.tu-berlin.de*
TU Berlin, Center for Technology and Society, Hardenbergstr. 16-18, D-10623 Berlin (Germany)

Abstract
In spite of recent advances in field delineation methods, enduring problems such as the impossibility to justify necessary thresholds and the difficulties in comparing thematic structures obtained by different algorithms leave bibliometricians with a sense of uneasiness about their methods. In this paper, we propose and demonstrate a new approach to the delineation of thematic structures that attempts to fit the methods for topic delineation to the properties of topics. We derive principles of topic delineation from a theoretical discussion of thematic structures in science. Applying these principles, we cluster citation links rather than publication nodes, use predominantly local information and grow communities of links from seeds in order to allow for pervasive overlaps of topics. The complexity of the clustering task requires the application of a memetic algorithm that combines probabilistic evolutionary strategies with deterministic local searches. We demonstrate our approach by applying it to a network of 14,954 Astronomy & Astrophysics papers and their cited sources.

Conference Topic
Methods and techniques (special session on algorithms for topic detection)

Introduction
The identification of thematic structures (topics or fields) in sets of papers is one of the recurrent problems of bibliometrics. It was deemed one of the challenges of bibliometrics by van Raan (1996) and is still considered as such despite the significant progress and a plethora of methods available. Major developments since van Raan's paper include approaches that cluster the whole Web of Science based on journal-to-journal citations, co-citations, or direct citations, the advance of hybrid approaches that combine citation-based and term-based techniques, and term-based probabilistic methods (topic modelling). However, methodological problems endure and leave bibliometricians with a sense of uneasiness about their methods. Advanced methods still apply thresholds that must be arbitrarily set and adapted to the specific structures that shall be obtained. The relevance of the structures identified by bibliometric methods are difficult to verify independently, and the relationships between thematic structures are difficult to assess. A recent analysis by Hric et al. (2014) found that current algorithms for the detection of communities in network of papers respond to topological properties of networks but not necessarily to the underlying real-world properties of nodes clustered. This observation casts further doubts on the fundamental assumption underlying bibliometric methods for topic delineation, namely that the topics reconstructed using *structural properties of networks of papers* reflect *thematic properties of the research published in those papers*.

In this paper, we propose and demonstrate a new approach to the delineation of thematic structures. We derive principles of topic delineation and criteria for the assessment of algorithms from a theoretical discussion of properties of thematic structures in science. Applying these principles, we cluster citation links rather than publication nodes, use predominantly local information, and grow communities from seeds in order to allow for

pervasive overlaps of topics. The complexity of the clustering task requires the application of a memetic algorithm that combines nondeterministic evolutionary strategies with deterministic local searches. We demonstrate our approach by applying it to a network of 14,954 Astronomy & Astrophysics papers and their cited sources.

Strategy, Methods and Data

Theoretical considerations and strategy

We define topics as theoretical or empirical knowledge about objects or methods of research that is a common focus for a set of research processes because it provides a reference for the decisions of researchers – the formulation of problems, the selection of methods or objects, the organisation of empirical data, or the interpretation of data (on the social ordering of research by knowledge see Gläser 2006). This definition resonates with Whitley's (1974) description of research areas but abandons the assumption that topics form a hierarchy. It only demands that some scientific knowledge is perceived similarly by researchers and influences their decisions.

This weak definition is linked to three properties of topics that create the problems for bibliometrics:

1) The fractal nature of knowledge has been described by van Raan (1991) and Katz (1999). Topics can have any 'size' (however measured) between the smallest (emerging topics that just concern one researcher) and very large thematic structures (fields or even themes cutting across several fields). Methods for topic identification should thus not be biased against any particular topic size.

2) Given the multiple objects of knowledge that can serve as common reference for researchers, topics inevitably overlap. Publications commonly contain several knowledge claims, which are likely to address different topics (Cozzens, 1985; Amsterdamska & Leydesdorff, 1989). Methods for topic identification should thus take into account that bibliometric objects (publications, authors, journals, and cited sources) are likely to belong to several topics simultaneously. Methods also should enable the reconstruction of topics that overlap pervasively (i.e. not only in their boundaries).

3) All topics emerge from coinciding autonomous interpretations and uses of knowledge by researchers (see e.g. the case studies discussed by Edge and Mulkay, 1976, pp. 350-402). While individual researchers may launch topics and advocate them, the latter's content and fate depends on the ways in which they are used by others. From this follows that topics are local in the sense that they are primarily topics to the researchers whose decisions are influenced by and who contribute to them. Methods for topic identification can reconstruct this insider perspective by using local information. Global approaches create different representations of topics by finding a compromise between insider perspectives and all outsider perspectives on topics.

Methods

For a detailed description of the method see Havemann, Gläser, & Heinz (2015). We operationalise 'topic' as a set of thematically related papers but cluster citation links instead of papers because the former can be assumed the thematically most homogenous bibliometric objects (see Evans & Lambiotte, 2009; and Ahn, Bagrow & Lehmann, 2010 on link clustering).

Cost Function: We followed the suggestion by Evans and Lambiotte (2009) to obtain link clusters by clustering vertices in a network's line graph and defined a local cost function $\Psi^*(L)$ of link set L in the line-graph approach. The internal degree $k_i^{\text{in}}(L)$ of node i is defined as the number of links in L attached to i. The external degree of a node is obtained by

subtracting the internal from the total degree: $k_i^{out}(L) = k_i - k_i^{in}(L)$. External degrees k_i^{out} are weighted with subgraph membership-grade k_i^{in}/k_i of boundary node i to obtain a measure of external connectivity of link set L:

$$\sigma(L) = \sum_{i=1}^{n} \frac{k_i^{out}(L) k_i^{in}(L)}{k_i} \quad (1)$$

where n is the number of all nodes. The sum can be restricted to boundary nodes because only for boundary nodes of L is $k_i^{out} k_i^{in} > 0$. A simple size normalization that accounts for the finite size of the network is achieved by adapting the ratio cut suggested by Wei and Cheng (1989) for link communities, which leads us to the cost function *ratio node-cut* $\Psi^*(L)$:

$$\Psi^*(L) = \frac{\sigma(L)}{k_{in}(L)(1 - \frac{k_{in}(L)}{2m})} \quad (2)$$

where m is the number of all links and $k_{in}(L)$ is the sum of all internal degrees $k_i^{in}(L)$. $\Psi^*(L)$ essentially relates external to total connectivity of link set L. It can be used to identify link communities (sets of links that are well connected internally and well separated from the rest of the graph) by finding local minima in the cost landscape.

Since the cost landscape is often very rough—has many local minima that sometimes correspond to very similar subgraphs—the resolution of the algorithm must be defined by setting a minimum distance (number of links that differ) between subgraphs corresponding to different local minima. We define the range of a community as the environment in which no subgraph exists that has a lower Ψ^* value. For our experiments with the citation network of astrophysical papers we set a community's minimum range at one third of its size.

Algorithm: The cost function Ψ^* is used in a clustering algorithm that grows communities from seeds. This approach fulfils two more principles derived from our definition of a topic. The independent construction of each community prevents a size bias of the algorithm and enables pervasive overlaps.

```
choose a connected subgraph as a seed
initialize population P by mutating the seed with high variance several times and adapt mutants
while the best community is not too old do
    mutate the best community with low variance and adapt the mutants
    if a mutant is new and its cost is lower than highest cost then
        add it to population P
    end if
    cross the best community with other communities and adapt the offspring
    if offspring is new and its cost is lower than highest cost then
        add it to population P
    end if
    select the best individuals so that the population size remains constant
    if there is no better best community for some generations and innovation rate is low then
        renew the population (mutate the best community with high variance and adapt it)
        select the best individuals so that the population size remains constant
    end if
end while
```

Figure 1. Pseudocode of memetic evolution.

The task of finding communities in large networks is always very complex and requires the use of heuristics. We chose a memetic algorithm that accelerates the search by combining non-deterministic evolution with a deterministic local search in the cost landscape (Neri,

Cotta, & Moscato, 2012). In our algorithm, populations of subgraphs evolve because after a random initialization of a population of some definite size, the genetic operators of crossover, mutation, and selection are repeatedly applied (Fig. 1). Each crossover and mutation is followed by a local search.

Data

The algorithm is applied to the citation network of 14,954 papers published 2010 in 53 journals listed in the category Astronomy & Astrophysics of the Journal Citation Reports 2010 (the journal *Space Weather* with 45 articles was accidentally left out). We downloaded all articles, letters and proceedings papers from the Web of Science. Reference data had to be standardised with rule-based scripts. To reduce the complexity of the network, we omitted all sources that are cited only once because they do not link papers and their removal should not unduly influence clustering. We excluded 184 papers that are not linked to the giant component of the citation network and proceeded with a network of 119,954 nodes that are connected by 536,020 citation links. We neglected the direction of citation links and analysed an undirected unweighted connected graph.

Experiments and Preliminary Results

Constructing the seed population

Since topics can assume all possible sizes, the algorithm should start from differently sized seed graphs. In our experiments, we combined two strategies for obtaining seeds. First, we used Ward clustering with a similarity measure derived from theoretical considerations (Gläser, Heinz & Havemann, 2015). We ordered all hard clusters by their stability (the length of their branch in the dendrogram) and selected the most stable but not too large clusters (a total of 63) as seeds. In addition, we used the citation links of 969 randomly selected papers as seed graphs.

Each seed was first adapted by a local search and then used to initialise the population of 16 different communities by mutating the seed with a variance of 15%.

Owing to the randomness of the evolutionary mechanisms the choice of seed graphs is unlikely to affect the clustering results. However, it is likely to effect the efficiency of the algorithm.

Running the memetic algorithm

Up to ten experiments were run with each seed. The standard mutation variance in each experiment was 5%, i.e. up to 5% of the nodes were randomly exchanged. The variance was increased to 15% for one mutation if Ψ^* values did not improve for 10 generations. Again, we assume these parameters to effect the algorithm's efficiency rather than its outcomes.

Table 1. Examples of experiments with the memetic algorithm.

Community	*Seed sub-graph*		*Number of generations*	*Community*		*Remaining nodes from seed*
	Size	*Ψ* value*		*Size*	*Ψ* value*	
1	13,469	.0692	339	10,586	.0339	10,380
2	19,697	.1174	233	35,159	.0397	18,860
3	35	.4075	232	33	.0047	0
4	76	.5498	203	28	.0975	0

Experiments with the seeds described above resulted in a total of 3,944 distinct communities, 1,375 of which were disregarded because there were better communities within a distance of

less than one third of their size. The remaining 2,569 communities were ordered by increasing Ψ^* values. Table 1 provides exemplary descriptions of some of the experiments. We then calculated the relative coverage of the network as a function of Ψ^* by successively uniting the *L*-sets of the ranked communities. Relative coverage is the ratio of the union's size to the number of all links *m* (Fig. 2). This function has a sharp bend at Ψ^*=0.10458, shortly below maximum coverage. We used this Ψ^* value as cutoff point, which gives us a preliminary result of 154 communities that cover 98.9 % of all links.

Currently, each of these 154 best communities is used as a seed for a refined local search that adds or removes single links instead of nodes with all their links. For some of the 154 communities this additional local search has already led to better communities.

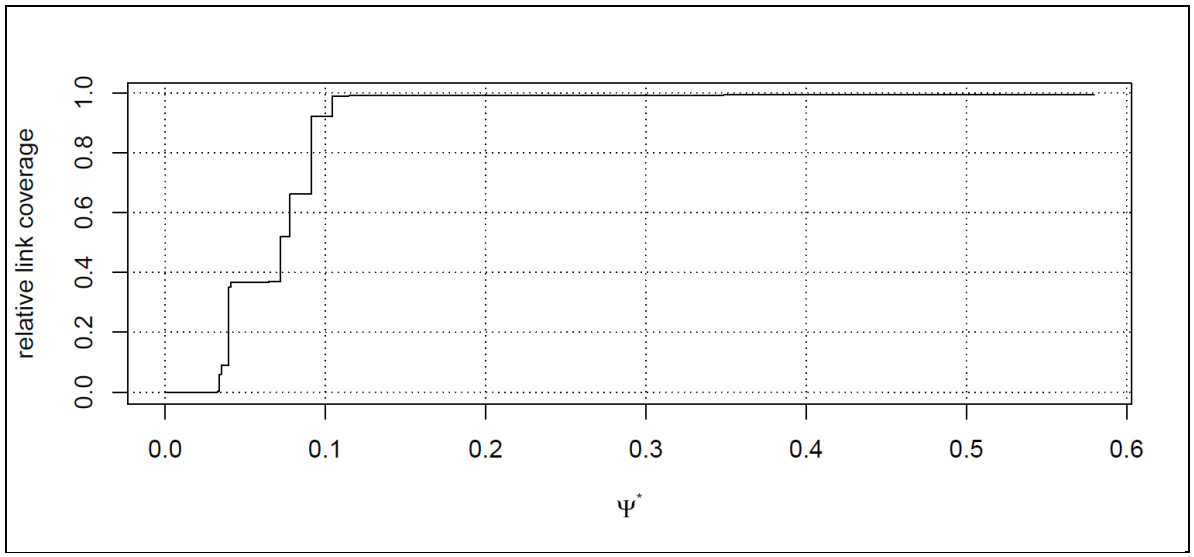


Figure 2. Relative coverage of the network by communities as a function of a Ψ* threshold.

Preliminary results

The 154 communities vary in their size between 9 and 49,324 nodes. Some of the communities overlap pervasively. Seventy communities were not a subset of any other community. The other 84 communities were subsets of one (12 communities) to 28 other communities (1). In Figure 3 we plot sizes and cost of the 154 best communities. Blue circles represent communities that are subsets of others. Green circles represent communities that overlap with another community in 95% of their nodes. All other communities are represented by red circles. The numbers in four circles refer to the communities described in Table 1.

The communities form a poly-hierarchy because some smaller communities are subsets of two larger communities that have no hierarchical subset relation. A community can also have a rest of nodes which are not members of any of its sub-communities.

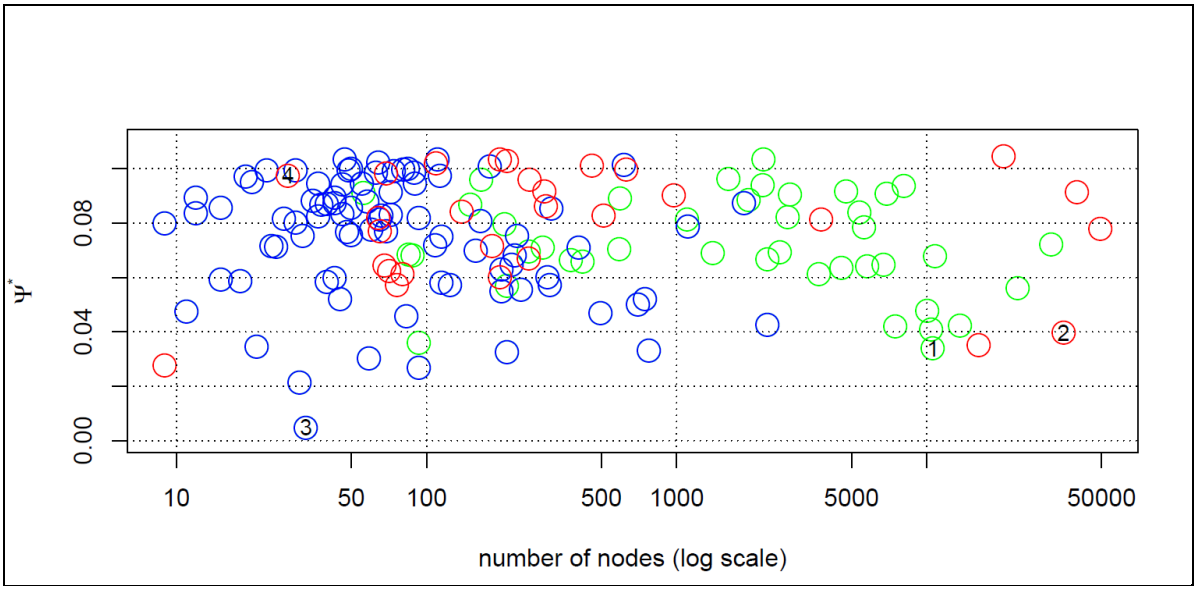


Figure 3. Sizes and Ψ* values of a set of communities covering 98.9% of the graph.

Conclusions

The communities have the structural properties of topics that were derived from the definition. Comparisons with other cluster solutions and tagging of communities will show whether the communities are consistent. We will test the dependence of results on parameter and seed choice with a smaller network. Ultimately, only a discussion with experts can show whether the communities obtained provide one of the possible scientifically meaningful cluster solutions of the astronomy and astrophysics dataset.

Acknowledgments

The work published here was funded by the German Research Ministry (01UZ0905). We thank Andreas Prescher for programming a fast C++-based R-package for parallel node-wise memetic search in the Ψ^* - landscape.

References

Ahn, Y., Bagrow, J. & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, *466*(7307), 761-764.

Amsterdamska, O. & Leydesdorff, L. (1989). Citations: Indicators of Significance? *Scientometrics*, 15, 449-471.

Cozzens, S. E. (1985). Comparing the Sciences: Citation Context Analysis of Papers from Neuropharmacology and the Sociology of Science. *Social Studies of Science, 15*(1), 127-153.

Edge, D. & Mulkay, M. J. (1976). *Astronomy Transformed: The Emergence of Radio Astronomy in Britain*. New York: John Wiley & Sons, Inc.

Evans, T. & Lambiotte, R. (2009). Line graphs, link partitions, and overlapping communities. *Physical Review E*, *80*(1), 16105.

Gläser, J. (2006). *Wissenschaftliche Produktionsgemeinschaften. Die soziale Ordnung der Forschung*. Frankfurt a. M.: Campus.

Gläser, J., Heinz, M. & Havemann, F. (2015). Measuring the diversity of research. Paper submitted to the 15th International Conference on Scientometrics and Informetrics, Istanbul, 29 June -4 July 2015.

Havemann, F., Gläser, J. & Heinz, M. (2015). Detecting Overlapping Link Communities by Finding Local Minima of a Cost Function with a Memetic Algorithm. Part 1: Problem and Method. arXiv:1501.05139.

Hric, D., Darst, R. K. & Fortunato, S. (2014). Community detection in networks: Structural communities versus ground truth. *Physical Review E*, *90*, 062805.

Katz, J. S. (1999). The self-similar science system. *Research Policy*, *28*, 501-517.

Neri, F., Cotta, C. & Moscato, P. (Eds.) (2012). *Handbook of Memetic Algorithms*, Volume 379 of Studies in Computational Intelligence. Berlin: Springer.

Van Raan, A. F. J. (1991). Fractal Geometry of Information Space as Represented by Co-Citation-Clustering. *Scientometrics*, *20*, 439-449.

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, *36*, 397-420.

Wei, Y.-C. & Cheng, C.-K. (1989). Towards efficient hierarchical designs by ratio cut partitioning. In *IEEE International Conference on Computer-Aided Design, 1989. ICCAD-89. Digest of Technical Papers*, pp. 298–301.

Whitley, R. D. (1974). Cognitive and social institutionalization of scientific specialties and research areas. In: R. Whitley (Ed.), *Social Processes of Scientific Development* (pp. 69-95), London: Routledge & Kegan Paul.

Re-citation Analysis: A Promising Method for Improving Citation Analysis for Research Evaluation, Knowledge Network Analysis, Knowledge Representation and Information Retrieval

Dangzhi Zhao[1] and Andreas Strotmann[2]

[1] *dzhao@ualberta.ca*
School of Library and Information Studies, University of Alberta, Edmonton (Canada)

[2] *andreas.strotmann@gmail.com*
ScienceXplore, F.-G.-Keller-Str. 10, D-01814 Bad Schandau (Germany)

Abstract
Citation analysis is used in research evaluation exercises around the globe, directly affecting the lives of millions of researchers and the expenditure of billions of dollars. It is therefore crucial to seriously address the problems and limitations that plague it. Central amongst critiques of the common practice of citation analysis has long been that it treats all citations equally, be they crucial to the citing paper or perfunctory. Weighting citations by their value to the citing paper has long been proposed as a theoretically promising solution to this problem. *Re-citation analysis* proposes to tune out the large percentage of perfunctory citations in a paper and tune in on crucial ones when performing citation analysis, by ignoring uni-citations (mentioned just once in a paper) and counting and analyzing only re-citations (used again and again in a citing paper). By focusing on core connections in knowledge networks, re-citation analysis can help research evaluation become more sensitive to the distinction between essential and perfunctory impact of research. It may benefit citation-link based knowledge representation and retrieval systems with improved precision by better capturing "aboutness" of articles, the essence of subject indexing in knowledge representation and retrieval, rather than merely providing "relatedness" information.

Conference Topic
Theory; Methods and techniques

Introduction
Citation analysis is used in research evaluation exercises around the globe, directly affecting the work and lives of millions of researchers and the expenditure of billions of dollars. It is therefore crucial to seriously address the problems and limitations that plague it. Central amongst critiques of the current practices of citation analysis has long been that it treats all citations equally, be they crucial to the citing paper or perfunctory. This problem is especially serious when tracing or assessing research impact.

Weighting citations by how they are used in the citing paper has therefore long been proposed as a theoretically promising solution to this problem, but in practice it has not been studied closely at a large scale until recently. Increasingly available digital full-text documents and advances in text processing technologies are now making it feasible to conduct large-scale studies on citation counting weighted by in-text citation frequency, location or context. As a result, interest in this type of studies is growing.

Re-citation analysis as defined here may be viewed as a large sub-class of the class of in-text frequency weighted citation analysis schemes, a class which has recently been found to be the most effective one among many features of in-text citations at characterizing essential citations (Zhu, Turney, Lemire, & Vellino, 2014). We discuss in this paper why we consider re-citation analysis a promising method for improving citation analysis for research evaluation, knowledge network analysis, knowledge representation and information retrieval.

Weighted Citation Counting

Citation analysis examines citation patterns and networks in the scholarly literature through statistical analysis and network visualization. It is applied widely in the social sciences to trace knowledge flows, to evaluate research impact, to study the characteristics of scholarly communities and knowledge networks, and to create citation link based knowledge representation and retrieval systems (Borgman & Furner, 2002; Hall, Jaffe, & Trajtenberg, 2005).

The basic assumption underlying citation analysis is that a citation represents the citing author's use of the cited work, and that it therefore indicates that the citing and cited works are related in subject matter or methodological approach (Garfield, 1979; White, 1990). The total number of citations that a document or any aggregate of documents (e.g., author oeuvre, journal) receives (or a score derived from it, e.g., h-index) is therefore used to assess its impact on research in research evaluation. Citation links are used to signify knowledge flow from the cited to the citing group and, along with scores derived from these links, to measure the relatedness between documents or their aggregates in the study of knowledge networks and in the representation and retrieval of related documents.

The assumptions of citation analysis are believed to be in line with Merton's normative view of science (Garfield, 1979; Merton, 1942; White, 1990). Like other activities of science, citation behaviour is assumed to be governed by a set of norms which require authors to cite documents that have influenced them in developing their current works in order to give credit where credit is due (Edge, 1979; Griffith, 1990; Peritz, 1992; Tranöy, 1980). Although citations for reasons other than giving due credit do exist (Cronin, 1984; Edge, 1979), citation analysis has generally been found to produce valid results because it is based on a statistical analysis of the collective perceptions of large numbers of citing authors, most of whom do adhere to the norms most of the time (Small, 1977; White, 1990). This is especially true with citation network analysis and citation link based knowledge representation and retrieval, as even non-normative citations will not refer to unrelated works.

Researchers do cite for various reasons and citations do serve many different functions in citing papers, however (Brooks, 1985, 1986; Case & Higgins, 2000; Chubin & Moitra, 1975; Liu, 1993; Moravcsik & Murugesan, 1975; Shadish, Tolliver, Gray & Sengupta, 1995; Vinkler, 1987). Small (1982), for example, identified five typical distinctions in citation classification schemes: (1) negative or refuted, (2) perfunctory or noted only, (3) compared or reviewed, (4) used or applied, and (5) substantiated or supported by the citing work.

The importance of weighing citations by their role in the text has therefore long been recognized (Herlach, 1978; Narin, 1976). In recent years, with increasingly available digital full-text documents and advances in technologies for text processing, interest in studying weighted citations has finally picked up. Studies have experimented with weighing citations by the frequency with which they are referred to in the text (e.g., Ding, Liu, Guo, & Cronin, 2013; Hou, Li, & Niu, 2011; Zhu, Turney, Lemire, & Vellino, 2014), by the citation impact of citing papers (Ding & Cronin, 2011), or by the location and context in which they are cited (Boyack, Small, & Klavans, 2013; Jeong, Song, & Ding, 2014). It has been found that frequency-weighted citation ranking can outperform traditional citation ranking of top authors, and that in-text citation frequency was the best of many other full-text features to help spot citations that were considered crucial to the citing papers by their authors, at least in a hard science field studied (Zhu, Turney, Lemire, & Vellino, 2014).

Depending on what functions they serve in a given citing paper, citations likely appear more or less frequently there: perfunctory ones once only, negative or contrastive ones a couple of times, and used or substantiated ones many times. By weighing citations by their frequency of appearance in a scholarly paper, it is hoped that essential citations could be assigned greater weight than perfunctory ones so that citation analysis can focus on the more profound

influences and on organic relationships. If so, this could improve traditional citation analysis significantly as a high incidence of perfunctory citations has been observed (Small, 1982). For example, Teufel, Siddharthan, & Tidhar (2006) found that only a fifth of the references are essential for the citing papers, and Moravcsik & Murugesan (1975) noted that 40% references were perfunctory, frequently simply copied from other papers without ever having been read (Dubin, 2004).

Re-citation analysis: motivation and innovation

Perfunctory citations can thus be considered a serious source of noise if the signal that one wants to detect is the direct and substantial flow of knowledge in the literature. There are two obvious types of approaches to dealing with this problem: (1) to amplify the signal or (2) to filter out the noise. The ultimately best approach is likely some combination of the two. All frequency-based weighing schemes studied so far used the former approach by assigning a weight based on the in-text citation frequency such as assigning a weight of N or N^2 to a citation that appears N times in a citing paper.

By contrast, re-citation analysis, a concept we introduced recently (Zhao & Strotmann, 2015), uses the latter approach: it attempts to filter out perfunctory citations from the analysis by removing uni-citations (i.e., documents referenced only once in the text of a work) in order to analyze only re-citations (i.e., references that appear more than once in the text of a citing paper). The degree to which a cited work is used or has impacted research can be further differentiated by assigning weights to different re-citation frequencies. Re-citation analysis can thus combine the noise filtering and signal amplification approaches, offering the potential to find an optimal weighing scheme for in-text citation frequency.

Thus, the fundamental difference between re-citation analysis and all other frequency-based weighing schemes and hence the innovation of re-citation analysis is that the former attempts to make the fundamental qualitative distinction between those citations that represent real use by, or core impact on, the citing paper (which it tends to retain for analysis) and those that are merely mentioned in passing as related work that the author is aware of but did not directly rely on (which it tends to remove). The basic assumption of re-citation analysis is that papers are very likely to be cited again and again in a publication that relies heavily on them, while perfunctory citations should appear once only in a citing paper almost by definition.

Re-citation analysis can also avoid potential technical problems associated with simply amplifying multi-citations. Since the noise created by perfunctory citations is very strong (40% or more), the signal amplification required to counter it tends to be so strong that it can cause serious distortions. For example, Zhao & Strotmann (2015) found that a simple weight of N does not suffice to make non-perfunctory citations stand out. N^2 is the minimal power of N that fulfills this requirement, but tends to be seriously affected by ultra-meticulous in-text citing styles of a few authors as it overweighs high in-text frequencies. Weighing re-citations avoids this problem.

Promises of Re-citation Analysis

Re-citation analysis can be expected to contribute significantly to the theory and methods of citation analysis. It addresses head-on an old and fundamental concern with citation analysis, especially with evaluative citation analysis. By proposing to filter out the strong noise caused by a high incidence of perfunctory citations rather than simply amplifying multi-citations, it also opens up a new way of thinking about weighing citations at a time when the study of weighted citation counting based on full-text analysis is still in its infancy.

Re-citation analysis is promising in improving citation analysis for research evaluation, knowledge network analysis, knowledge representation and information retrieval.

- Evaluative citation analysis ranks authors, journals, institutions or other components of the scholarly communication system by their citation counts or by derivative scores such as the h-index. Scores based on re-citation counting can be expected to boost those researchers or groupings whose publications receive close scrutiny and to introduce a bias against those whose work mainly provides convenient background information. Such re-citation metrics should thus be better at measuring research impact than traditional citation metrics.
- In citation-based knowledge network analysis and visualization, results based on re-citations can be expected to be significantly more detailed and "crisp" than those based on citations since re-citation based relations (e.g., direct re-citation, co-recitation, or re-citation coupling) should represent core relationships where citation-based relations include many peripheral ones. The price might be an underestimation of interrelatedness between distant parts of a science map.
- For information retrieval (IR), re-citation based similarity metrics can likely provide a considerably enhanced precision of the "Similar documents" or "More like this" feature that many IR systems provide nowadays, compared to citation-based ones. The latter can be expected to show better recall, however, so that a (weighted) combination of the two may work better than either one alone.
- For knowledge representation, it is well understood that citations in scholarly publications serve as concept symbols (Small, 1978). One would expect the presence of a certain set of citations in a paper to translate fairly straightforwardly to the assignment of that paper to a specific subject category. However, subject categories are meant to capture the paper's "aboutness", but a large percentage of citations merely provide "relatedness" information. We suspect that re-citations, on the other hand, do correspond to a considerable degree to concept symbols with an "aboutness" semantics. A re-citation based form of computer-aided subject indexing might therefore be feasible.

Re-citation analysis may thus have a profound impact on the future of the scholarly communication system and of Scientometrics as re-citation analysis values and thus encourages research that is worth following in depth, whereas traditional citation analysis has encouraged review publications that tend to be cited widely.

Finally, as they rely on access to the full text of scholarly publications rather than on citation databases such as Web of Science and Scopus, re-citation analysis methods and metrics are as easily available to the study and evaluation of the social sciences and humanities as to that of the natural and life sciences. Unlike the latter, the former have never been treated fairly by traditional citation analysis due to the insufficient coverage of their literature by these databases.

References

Borgman, C.L. & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology, 36*, 3-72.

Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co‐citation clustering using full text. *Journal of the American Society for Information Science and Technology, 64*(9), 1759-1767.

Brooks, T. A. (1985). Private acts and public objects: an investigation of citer motivations. *Journal of the American Society for Information Science, 36*(4), 223-229.

Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science, 37*(1), 34-36.

Case, D. O. & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science, 51*(7), 635-645.

Cronin, B. (1984). *The Citation Process. The Role and Significance of Citations in Scientific Communication*. London: Taylor Graham.

Chubin, D. E. & Moitra, S. D. (1975). Content analysis of references: adjunct or alternative to citation counting? *Social Studies of Science, 5*(4), 423-441.

Ding, Y. & Cronin, B. (2011). Popular and/or prestigious? Measures of scholarly esteem. *Information Processing and Management, 47*, 80–96.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics, 7*(3), 583-592.

Dubin, D. (2004). The Most Influential Paper Gerard Salton Never Wrote. *Library trends, 52*(4), 748-764.

Edge, D. (1979). Quantitative measures of communication in science: A critical review. *History of Science Cambridge, 17*(36), 102-134.

Garfield, E. (1979). *Citation indexing – Its Theory and Application in Science, Technology, and Humanities.* New York: John Wiley & Sons.

Griffith, B. C. (1990). Understanding science: Studies of communication and information. In C. L. Borgman (ed.). *Scholarly Communication and Bibliometrics*, 33-45. Newbury Park, CA: Sage Publications, Inc.

Hall, B.H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics, 36* (1), 16–38.

Herlach, G. (1978). Can retrieval of information from citation indexes be simplified? Multiple mention of a reference as a characteristic of the link between cited and citing article. *Journal of the American Society for Information Science, 29*(6), 308-310.

Hou, W., Li, M., & Niu, D. (2011). Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution. *BioEssays, 33*, 724-727.

Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics, 8*(1), 197-211.

Liu, M. (1993). The complexities of citation practice: A review of citation studies. *Journal of Documentation, 49*, 370–408.

Merton, R. K. (1942). Science and technology in a democratic order. *Journal of Legal and Political Sociology, 1*, 115-126.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science, 5*(1), 86-92.

Narin, F. (1976). *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Washington, D. C.: Computer Horizons.

Peritz, B. C. (1992). On the objectives of citation analysis: Problems of theory and method. *Journal of the American Society for Information Science, 43*(6), 448-451.

Shadish, W. R., Tolliver, D., Gray, M., & Gupta, S. K. S. (1995). Author judgements about works they cite: three studies from psychology journals. *Social Studies of Science, 25*(3), 477-498.

Small, H. (1977). A co-citation model of a scientific specialty: A longitudinal study of collagen research. *Social Studies of Science, 7*(2), 139-166.

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*(3), 327-340.

Small, H. (1982). Citation context analysis. *In* B. J. Dervin & M. J. Voigt (eds.)*, Progress in Communication Sciences, 3* (pp. 287-310). Norwood, NJ: Ablex.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function*. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (pp. 103-110)*. Stroudsburg, PA, USA.

Tranöy, K. E. (1980). Norms of inquiry: Rationality, consistency requirements and normative conflict. In *Rationality in Science* (pp. 191-202). Springer Netherlands.

Vinkler, P. (1987). A quasi-quantitative citation model. *Scientometrics, 12*(1), 47-72.

White, H. D. (1990). Author co-citation analysis: Overview and defense. *In* C. L. Borgman (ed.), *Scholarly Communication and Bibliometrics* (pp. 84-106). Newbury Park, CA: Sage.

Zhao, D. & Strotmann, A. (2015). Dimensions and uncertainties of author citation rankings: Lessons learned from frequency-weighted in-text citation counting. *Journal of the Association for Information Science and Technology*, doi: 10.1002/asi.23418.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2014). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*. Early view (DOI: 10.1002/asi.23179).

Topic Affinity Analysis for an Astronomy and Astrophysics Data Set

[1]Theresa Velden, Shiyan Yan, and Carl Lagoze

tvelden@umich.edu, shiyansi@umich.edu, clagoze@umich.edu
[1]School of Information, University of Michigan, 105 S. State Street, Ann Arbor, MI 48109 (USA)

Abstract
In this paper we map the affinity between topics extracted from a body of literature published in Astronomy and Astrophysics journals between 2003-2010. The topics are extracted using the popular information theoretic Infomap clustering algorithm (Rosvall & Bergstrom, 2008) iteratively on the giant component of the direct citation network constructed from the data. The affinity network shows what topics are disproportionally well connected (by citations) to other topics. The topology of the network highlights a large division into astrophysics versus astronomically oriented publications. Bridging between those two domains is a population of smaller topics. Going forward, we plan to create and analyze topic affinity network maps for alternative solutions to the topic extraction challenge on that same data set that are produced by our colleagues and that will be discussed and compared at the proposed special session on 'Same data? Different results? The performative nature of algorithms for topic detection in science' at ISSI 2015. We expect that topic affinity mappings will help to examine the nature of differences between different topic extraction solutions.

Conference Topic
Methods and techniques (special session on algorithms for topic detection)

Introduction
The mapping of research topics and collaborative ties in scientific research fields (Morris 2008) is flourishing for a number of reasons. Increasingly, scholarly publications and their metadata are available from a variety of sources (digital libraries, institutional and disciplinary repositories, along with bibliographic abstracting services such as the long established Web of Knowledge and more recently, Scopus). Complementing this is the emergence of sophisticated algorithms for the analysis of complex networks (Newman 2003b) and the wide availability of advanced user-friendly network analysis and visualization tools like pajek, gephi, or VOS Viewer.

However, many different algorithms for community extraction and topic detection exist and offer different suggestions what the most prominent groupings of publications or authors may be. The special session at ISSI 2015 sets out to systematically compare and evaluate the origin, extent, and implication of differences between topic extraction methods. In this paper we describe the results of our approach to topic detection and topic affinity analysis to the shared 'astronomy and astrophysics' data set. This approach has emerged from research program on studying behavioral patterns in scientific communities and comparing them across fields, and may help to shed light on the nature of differences between topic extraction solutions.

Background
As described in (Velden 2009), we take a mixed method approach to studying field-specific practices and cultures of scientific communities, integrating ethnographic field studies with network analytic methods. The network analytic method we apply here to the 'astronomy and astrophysics' data set is part of an ongoing effort to combine network analytic with ethnographic methods (Velden, Haque & Lagoze, 2010; Velden, 2013). This evolves a tradition of close-up analysis of scientific networks and communication practices started by Crane's work (1972) on invisible colleges and taken up more recently by Zuccala (2006).

Scientific research specialties are a complex social and cognitive phenomenon. Sociologically, they can be characterized as collective production communities that emerge from the indirectly coordinated activity of autonomous actors (research groups) who aim to contribute to a shared knowledge base (Gläser, 2006; Velden, 2013). Therefore, the combined analysis of social and cognitive structures is of particular interest (Ding, 2011). In our work we achieve this in two steps: first by algorithmically extracting major research topics in a research specialty from the direct citation network and generating an affinity network that shows what topics are disproportionally well connected through citations to other topics. In a second step, we overlay the topic information on the group collaboration network (Velden, Haque & Lagoze, 2010) extracted from the co-author network of the research specialty. The resulting maps show how collaborative ties connect groups active in a particular topic area. This paper reports work in progress. At this point, we have produced and analyzed the topic affinity network. Producing the overlay with the group collaboration network will be one of the next steps.

Method

Our approach to topic extraction and topic affinity analysis is discussed in detail in Velden (2013). Below we briefly review the relevant details for the analysis reported in this paper.

Data

The data set used in this study includes papers published 2003-2010 in 59 astrophysical journals indexed by Web of Science. By accepting only documents of type 'Article', 'Letter', and 'Proceedings Paper', the data set comprised the bibliographic data of 111,616 publications.

Network construction

Various citation-based approaches have been used in the past to detect topics in research fields. These include bibliographic coupling, co-citation and direct citation, including or excluding citation environments. The advantages and disadvantages of these approaches have been discussed in Boyack (2010). We base our topic extraction on the direct citation network.

Clustering

We use the Infomap clustering algorithm (Rosvall & Bergstrom, 2008) twice to iteratively extract clusters of clusters of documents. The repeated clustering is necessary to obtain sufficiently large entities (topics) for further visual inspection and analysis. In the resulting topic network, nodes represent clusters of publications based on the direct citation links between them.

Topic affinity network

We evaluate the strength of citation links between topic areas relative to a null model that assumes a random distribution of citation links proportional to topic area sizes. Hence, the existence of a link between topics in the affinity indicates a surplus of connectivity between the two topic areas in question, whereas the absence of a link may either mean 'normal' (random) background connectivity or a negative affinity value ('antagonism').

The affinity between a source topic area and a target topic area is calculated as shown in Figure 1 below.

Assume:
A_{11-i}: Top 11 Areas expect area i
$N_{p(j)}$: Number of papers in topic area j
C_{ij}: Number of Citation from topic area i to topic area j

We define the citation based affinity A between two topic areas i and j as the residual:

$$A_{ij} = \frac{\text{Actual Count}_{ij} - \text{Expected Count}_{ij}}{\sqrt{\text{Expected Count}_{ij}}}$$

where:

$$\text{Actual Count}_{ij} = C_{ij}$$

$$\text{Expected Count}_{ij} = \frac{N_{p(j)}}{\sum_{k \in A_{11-i}} N_{p(k)}} \times \left(\sum_{k \in A_{11-i}} C_{ik}\right)$$

Figure 1. Affinity between a source topic area and a target topic

Topic affinity as defined here is a relative property. It expresses the relative preference for documents in one topic area to cite documents in another area given the choice of topic areas included in the data set and in the affinity calculation. Theoretically, the relative affinity to document clusters outside the set of topic areas selected for this analysis or even outside of the data set (external citations) could be greater than to the ones in the set.

Topic Labeling

To support the interpretation of the resulting topic affinity network, we use a semi-automatic approach to labeling topic areas. To this end, we analyze the frequency of journals that the documents in each topic area are published in. Using a measure based on the concept of *term frequency - inverse document frequency (tf-idf)* to combine popularity with distinctiveness of a journal title within the data set, we produce a ranked list of the 15 most popular journals in each topic area. From those journal titles we then derive labels that typically reflect sub disciplinary orientation of topic areas. A more detailed and specific identification of topic area content either algorithmically or through expert evaluation or would be desirable.

Results

The topic extraction from the giant component of the direct citation network results in 22 document clusters ('topics'). For pragmatic reasons, to support interpretation of the visualized network, we include only the largest eleven topic areas in the affinity network. Given the uneven size distribution of clusters (Fig. 1), these largest clusters account for the large majority of publications in the giant component of the direct citation network, namely 84% (see Table 1 for details on the sizes of various network components).

Table 1. Properties of direct citation network.

	# of nodes (documents)	**% of network**	**% of giant component**
entire network	111,616	100	N.A.
giant component	101,831	91.2	100
11 largest topic areas	85,562	84.0	76.7

The topic affinity network for the largest 11 document clusters is shown in Figure 2. The most striking topological feature regards the relationship between the three largest topics. Notably,

topic 3 (Astronomy/Solar System) is not directly connected with the other two topics, topic 1 (Astronomy/Astrophysics) and topic 2 (Gravitational Physics, Cosmology). Topic 2 has a strong directed link to topic 1, indicating that it borrows disproportionally from the literature in topic 2. Topics 1 and 3 are indirectly linked, via small, astronomically oriented 'proxy topics', essentially topics 7 and 9, and to l lesser degree topics 10 and 11. However, there exists only a very faint indirect affinity link between topic 2 and topic 3, via topic 11.

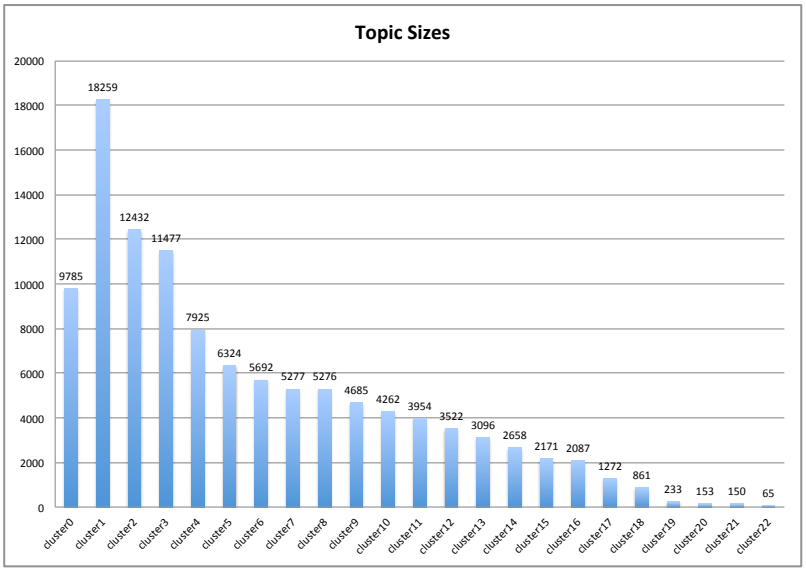


Figure 1. Sizes of the 22 document clusters ('topics') that constitute the giant component of the direct citation network. Cluster '0' shows the number of documents not included in the giant.

Discussion

Based on our own, if limited, expertise in this larger domain of research, we would offer the following speculations about the interpretation of the tripartite structure of the current 2003-2010 literature in the astronomy and astrophysics data set that is suggested by the topology of the affinity network in figure 2. The literature is subdivided into three large domains, with distinct research focus, namely astrophysics - the quest for developing a theoretical understanding of physical and chemical properties of celestial bodies (topic 1), gravitational physics - the quest for understanding the workings of gravitational forces in the universe (topic 2), and planetary science - the quest for understanding the composition, dynamics and history of planets and solar systems (topic 3). As reflected by the affinity network, in the 2003-2010 period, the three domains rely to varying degrees on astronomical observation; this is least the case for gravitational physics. An interesting open question is to what degree the observational astronomy literature has been integrated through citations into these larger topics rather than being identifiable as separate topics. The topic affinity network further underlines that whereas there are strong connections between astrophysics and gravitational physics (such as the role of gravitational forces in the formation of black holes and the puzzle of the nature of black matter), the cognitive links between gravitational physics and planetary science are weak.

Table 2. Ranking of the 15 most popular journals in each topic. This list of journal titles is used to help identify the subject matter of a topic in terms of its subdisciplinary orientation.

Journal titles	# of publications	tf*idf score
Area1		
ASTRONOMICAL JOURNAL	1098	0.104672985
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	4415	0.091614001
ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES	401	0.06435346
ASTRONOMISCHE NACHRICHTEN	314	0.062939775
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF AUSTRALIA	116	0.056289489
NEW ASTRONOMY REVIEWS	347	0.043675217
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF THE PACIFIC	152	0.037652069
ASTRONOMY REPORTS	164	0.032873003
CHINESE JOURNAL OF ASTRONOMY AND ASTROPHYSICS	171	0.027442498
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	284	0.027073887
ASTROPHYSICAL JOURNAL LETTERS	510	0.022086996
PHYSICAL REVIEW D	164	0.020641889
ASTROPHYSICS AND SPACE SCIENCE	290	0.006017681
ASTROPHYSICAL JOURNAL	5565	0
ASTRONOMY & ASTROPHYSICS	3148	0
Area2		
PHYSICAL REVIEW D	5616	0.700439718
JOURNAL OF COSMOLOGY AND ASTROPARTICLE PHYSICS	1416	0.533389555
CLASSICAL AND QUANTUM GRAVITY	1533	0.376292436
GENERAL RELATIVITY AND GRAVITATION	543	0.204541334
INTERNATIONAL JOURNAL OF MODERN PHYSICS D	655	0.081693023
GRAVITATION & COSMOLOGY	75	0.036063565
ASTROPARTICLE PHYSICS	78	0.023617218
NEW ASTRONOMY	46	0.017327627
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	783	0.016100189
NEW ASTRONOMY REVIEWS	122	0.015216105
ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES	49	0.007792228
ASTROPHYSICS AND SPACE SCIENCE	286	0.005880784
ASTROPHYSICAL JOURNAL LETTERS	40	0.001716582
ASTROPHYSICAL JOURNAL	506	0
ASTRONOMY & ASTROPHYSICS	325	0
Area3		
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF THE PACIFIC	364	0.160723328
ICARUS	150	0.129745662
ASTRONOMISCHE NACHRICHTEN	361	0.128983753
ASTRONOMICAL JOURNAL	732	0.124387179
NEW ASTRONOMY	107	0.072503543
ASTROPHYSICS	89	0.060306686
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	1461	0.054039787
ASTRONOMY REPORTS	108	0.038587937
ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES	111	0.031752855
ASTROPHYSICAL JOURNAL LETTERS	318	0.024548551
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	127	0.021580836
NEW ASTRONOMY REVIEWS	85	0.019070268
ASTROPHYSICS AND SPACE SCIENCE	385	0.014240464
ASTROPHYSICAL JOURNAL	2773	0
ASTRONOMY & ASTROPHYSICS	3122	0
Area4		
SOLAR PHYSICS	1248	2.133094119
ANNALES GEOPHYSICAE	228	0.222784668
ADVANCES IN SPACE RESEARCH	372	0.153453831
GEOPHYSICAL AND ASTROPHYSICAL FLUID DYNAMICS	77	0.131609172
ASTRONOMISCHE NACHRICHTEN	187	0.096348379
SPACE SCIENCE REVIEWS	96	0.093804071
ASTRONOMY REPORTS	119	0.061312605
ASTROPHYSICAL JOURNAL LETTERS	333	0.037069606
CHINESE JOURNAL OF ASTRONOMY AND ASTROPHYSICS	77	0.031763293
ASTRONOMY LETTERS-A JOURNAL OF ASTRONOMY AND SPACE ASTROPHYSICS	95	0.03073523
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	102	0.024994227
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	189	0.010080921
ASTROPHYSICS AND SPACE SCIENCE	75	0.004000365
ASTROPHYSICAL JOURNAL	2165	0
ASTRONOMY & ASTROPHYSICS	1609	0
Area5		
ICARUS	2102	2.700439718
PLANETARY AND SPACE SCIENCE	850	1.091995129
ASTROBIOLOGY	258	0.454192886
EARTH MOON AND PLANETS	257	0.330167939
CELESTIAL MECHANICS & DYNAMICAL ASTRONOMY	170	0.299274383
SOLAR SYSTEM RESEARCH	167	0.29399307
SPACE SCIENCE REVIEWS	115	0.115737336
ADVANCES IN SPACE RESEARCH	263	0.111741818
ANNALES GEOPHYSICAE	104	0.104666808
ASTRONOMICAL JOURNAL	231	0.058301094
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	219	0.012031166
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	38	0.009590656
ASTROPHYSICAL JOURNAL LETTERS	72	0.008255273
ASTROPHYSICAL JOURNAL	286	0
ASTRONOMY & ASTROPHYSICS	598	0
Area6		
PHYSICAL REVIEW D	4101	0.700439718
JOURNAL OF COSMOLOGY AND ASTROPARTICLE PHYSICS	353	0.182093016
ASTROPARTICLE PHYSICS	430	0.178295313
CLASSICAL AND QUANTUM GRAVITY	33	0.011092632
ADVANCES IN SPACE RESEARCH	45	0.00979976
ASTROPHYSICS	16	0.008253508
NEW ASTRONOMY REVIEWS	45	0.007685878
INTERNATIONAL JOURNAL OF MODERN PHYSICS D	45	0.007685878
COMPTES RENDUS PHYSIQUE	16	0.006634244

Journal titles	# of publications	tf*idf score
Area 6 (contd)		
ASTRONOMY LETTERS-A JOURNAL OF ASTRONOMY AND SPACE ASTROPHYSICS	15	0.002561959
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	69	0.001942923
ASTROPHYSICAL JOURNAL LETTERS	29	0.001704276
ASTROPHYSICS AND SPACE SCIENCE	25	0.000703958
ASTROPHYSICAL JOURNAL	241	0
ASTRONOMY & ASTROPHYSICS	107	0
Area7		
BALTIC ASTRONOMY	64	0.093118611
REVISTA MEXICANA DE ASTRONOMIA Y ASTROFISICA	39	0.077757085
ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES	131	0.063035619
ASTRONOMICAL JOURNAL	218	0.062312612
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	686	0.042681773
ASTRONOMY REPORTS	65	0.039065743
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF THE PACIFIC	45	0.033422965
SPACE SCIENCE REVIEWS	26	0.029634918
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	90	0.02572539
ASTROPHYSICAL JOURNAL LETTERS	160	0.02077656
ASTRONOMY LETTERS-A JOURNAL OF ASTRONOMY AND SPACE ASTROPHYSICS	36	0.013586115
CHINESE JOURNAL OF ASTRONOMY AND ASTROPHYSICS	23	0.011067322
ASTROPHYSICS AND SPACE SCIENCE	176	0.010950426
ASTROPHYSICAL JOURNAL	1856	0
ASTRONOMY & ASTROPHYSICS	1359	0
Area8		
PHYSICAL REVIEW D	5208	0.700439718
INTERNATIONAL JOURNAL OF MODERN PHYSICS D	31	0.004169284
CLASSICAL AND QUANTUM GRAVITY	8	0.002117529
JOURNAL OF COSMOLOGY AND ASTROPARTICLE PHYSICS	5	0.002030988
GENERAL RELATIVITY AND GRAVITATION	3	0.001218593
ASTROPHYSICS	3	0.001218593
NUOVO CIMENTO DELLA SOCIETA ITALIANA DI FISICA C-GEOPHYSICS AND SPACE PHYSICS	3	0.001218593
COMPTES RENDUS PHYSIQUE	3	0.000979516
ASTROPARTICLE PHYSICS	2	0.000653011
GRAVITATION & COSMOLOGY	1	0.000518518
CHINESE JOURNAL OF ASTRONOMY AND ASTROPHYSICS	3	0.00051445
NEW ASTRONOMY REVIEWS	1	0.000134493
ASTROPHYSICS AND SPACE SCIENCE	2	4.43E-05
ASTRONOMY & ASTROPHYSICS	2	0
ASTROPHYSICAL JOURNAL	1	0
Area9		
ASTRONOMICAL JOURNAL	571	0.282314914
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF AUSTRALIA	86	0.216437853
ACTA ASTRONOMICA	50	0.172434283
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF THE PACIFIC	104	0.133611565
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	909	0.097827391
NEW ASTRONOMY	48	0.094634582
ASTRONOMISCHE NACHRICHTEN	79	0.0821274
ASTRONOMY REPORTS	43	0.044702256
ASTRONOMY LETTERS-A JOURNAL OF ASTRONOMY AND SPACE ASTROPHYSICS	58	0.037861606
ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES	45	0.037454628
ASTROPHYSICAL JOURNAL LETTERS	159	0.035713223
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	42	0.020765721
ASTROPHYSICS AND SPACE SCIENCE	81	0.008717292
ASTROPHYSICAL JOURNAL	1073	0
ASTRONOMY & ASTROPHYSICS	1051	0
Area10		
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	217	0.086427698
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	783	0.067881878
CHINESE JOURNAL OF ASTRONOMY AND ASTROPHYSICS	82	0.054979695
ASTRONOMISCHE NACHRICHTEN	65	0.054433948
ADVANCES IN SPACE RESEARCH	72	0.048274854
ASTRONOMY LETTERS-A JOURNAL OF ASTRONOMY AND SPACE ASTROPHYSICS	64	0.033654761
ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES	49	0.03285372
NEW ASTRONOMY REVIEWS	58	0.030499627
PHYSICAL REVIEW D	49	0.025766927
INTERNATIONAL JOURNAL OF MODERN PHYSICS D	49	0.025766927
ASTROPHYSICAL JOURNAL LETTERS	135	0.024426497
ASTRONOMICAL JOURNAL	50	0.019914216
ASTROPHYSICS AND SPACE SCIENCE	106	0.009189628
ASTROPHYSICAL JOURNAL	1332	0
ASTRONOMY & ASTROPHYSICS	897	0
Area11		
NUOVO CIMENTO DELLA SOCIETA ITALIANA DI FISICA C-GEOPHYSICS AND SPACE PHYSICS	105	0.152244762
PHYSICAL REVIEW D	117	0.0561696
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF THE PACIFIC	59	0.055745158
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	596	0.047172325
CHINESE JOURNAL OF ASTRONOMY AND ASTROPHYSICS	70	0.042848482
ASTRONOMICAL JOURNAL	88	0.031998146
ASTRONOMY LETTERS-A JOURNAL OF ASTRONOMY AND SPACE ASTROPHYSICS	58	0.027844759
INTERNATIONAL JOURNAL OF MODERN PHYSICS D	56	0.026884595
ASTROPHYSICAL JOURNAL LETTERS	162	0.026760324
ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES	43	0.026321211
NEW ASTRONOMY REVIEWS	51	0.024484185
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	49	0.017817149
ASTROPHYSICS AND SPACE SCIENCE	115	0.009102042
ASTROPHYSICAL JOURNAL	1459	0
ASTRONOMY & ASTROPHYSICS	589	0

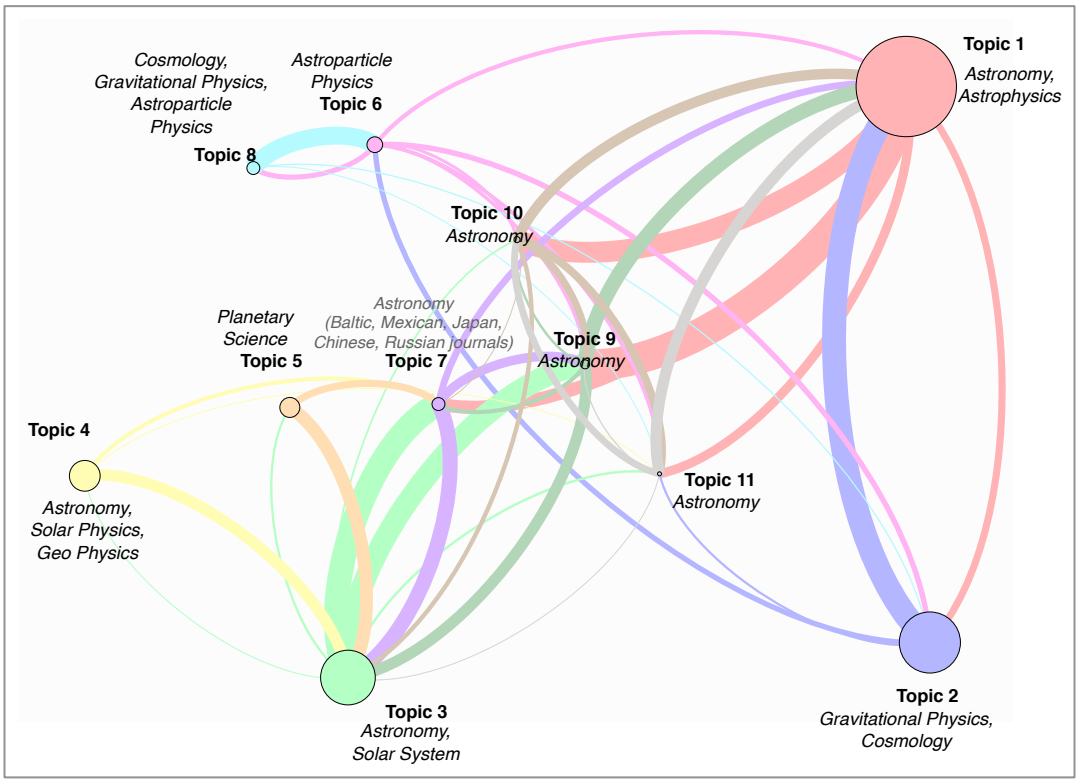


Figure 2: Topic affinity network. Node size indicates number of documents. Link strength indicates relative preference given by publications in one topic to cite publications in another. Links are directed: they are colored by their source node and curve clockwise away from it.

To further validate these hypotheses, a review of the topic contents and interpretation of the topic affinity links by experts could be insightful. Further, an extension of the data set backward in time to show the temporal evolution of affinity links could be informative. This would allow matching the evolution of affinity links over time to reports by experts about major research developments in this domain that may affect the interlinking between topics. One challenge in such an undertaking is that not just the linkages between topics evolve over time, but so does the identity of topics itself.

Conclusions

The topology of the affinity network highlights cognitive links between the topics extracted by our method from the astronomy and astrophysics data set. The interesting question in the context of the special session on the comparison of topic extraction algorithms will be what other cognitive features of this literature will be highlighted, if the affinity network is constructed for alternative groupings of documents into topics produced by other topic extraction algorithms. We suggest that this method of investigating the nature of differences between alternative topic extraction results is useful, in particular for cases where the topic size distribution is such that the large majority of documents, 80-90% is concentrated in 10-30 topics. For more granular topic extraction results the affinity network visualization is likely to become too unwieldy to interpret.

Acknowledgements

We gratefully acknowledge funding from SMA 1258891 EAGER: Collaborative Research: Scientific Collaboration in Time, as well as a travel grant by the intergovernmental framework for European Cooperation in Science and Technology (COST, Action: TD1210).

References

Boyack, K. W. & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404.

Crane, D. (1972). *Invisible Colleges - Diffusion of Knowledge in Scientific Communities*. The University of Chicago Press.

Ding, Y. (2011). Community detection: Topological vs. topical. *Journal of Informetrics*, *5*(4), 498–514.

Gläser, J. (2006). *Wissenschaftliche Produktionsgemeinschaften - die soziale Ordnung der Forschung*, Volume 906 of Campus Forschung. Frankfurt: Campus Verlag.

Morris, S. & Van der Veer Martens, B. (2008). Mapping research specialties. *Annual Review of Information Science and Technology*, *42*(1), 213–295.

Newman, M. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167–256.

Rosvall, M. & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123.

Velden, T. (2013). Explaining field differences in openness and sharing in scientific communities. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 445–458. ACM.

Velden, T. & Lagoze, C. (2013). The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication. *Journal of the American Society for Information Science and Technology*, *64*(12), 2405–2427.

Velden, T. & Lagoze, C. (2009). Patterns of collaboration in co-authorship networks in chemistry - mesoscopic analysis and interpretation. In Larsen, B. & Leta, J. (eds.), *Proceedings of ISSI 2009 - the 12th International Conference of the International Society for Scientometrics and Informetrics, Rio de Janeiro, Brazil, July 14-17, 2009* (2 volumes).

Velden, T., Haque, A. & Lagoze, C. (2010). A new approach to analyzing patterns of collaboration in co-authorship networks: mesoscopic analysis and interpretation. *Scientometrics*, *85*(1), 219–242.

Zuccala, A. (2006) Modeling the invisible college. *Journal of the American Society for Information Science and Technology*, *57*(2), 152 – 168.

Time & Citation Networks

James R. Clough and Tim S. Evans

{james.clough09, t.evans} @ imperial.ac.uk
Imperial College London, Centre for Complexity Science, South Kensington Campus, London SW7 2AZ (U.K.)

Abstract
Citation networks emerge from a number of different social systems, such as academia (from published papers), business (through patents) and law (through legal judgements). A citation represents a transfer of information, and so studying the structure of the citation network will help us understand how knowledge is passed on. What distinguishes citation networks from other networks is time; documents can only cite older documents. We propose that existing network measures do not take account of the strong constraint imposed by time. We will illustrate our approach with two types of causally aware analysis. We apply our methods to the citation networks formed by academic papers on the arXiv, to US patents and to US Supreme Court judgements. We show that our tools can reveal that citation networks which appear to have very similar structure by standard network measures, turn out to have significantly different properties. We interpret our results as indicating that many papers in a bibliography were not directly relevant to the work and that we can provide a simple indicator of the important citations. We suggest our methods may highlight papers which are of more interest for interdisciplinary research. We also quantify differences in the diversity of research directions of different fields.

Background
Bibliometrics has a long tradition of dealing with citation networks from a network point of view as Price's model (Price, 1965) shows. The recent explosion of interest in network analysis in other fields has led to development of existing methods and introduced many new techniques. However most network methods assume static graphs where time plays no explicit role even if the underlying data is almost always evolving. Time can be incorporated into a network representation in two main ways. If we assign a single time to each edge we have a *Temporal Edge Network*. Such networks have received considerable attention (Holme & Saramäki, 2012). For instance they form a useful representation for the pattern of communications between individuals. Alternatively in *Temporal Vertex Networks* each node carries a single time. The citation network provides a natural example of the latter as each paper has its publication date. Here then we will focus on the analysis of this second type of temporal network, using the bibliometric context of citation networks to motivate our work.

The causal structure of citations plays a central role in bibliometric analyses. At the simplest level understanding the different time scales for citation patterns seen in different research fields is known to be essential. In Price's model (Price, 1965) vertices appear in a fixed order, reflecting the order of publication of real citation networks. Price's model captures the essential nature of a citation; they are always from newer to older papers. Applying Price's growing network model to other contexts where time plays a different role makes no sense e.g. links between web pages are not constrained by the age of a web site.

The constraints imposed by time are very different from the spatial constraints. Network science has few tools specifically developed to work with temporal vertex networks. However as part of our work we adapt results found in other areas: discrete mathematics, quantum gravity, and in computer science. Bibliometrics asks very different questions about such networks so applying these ideas is not always straightforward.

Our hypothesis is that existing network measures do not account for the constraint of time. So we have embarked on a programme to develop new temporally aware network measures and to prove their utility in the context of citation networks.

Methods and Data

Our networks are defined such that each node has a unique time. Edges can only exist from a younger to an older node, see Figure 1. Citations between academic papers are a good example, patents and court rulings have similar citation structures. All edges are directed, but the arrow of time also ensures that such networks will have no loops (acyclic) provided you follow the direction of the edges. The formal name for such a network is a *Directed Acyclic Graph* or *DAG* for short.

In practice, citation data is not exactly a DAG but we found that citations in the 'wrong' direction form less than 1% of our data so they should have a limited effect on any conclusions. We construct a true DAG by dropping any such acausal citations.

We have used a variety of data sets in our work (Clough et al., 2015, Clough & Evans, 2014). We have used citation information on the arXiv repository taken from two independent different sources. This allows us to check that our results are robust against any differences in citation extraction. First we use the KDD cup data (2003) which covers the first ten years of the hep-ph and hep-th sections (theoretical and phenomenological particle physics respectively). We have also looked at a separate version which covers all sections of arXiv up to 2013 which was derived from paperscape.org they also form a citation network.

We have also studied the citation network of around 4,000,000 US patents between 1975 and 1999 (Hall, Jaffe, & Trajtenberg, 2001). Finally we worked with the network defined by about 25,000 judgements of the US Supreme court 1754 to 2002 (Fowler & Jeon, 2008).

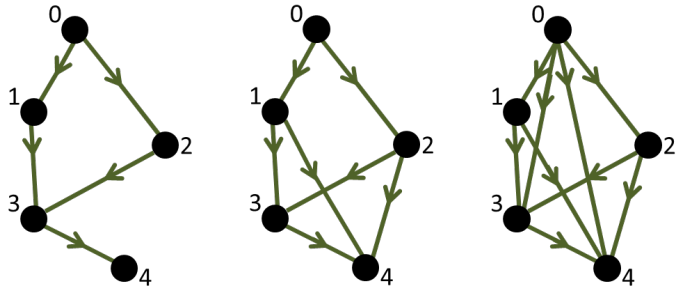


Figure 1 The unique transitively reduction (left) and transitive completion (right) of the citation network (a Directed Acylic Graph or DAG) shown in the centre. All casual relationships implied by an edge in the central network appear as an explicit edge in the right hand network. The edges in the left hand network are the least required to capture all these causal relationships.

Transitive Reduction (TR)

Our first example of a network operation, which takes account of the constraint of time, is Transitive Reduction (*TR*). In TR, links are removed provided that they leave the connectivity of every pair of nodes unchanged. That is if there was a path between a given pair of nodes (respecting the direction of the links) before TR, there will still be at least one such path after TR. This process can be defined on any network but for DAGs it is guaranteed to produce a unique result, see Figure 1. Algorithms for this procedure are well known in computer science but we found basic implementations in python were sufficient even for our largest networks (Clough et al., 2015)

Once we have this essential causal core of our citation network we illustrate our approach with two simple measures: the fraction of edges lost in the TR process and a comparison of the citation count of papers before and after TR.

Dimension

In bibliometrics, we often place papers in different fields as there is great interest in understanding the relationships between topics, as illustrated by maps-of-science (such as

Börner et al., 2012). It is natural to ask if we can assign a sense of dimension to such 'topic' spaces. A high dimension would indicate that researchers can develop work in several independent directions, a low dimension indicates that all the work in that field is tightly linked with little independence. There are some standard ways to assign an effective dimension to a network but these all assume that all directions are similar, just as moving left/right or forwards/backwards is the same for a ball on a flat table. Unfortunately, none of the measures used in the network science literature take account of time, which is a very different sort of dimension. Given that temporal information is an essential part of the definition of a citation network, we must work with a different type of measure. Our work (Clough & Evans, 2014) draws on inspiration from work in discrete mathematics on *posets* (partially ordered sets, e.g. Bollobás & Brightwell, 1991) and from the Causal Set programme of quantum gravity (e.g. Reid, 2003).

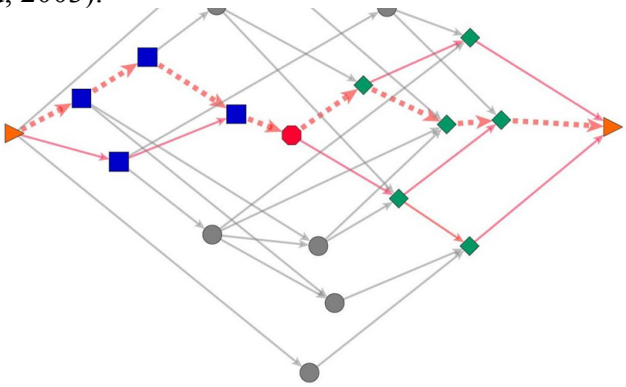


Figure 2 An illustration of the box counting method to find dimension. Here the source and the target papers (triangles at left and right respectively) define an interval of N=19 papers - the other vertices shown here. The edges represent the transitively reduced citation network of all twenty paper. The midpoint is shown as the red circle in the centre. It defines two sub-intervals N_1=4 (blue squares) on the left and N_1=6 on the right (green diamonds). This gives D=2.16 and D=1.61 as our dimension estimates. The example was generated by throwing points down with one space and one time coordinate chosen at random, i.e. D=2.

Our first approach is a simple box counting method (Reid, 2003). We first choose a pair of papers, the source and target nodes, at random. We then find the *interval* defined by the source and target nodes, which is the set of all N papers which lie on a path between source and target. As always our paths must respect the direction of time. Next we find the midpoint, a node chosen such that two sub-intervals defined by source and midpoint, and by midpoint and target nodes, are roughly equal size $N_1 \approx N_2$. It then follows that we should expect the 'length' scale of our two smaller intervals interval to be roughly half that of the large interval. Assuming papers are scattered at equal density in our data, we can use the number of points in an interval as a measure of the volume in the space-time. It then follows that the ratio of the number of points from small to large interval should scale as $N_1/N \approx N_2/N \approx 2^{-D}$. By analysing many intervals within one academic field the space-time dimension D (one time and $(D-1)$ topic space dimensions) of that field may be found.

The second method we use here is the Myrheim-Meyer dimension estimator (see Reid, 2003 for references). To do this we again pick a source and target paper. We then count the number causally connected pairs P in the interval defined by our source and sink which contains N nodes and these are related by $(P/N^2)=\Gamma(D+1)\,\Gamma(D/2)\,/\,(4\,\Gamma(3D/2))$ where $\Gamma(x)$ is the standard Gamma function. This formula is derived for a large N by assuming points are sprinkled at uniform density in Minkowskii space-time. We have also used the same approach

to show that in a different type of space, the cube box space of Bollobás & Brightwell (1991) the formula is simply $P=N(N-1)/2^D$.

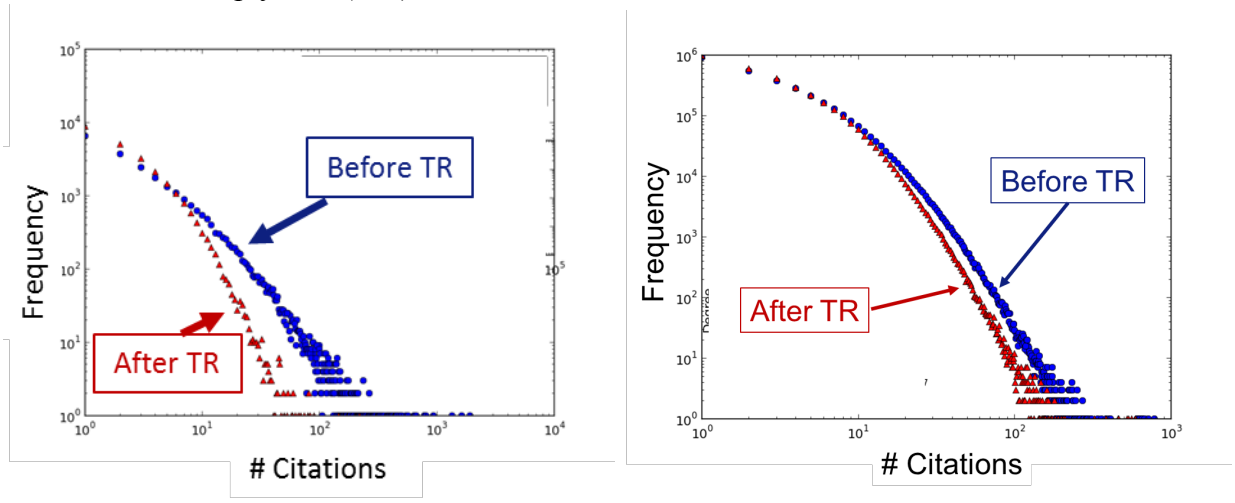


Figure 3 The citation count distribution before and after TR. On the left the results for the quant-ph section of arXiv (paperscape dataset) shows a significant change and an overall loss of around 80% of the edges. On the other hand, US patents shown on the right lose around 15% of edge and the citation distribution remains similar.

Findings

One of the most striking findings is that different types of citation network show very different behaviour under TR. All the citations networks of academic papers we have studied have shown a dramatic loss in the number of edges, typically around 70% to 80%. Further, it is the high cited papers which suffer the most as can be seen in Figure 3 for the hep-th arXiv where the citation distribution becomes noticeably steeper. On investigation it is clear that the edges which remain are those with the age difference between cited and citing papers. Interestingly citations in US supreme court judgements show a similar pattern (not shown) but US patents show only a moderate loss as shown in Figure 3.

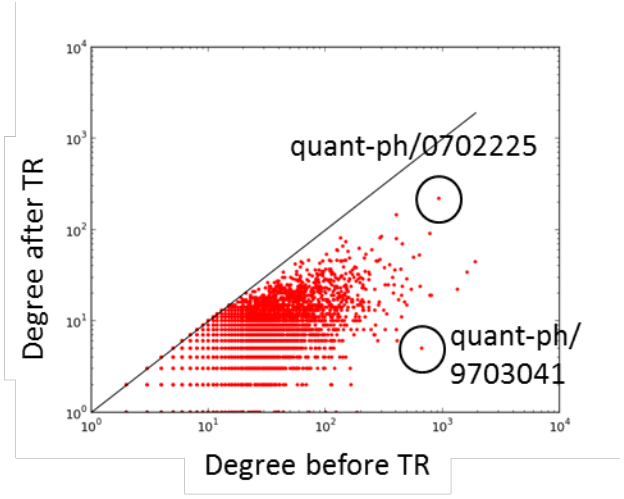


Figure 4 The citation count before and after TR for each paper in the quant-ph paperscape data.

Rather than looking at these bulk statistics we can look at the effect of TR on individual papers. Of course there are winners and losers. The example of the astro-ph arXiv section from paperscape.org highlights the different fates of two papers, see Figure 4. Paper quant-ph/9703041 (an older research paper on quantum entanglement) is one of the most highly cited papers with 664 citations yet TR shows that anyone using quant-ph/9703041 also took

information (directly or indirectly) from five other papers. On the other hand, paper quant-ph/0702225 (a more recent review of quantum entanglement) begins with a similar number of citations, 937, yet after TR it retains 219 of these.

We have also run our dimension measures on a variety of data sets. Our results are consistent whichever of the measures we use. What emerges is that we can generally give each field a well-defined dimension and that these are significantly different. For instance Figure 5 shows how papers in two parts of the arXiv repository have distinctive dimensions. For the arXiv we have found dimensions of about for hep-th (string theory), 3 for both hep-ph (particle physics) and quant-ph (quantum physics), and around 3.5 for while astro-ph (astrophysics).

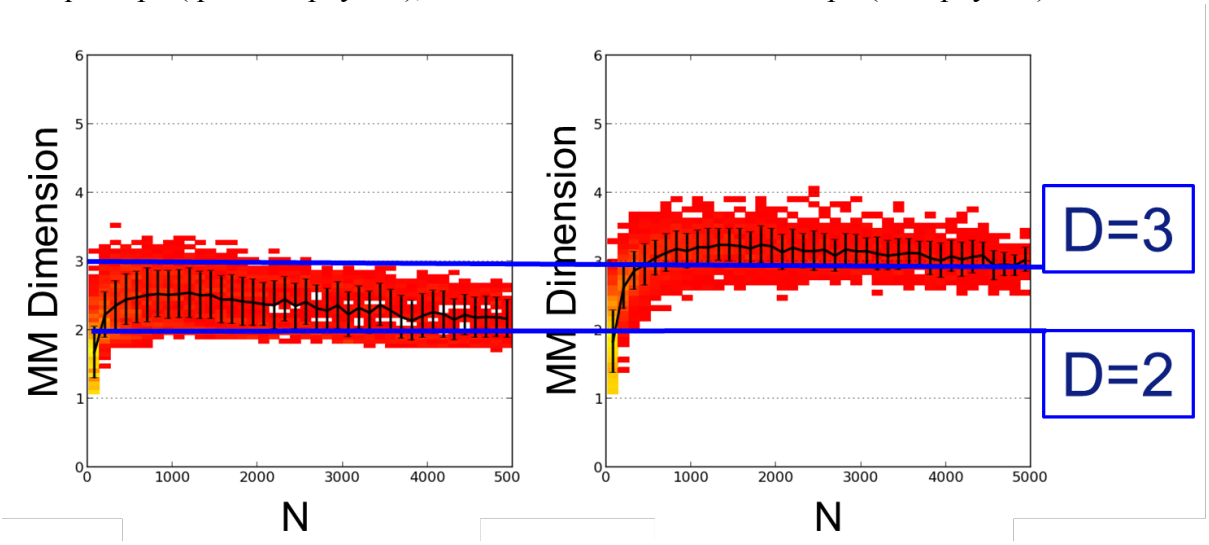


Figure 5 Dimension of two parts of the arXiv repository (KDD cup dataset) using the MM (Myrheim-Meyer) dimension estimator. Each point represents the dimension estimated from an number of intervals defined by two randomly chosen papers. On the left the hep-th section is seen to be of lower dimension than the hep-ph section shown on the right.

Discussion

For us TR captures the essential causal skeleton underlying the citation network. If information is flowing from older papers to newer papers and this is reflected in the bibliographies, then all the links in the transitively reduced network are the minimum needed for such a process. Of course in practice authors may use 'short cuts' and derive information directly from older papers, but equally such short cuts were not essential and therefore there is no reason to suppose they were important. We see TR as providing a lower bound on the actual route used by the flow of important information. To go beyond this, some sort of expensive semantic analysis is needed, be it via automatic methods or by hand.

In fact we believe the transitively reduced network may be much closer to the actual set of citations of direct relevance to a publication. We have found that around 80% of links between academic papers are removed by TR. Interestingly this matches the figure given by Simkin & Roychowdhury (2003, 2005) who suggest around 80% of citations are copied from intermediate works. Any citation which was copied will always be removed by TR.

Our suggestion is that TR could be an important way to reveal which papers were essential for the developments described in a new paper. Not surprisingly, these tend to be recent papers but it is still a surprise to find such a large fraction are removed. We have shown that there are big differences in the post-TR citation count of papers in similar fields with similar high citation counts. This could be a way to discriminate between papers and could provide an alternative basis for a recommendation system. For instance searches could be ordered by post-TR citation count. One hypothesis is that papers which retain a high citation count after

TR have been used across a wider range of topics. These are works which might be of more interest to researchers looking for papers outside their normal field of interest.

The behaviour of our patents and court citations also shows how TR can be a useful way to highlight different citation practices. The court data behaves in a way which is similar to that of academic papers with a large number of edges lost under TR. On the other hand, patents lose only a small fraction of their edges. The difference reflects the fact that for a patent, citations are a recognition of prior art, a legal necessity when writing a patent. However, as a patent is meant to be a novel development, they presumably try not to refer to earlier work so as to appear to be as different as possible from the literature. On the other hand, US Supreme Court judges seem to act like academic authors, citing older documents, which may have no direct relevance, along with the more recent documents, which have the latest distillation of this knowledge and are the real source of any innovation.

Our dimension measures again highlight difference between fields. We interpret the low dimension of the hep-th arXiv to suggest that string theory is a rather narrow field feeding off a few strands of research, at least when compared to hep-ph, quant-ph and astro-ph where research appears to be moving in a wider range of directions.

Conclusions

We have argued that citation networks require a new type of measure which takes account of the constraint imposed by time. We have given some examples of how this can be done and shown that they reveal some interesting features in real citation networks. We hope to add other measures and to improve the interpretation of our results by comparing them with non-network derived measures.

Acknowledgments

We would like to thank Damien George and Robert.Knegjens who provided us with access to their paperscape.org arXiv citation data. We also would like to acknowledge useful conversations with K.Christensen, J.Gollings, A.Hughes and T.Loach.

References

Bollobás B, & Brightwell G. (1991). Box-Spaces and Random Partial Orders. *Transactions of the American Mathematical Society*, 324, 59-72.

Börner, K.; Klavans, R.; Patek, M.; Zoss, A.M.; Biberstine, J.R.; Light, R.P.; Larivière, V. & Boyack, K.W. (2012). Design and update of a classification system: The UCSD map of science *PloS one*, 7, e39464

Clough, J.R.; Gollings, J.; Loach, T.V. & Evans, T.S. (2015). Transitive reduction of citation networks *Journal of Complex Networks,* 3, 189-203 http://dx.doi.org/10.1093/comnet/cnu039.

Clough, J.R. & Evans, T.S. (2014). What is the dimension of citation space? arXiv:1408.1274.

Fowler, J.H. &Jeon, S. (2008). The authority of Supreme Court precedent, *Social Networks*, 30, 16-30.

Hall, B., Jaffe, A. & Trajtenberg, M. (2001). The NBER Patent Citations Data File, NBER Working Paper Series.

Holme, P. & Saramäki, J. (2012). Temporal Networks. *Physics Reports*, 519, 97-125.

KDD Cup, 2003: *Network mining and usage log analysis*. Retrieved 1st October 2012 from http://www.cs.cornell.edu/projects/kddcup/datasets.html.

Price, D.S. (1965). Networks of Scientific Papers. *Science*, 149, 510-515.

Reid, D.D. (2003). Manifold dimension of a causal set: Tests in conformally flat spacetimes. *Phys. Rev. D*, 67, 024034

Simkin, M.V. & Roychowdhury, V.P. (2003). Read before you cite! *Complex Systems*, 14, 269-274.

Simkin, M.V. & Roychowdhury, V.P. (2005). Stochastic modeling of citation slips. *Scientometrics*, 62, 367-384.

Coming to Terms: A Discourse Epistemetrics Study of Article Abstracts from the Web of Science

Bradford Demarest[1], Vincent Larivière[2], Cassidy R. Sugimoto[3]

[1] *bdemares@indiana.edu*
Indiana University – Bloomington, School of Informatics and Computing, Bloomington, IN (United States)

[2] *vincent.lariviere@umontreal.ca*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, Montreal, QC (Canada)

[3] *sugimoto@indiana.edu*
Indiana University – Bloomington, School of Informatics and Computing, Bloomington, IN (United States)

Abstract
This study investigates the relative power and characteristics of a set of social and epistemic terms to distinguish among disciplines of research article abstracts, using a corpus of 928,572 abstracts from 13 disciplines indexed by Web of Science in 2011. Applying the machine-learning approach to discourse epistemetrics using a sequential minimal optimization (SMO) algorithm, and a feature set of terms derived from Hyland's (2005) metadiscourse studies per Demarest and Sugimoto (2014), the current paper reports subsets of terms that best (and least) distinguish among disciplines, finding that the terms least able to distinguish among disciplines are rarely used and overwhelmingly adjectival or adverbial markers of authorial attitude, reflecting personal positioning, while terms best able to distinguish disciplines are mostly verbs frequently used as engagement markers, framing the generation of knowledge for the readership in ways that are standardized within disciplines (while varying among them). We plan to analyze the findings of the current research-in-progress from discipline-based as well as term-based perspectives, incorporating both into a two-mode network, as well as incorporating finer grained data for specific specializations to compare with the current higher-level disciplinary findings.

Conference Topic
Methods and techniques, altmetrics

Introduction
Understanding and depicting the relationships among different academic realms (whether disciplines, fields, specialisms, or a host of other divisions using some combination of social, epistemological, and institutional aspects) is a well-studied subarea of scientometric (Leydesdorff & Rafols, 2009). Initial forays into modeling disciplinary differences based on a core set of social and epistemic terms have yielded potentially promising results (Demarest & Sugimoto, 2013; Demarest & Sugimoto, 2014). However, no studies to date have used computational approaches to compare the abilities of specific social and epistemic terms to distinguish among disciplines. The current work-in-progress seeks to enact such a comparison, using a machine-learning approach to derive term differences between pairs of disciplines and by extension between a given discipline and all other disciplines under study. In finding the social and epistemic terms that best distinguish among academic disciplines, we hope to open new dimensions of analysis of the sciences through their texts.

Literature Review
There have been very few previous attempts to map the relatedness of academic disciplines based upon common social and epistemic terms. However, previous research of social and epistemic discourse usage in different academic disciplines as well as previous studies of document, journal, author, and discipline similarity or relatedness based on a variety of other measures guide the current study.

Differences in how academic disciplines employ language that positions the author in relation to the reader, the text itself, and previous scholars and works have been studied under various monikers, including stance (Biber & Finegan, 1989), metadiscourse (Hyland & Tse, 2004), appraisal (Martin & White, 2008), and attitude (Halliday, 1985). For the most part these differences have not been studied using automated quantitative methods (although cf. Argamon and Dodick, 2004), and in no cases have the resulting metrics been used as a basis for mapping the relatedness of disciplines. The current study draws upon Hyland's (2005) study of metadiscourse in a number of different disciplines, leveraging a set of words and phrases that Hyland (2005) found to be widely occurring in academic writing as our feature set for machine learning-based modeling of term differences among disciplines.

Previously, scholars have sought to map science based upon patterns of co-citation (Boyack, Klavans, & Börner, 2005) as well as topic, via ISI subject headings (e.g., Leydesdorff & Rafols, 2009). Other studies of similarity or relatedness have sought to compare multiple kinds of networks, including "bibliographic coupling, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks" (Yan & Ding, 2012, p. 1313). While the current work-in-progress focuses on a single type of similarity, it is with the intention of eventually adding to and comparing with these previously established measures of comparison. Furthermore, in order to create results that are comparable to previous work, we will also draw our data from the Web of Science, focusing specifically on the genre of scholarly articles, and use the high-level subject categories (although in future iterations of this study we hope to look at both higher and lower-level subject categories).

Methods

The current study analyzes all journal article abstracts from 13 disciplines contained in the Web of Science from 2011, totaling 928,572. Table 1 provides an overview of disciplines and counts of abstracts in the data corpus.

Table 1. Counts of abstracts by discipline.

Discipline	Abstracts
Engineering and Tech	172949
Biomedical Research	153166
Chemistry	129685
Physics	121702
Biology	93765
Earth and Space	70018
Mathematics	42685
Social Sciences	40463
Professional Fields	34590
Health	28343
Psychology	25802
Humanities	13673
Arts	1731
TOTAL	**928572**

For each abstract, relative frequencies were computed for 307 words or phrases taken from Hyland (2005). These terms fall into one or another of the following categories: hedges, boosters, attitude markers, engagement markers, and self-mentions. Hedges (e.g., "perhaps", "possible", "approximately") mitigate the certainty of an assertion, while boosters (e.g., "clearly", "obvious") amplify it. Attitude markers, such as "unexpectedly" or "unfortunately", frame assertions affectively, expressing the author's emotion regarding the

asserted facts, as distinct from their assurance of the facts' certainty. Engagement markers (such as "the reader" and "you", but also imperative verbs such as "consider" or "observe") address the reader explicitly or implicitly, and guide the reader to specific social and epistemic framing of an assertion (e.g., as an externally observable fact or as an idea intended for mental simulation). Finally, self-mentions, such as "I", "we", or "the author", serve as means for authors to insert themselves into the text, either as subjective actors or as social players (whether alone or as part of an authorial cohort).

After preparing the data, the Sequential Minimal Optimization algorithm (SMO) (Platt, 1998), a support-vector model classifier implemented in the WEKA v3.6.6 tool (Hall et al., 2009), was employed to create models distinguishing between each pair of disciplines based on the socio-epistemic features' relative frequencies. The resulting term weights for each model of discipline pairs were then normalized across the model, such that the absolute values of weights for a given discipline pair model would sum to 1. Model-normalized weights for each term were then averaged for each discipline across all discipline pairs for which the given discipline was a pair member. For the sake of standardization, negative term weights indicate a positive correlation with a given discipline (i.e., the more frequently the term appears in a text, the more likely this text belongs to the given discipline), while positive term weights indicate a negative correlation (i.e., the more frequently the term appears in the text, the less likely this text belongs to the given discipline).

Results

Due to space limitations, we eschew reporting the full 307 term set of results, focusing instead on the terms that most and least distinguish among disciplines. We discern these terms based upon the standard deviation of model-normalized average weights, as terms that discern well among disciplines will result in strong positive as well as negative weights, depending on which discipline is being modeled, while terms whose weights have small absolute values will in turn have smaller standard deviations, as all weights approach the 0 point.

Table 2 reports the 20 terms with the highest standard deviations of model-normalized average weights, as well as the 20 terms with the lowest standard deviations. While the results might at first blush suggest that the terms with the lowest standard deviations are part of a universal academic discourse, it is worth noting that many of the terms in the Bottom 20 list are exceedingly rare in the sample – out of 928,572 abstracts, "unbelievable" appears in 3 of them (although "shockingly" also appears in 3 abstracts; however, "unbelievable" is found in 2 engineering abstracts and one humanities abstract, suggesting that the scant data that exists shows no distinction between two otherwise fairly different disciplines). Also worth noting is that any terms that appeared in no abstracts at all are eschewed from the reported results.

However, the bottom 20 terms do provide some information about scholarly writing across the disciplines – the vast majority of these terms (19 out of 20) act as attitude markers; given the wide range of adjectives and adverbs available to describe the affective state of the author (and given that adjectives and adverbs are linguistic "open classes", i.e., new words can and are generated for these classes regularly), it is not surprising that such terms would be diffuse, rare, and not strongly indicative as individual terms.

Pivoting to consider the top 20 terms, the first notable characteristic is that where the bottom 20 terms tend toward adjectives and adverbs (as well as attitude markers), 19 of the top 20 terms are either self-mentions or engagement markers (and the latter for the most part are verbs). While nouns and verbs are also linguistic open classes, the use of verbs to describe the epistemic frame of scientific work here as well as the terms with which scientific authors refer to themselves can be seen to be more standardized within disciplinary communities, whereas the attitude markers of the bottom 20 terms are more personalized. The indicative

strength of self-mentions such as “we”, “my”, and “author”, as well as verbs like “argue” and “measure” also resonates with previous findings of Demarest and Sugimoto (2014), with “argue” and “my” serving as a strong indicator of philosophy and “measure” and “we” a better indicator of psychology and physics in dissertation abstracts as well.

Table 2. The top and bottom 20 social and epistemic terms for distinguishing among disciplines (ranked by standard deviation).

Top 20		Bottom 20	
Term	Standard Deviation	Term	Standard Deviation
we	0.009848	shockingly	0.0009166
argues	0.009686	view	0.0008793
prove	0.009614	disappointed	0.0008707
argue	0.009098	astonishingly	0.0008043
author	0.009063	!	0.0007801
showed	0.008494	incontestable	0.0007541
about	0.008138	knowledge	0.0007406
let	0.008044	incontrovertible	0.0007283
proved	0.008019	presumable	0.0007005
my	0.007908	unclearly	0.0006577
recall	0.007684	desirably	0.0006524
estimate	0.007646	amazed	0.0006068
review	0.007592	disappointingly	0.0006046
measure	0.007268	uncertainly	0.0004573
pay	0.007173	undisputedly	0.0003956
thought	0.007102	unbelievably	0.0003247
claims	0.006978	incontrovertibly	0.0002968
consider	0.006879	incontestably	0.0002821
shown	0.006687	astonished	0.0002649
set	0.006672	unbelievable	0.0001121

Another aspect of the findings to consider is that while the standard deviation values derive from the full set of model-normalized average weights, in some circumstances high standard deviation values can derive from a single outlier, while in others it derives from a more uniform spread of weights. Figure 1 depicts the model-normalized average weights for the top 20 terms ranked by standard deviation. Visual inspection reveals terms whose weights are more uniformly distributed (e.g., “author”), which suggest that they may serve as robust terms to distinguish among a variety of disciplines, while other terms (e.g. “let”, “prove”, and “proved”) serve as strong indicators of a single outlier discipline, with all other disciplines much more tightly clustered. As it happens, the terms “let”, “prove”, and “proved” provide a strong indication of mathematics as they occur more frequently in a text, in contrast to all other disciplines.

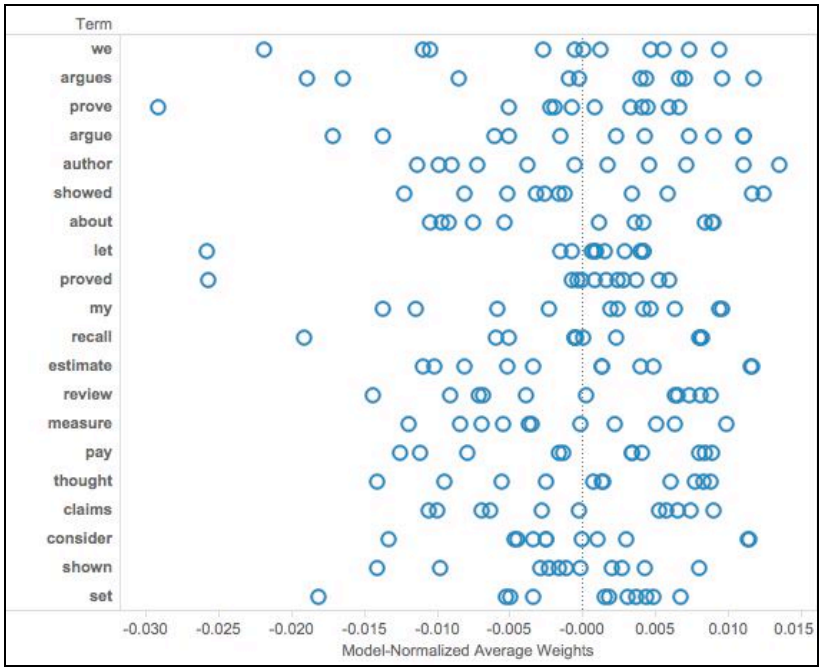


Figure 1. Model-normalized average weights (Top 20, ranked by standard deviation).

Future Directions

While the results of the current study-in-progress have focused on summary ranking and overall patterns of distribution of weights per term, our next goals in the near term are to more deeply tease apart trends as they appear for single disciplines as well as groups of disciplines, including the traditional groupings of soft vs. hard and pure vs. applied (Biglan, 1973). Further, we can derive overall measures of similarity among disciplines from the overall accuracy measures of the machine-learning models from which these terms are taken (per Demarest & Sugimoto, 2014), or more ambitiously we could seek to cast disciplines and terms in a bipartite network, to more fully grasp the interplay between different disciplinary communities and the words they use.

More distantly, we intend to use this same approach, in light of patterns and trends perceived at the current level of aggregations, to consider specializations, so that we may ask questions such as how broad the social and epistemic spread of specialized areas of study are within disciplines – are some disciplines more socially or epistemically diverse, and others more centralized? Do these degrees of variety reflect patterns of fragmentation and specialization in subject area? It is questions such as these that compels the current research-in-progress.

References

Argamon, S. & Dodick, J. (2004). Conjunction and modal assessment in genre classification: A corpus-based study of historical and experimental science writing. In *AAAI Spring Symposium on Attitude and Affect in Text*. Retrieved from http://www.aaai.org.proxyiub.uits.iu.edu/Papers/Symposia/Spring/2004/SS-04-07/SS04-07-001.pdf?origin=publication_detail

Biber, D. & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text*, *9*(1), 93–124.

Biglan, A. (1973). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology*, *57*(3), 195.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351–374. doi:10.1007/s11192-005-0255-6

Demarest, B. & Sugimoto, C. R. (2013). Interpreting epistemic and social cultural identities of disciplines with machine learning models of metadiscourse. *In Proceedings of ISSI 2013 (Vol. 2, pp. 2027–2030)*. Vienna. Demarest, B., & Sugimoto, C. R. (2014). Argue, observe, assess: Measuring disciplinary identities and

differences through socio-epistemic discourse. *Journal of the Association for Information Science and Technology*. doi: 10.1002/asi.23271

Diermeier, D., Godbout, J.-F., Yu, B., & Kaufmann, S. (2011). Language and Ideology in Congress. *British Journal of Political Science*, *42*(01), 31–55. doi:10.1017/S0007123411000160

Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Edward Arnold Press.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London: Continuum International Publishing Group.

Hyland, K. & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied Linguistics*, *25*(2), 156–177.

Klavans, R. & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, *60*(3), 455–476.

Leydesdorff, L. & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348–362.

Martin, J. R. & White, P. R. R. (2008). *Language of Evaluation: Appraisal in English* (First Edition.). Palgrave Macmillan.

Platt, J. C. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. In *Advances in Kernel Methods - Support Vector Learning*. Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.4376

Yan, E. & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313–1326. doi:10.1002/asi.22680

Using Hybrid Methods and 'Core Documents' for the Representation of Clusters and Topics: The Astronomy Dataset

Wolfgang Glänzel[1,2] and Bart Thijs[1]

[1] *wolfgang.glanzel@kuleuven.be, bart.thijs@kuleuven.be*
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
[2]Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

Abstract
Based on a dataset on Astronomy & Astrophysics a hybrid cluster analysis has been conducted. Hybrid clustering was based on a combination of bibliographic coupling and textual similarities using Louvain method at two resolution levels. The procedure resulted in seven and thirteen clusters, respectively. The statistics reflect a high quality of classification. For labelling and interpreting clusters, *core documents* are used. The results of these two scenarios are presented, discussed and compared with each other. The two scenarios clearly result in hierarchical structures that are analysed with the help of a concordance table. Furthermore, the core documents help depict the internal structure of the complete network and the clusters.
This work has been done as part of the international project 'Measuring the Diversity of Research' and in the framework a special workshop on the comparative analysis of algorithms for the identification of topics in science organised in Berlin in August 2014.

Conference Topic
Methods and techniques (special session on algorithms for topic detection)

Introduction
Within the framework of the event series on 'Measuring the Diversity of Research' a special workshop on the comparative analysis of algorithms for the identification of topics in science was organised in Berlin in August 2014. A dataset downloaded from Thomson Reuters Web of Science covering the annual volumes 2003–2010 was shared with all contributors in order to test the various algorithms and techniques and to compare the results of the different approaches. On the basis of the shared Astronomy & Astrophysics dataset the following analysis has been conducted at our institute. In particular, the topic structure of the subject defined by the set was analysed using two different but related techniques. A cluster analysis was based on bibliographic coupling and textual similarity. And *core documents* (Glänzel & Czerwon, 1996) defined on the same links were used to represent topics within the subject and to depict the internal structures of both subject and clusters (cf. Glänzel & Thijs, 2011). Main results are presented in the following, but changing parameters of the algorithm and of the combination of the components leads to further results.
Currently a new and more robust method for the measurement of textual similarities and thus for the revision of the lexical component is in development. A comparison of the results of the present study with those of the new algorithm is part of the ongoing project and will be presented on a later occasion, when available.

Methodological aspects
The advantage of using hybrid lexical–citation based methods, notably of combinations of term-frequency and bibliographic coupling, has already been discussed in previous studies (e.g., Glenisson et al., 2005; Boyack & Klavans, 2010). However, at this level of aggregation (topics within the same field or discipline) we have encountered several specific problems that have already been reported in earlier studies in the context of the detection of emerging topics (e.g., Glänzel & Thijs, 2012). Terms and phrases might become less specific since they express common knowledge base and vocabulary while others might gain more 'information

value'. The most important TF-IDF keywords and terms alone are often not specific enough for topic description and labelling. Thus a larger set of terms is needed to describe topics at this level. A possible solution has already be discussed already in earlier studies (e.g., Glänzel & This, 2011): On one hand, depending on the level of aggregation *and* the discipline under study, the weight of the two components can be adjusted and, on the other hand, instead of the best TF-IDF terms *core documents* can be used to describe and label clusters. In order to apply the hybrid clustering we have only vertices with *positive* degree (i.e., documents with at least one link) taken into account. Furthermore, we have removed all papers with publication years outside the period 2003–2010. Table 1 shows the description of the dataset.

Table 1. The input dataset.
[Data sourced from Thomson Reuters Web of Science Core Collection]

Data	Documents	Percentage
Original dataset	111514	100.00
Not present in ECOOM Database	103	0.09
Publications in 2003-2010	110412	99.01
Excluded from all analysis	1205	1.08

We applied Louvain method (Blondel et al., 2008) using Pajek (Batagelj & Mrvar, 2003) to this dataset. The reason for this choice was that hierarchical clustering with Ward used in previous projects (e.g., Thijs et al., 2013) often results in a heterogeneous "hotchpotch" cluster of objects that can otherwise not be assigned. Therefore we decided to apply Louvain method. We conducted a hybrid clustering with two components: *bibliographic coupling* (BC) and *textual similarity* (TS), where we used a weight of 0.75 for BC and 0.25 for TS according to the algorithm described in Glänzel & Thijs (2011). In particular, the underlying similarity measure r is defined as the cosine of the linear combination of the underlying angles between the vectors representing the corresponding documents in the vector space model, i.e.,

$$r = \cos\left(\lambda \cdot \arccos(\eta) + (1-\lambda) \cdot \arccos(\xi)\right), \quad \lambda \in [0,1],$$

where η is the similarity defined on bibliographic coupling and ξ the textual similarity. The λ parameter defines the convex combination, $\arccos(\eta)$ and $\arccos(\xi)$, respectively, denote the two underlying angles. Furthermore, we have conducted the clustering at two resolution levels, namely 0.7 and 1.4. The results of these two scenarios will be presented and briefly discussed in the following section.

Results

The results using both resolution levels are briefly summarised in Table 2. The number of documents, that could not been clustered, is marginal. The number of clusters has almost doubled (from 7 to 13) with growing resolution. The solutions for the two resolution levels are presented in Tables 3 and 4. Except for the tiny cluster (#13) on atmospheric turbulence in the second solution, all clusters are of reasonable size. This is expressed by the frequency, i.e., the number of documents per cluster (columns 2–4). The description of the clusters, shown in the last column of the tables, have been derived from the most important TF-IDF terms and the titles of the *core documents*, where the core documents have been determined according to see Glänzel (2012) on the basis of the *degree h-index* of the hybrid document network. In particular, core documents are represented by core nodes, which, in turn, are defined as nodes with at least h degrees each, where h is the h-index of the underlying graph. Or, to express this simpler, degrees of documents are ranked in descending order and the h-core is formed by the documents the degrees of which do not undercut their rank value. This method has proved

efficient in *local* clustering, that is, in clustering of fields or disciplines, where the network h-core usually represents the order of magnitude of 1% of the total document set (see Glänzel, 2012).

Table 2. Description of parameters and results. [Data sourced from Thomson Reuters Web of Science Core Collection].

Number of vertices		108937
Number of edges		87602281
Density		1.5%
All Degree Centralization		0.13
Method		Louvain (Pajek)
Hybridity parameter		$\lambda = 0.75$
Resolution	0.7	1.4
Number of Clusters	7	15
Documents not Clustered	360	360
Modularity	0.61	0.49

Table 3. Scenario 1 (description of structures in the seven-cluster structure). [Data sourced from Thomson Reuters Web of Science Core Collection].

Cluster	Freq	Freq%	CumFreq	CumFreq%	Label
1	20634	18.7%	20634	18.7%	Star Clusters
2	12149	11.0%	32783	29.7%	Terrestrial planets/Extra Solar Planets
3	14365	13.0%	47148	42.7%	Solar Flares
4	17036	15.4%	64184	58.2%	Star Formation
5	20173	18.3%	84357	76.5%	Dark Energy
6	15023	13.6%	99380	90.1%	Gamma Ray Burst
7	10820	9.8%	110200	99.9%	Neutrino

Table 4. Scenario 2 (description of structures in the 13-cluster structure). [Data sourced from Thomson Reuters Web of Science Core Collection].

Cluster	Freq	Freq%	CumFreq	CumFreq%	Label
1	11569	10.5%	11569	10.5%	Star Clusters / Globular Clusters
2	9470	8.6%	21039	19.1%	Disk around a brown dwarf or young star
3	12163	11.0%	33202	30.1%	Extrasolar planetary systems
4	15060	13.7%	48262	43.8%	Solar Flares
5	6481	5.9%	54743	49.6%	Dark Matter Halo: Formation of galaxies
6	10075	9.1%	64818	58.8%	Star formation
7	7523	6.8%	72341	65.6%	Dark Energy
8	9005	8.2%	81346	73.8%	Astrophysical jets and accretion discs
9	10298	9.3%	91644	83.1%	Brane-world black hole
10	5503	5.0%	97147	88.1%	Radio Pulsars
11	2336	2.1%	99483	90.2%	Gamma Ray Burst
12	10224	9.3%	109707	99.5%	Neutrino
13	477	0.4%	110184	99.9%	Atmospheric turbulence

Table 5. Core-document representation of Cluster #5 based on h-core. [Data sourced from Thomson Reuters Web of Science Core Collection].

UT	Degree	Rank	Title
000261696000006	111	1	Non-linear isocurvature perturbations and non-Gaussianities
000278201600003	99	2	Non-Gaussianity of quantum fields during inflation
000260529800008	96	3	Conditions for large non-Gaussianity in two-field slow-roll inflation
000261260200020	88	4	A curvaton with a polynomial potential
000278201600004	86	5	Local non-Gaussianity from inflation
000238060100019	84	6	Non-Gaussianities in two-field inflation
000186983100013	83	7	Generalized chaplygin gas with alpha=0 and the Lambda CDM cosmological model
000246571300004	82	8	Cleaned 3 year Wilkinson Microwave Anistropy Probe cosmic microwave background map: Magnitude of the quadrupole and alignment of large-scale modes
000253980700030	82	9	Non-Gaussianity analysis on local morphological measures of WMAP data
000276102300001	81	10	Scale dependence of local f(NL)
000270036800016	79	11	Non-Gaussianity beyond slow roll in multi-field inflation
000235669800017	78	12	Testing primordial non-Gaussianity in CMB anisotropies
000259692800055	77	13	Anomalous CMB North-South asymmetry
000185760100005	76	14	WMAP and the generalized Chaplygin gas
000250363000004	75	15	Alignment and signed-intensity anomalies in Wilkinson Microwave Anisotropy Probe data
000221258900057	74	16	Numerical analysis of quasinormal modes in nearly extremal Schwarzschild-de Sitter spacetimes
000264762500065	74	17	Modeling gravitational recoil from precessing highly spinning unequal-mass black-hole binaries
000220092300012	73	18	Non-Gaussianity in the curvaton scenario
000242409800004	72	19	Non-Gaussianity of the primordial perturbation in the curvaton model
000242449600008	72	20	A numerical study of non-Gaussianity in the curvaton scenario
000245928000021	70	21	Exploring the properties of dark energy using type-Ia supernovae and other datasets
000248953800006	70	22	Primordial non-Gaussianity in multi-scalar slow-roll inflation
000253764800075	70	23	Further insight into gravitational recoil
000243171800001	68	24	Inflationary trispectrum for models with large non-Gaussianities
000252864000020	68	25	Non-Gaussianity in the modulated reheating scenario
000243171800040	67	26	Primordial trispectrum from inflation
000275514800001	67	27	Disks in the sky: A reassessment of the WMAP ""cold spot""
000278201600005	67	28	Use of delta N formalism-difficulties in generating large local-type non-Gaussianity during inflation
000221258900023	66	29	Curvature and isocurvature perturbations in a three-fluid model of curvaton decay
000221277400044	66	30	Dirac quasinormal modes of the Reissner-Nordstrom de Sitter black hole
000266501900050	66	31	Trispectrum versus bispectrum in single-field inflation
000272271900003	66	32	The subdominant curvaton
000243725400002	65	33	The non-Gaussian cold spot in the 3 year Wilkinson Microwave Anisotropy Probe data
000244679500013	65	34	Mapping the large-scale anisotropy in the WMAP data
000255424300029	65	35	Generation and characterization of large non-Gaussianities in single field inflation
000235939700023	64	36	On the large-angle anomalies of the microwave sky
000250954900032	64	37	A note on the large-angle anisotropies in the WMAP cut-sky maps
000257290600085	64	38	Anti-de Sitter universe dynamics in loop quantum cosmology
000245405900001	63	39	Constraints on the generalized Chaplygin gas model from recent supernova data and baryonic acoustic oscillations
000183377200050	62	40	Generation of dark radiation in the bulk inflaton model
000188864800011	62	41	Large scale structure and the generalized Chaplygin gas as dark energy
...	...	...	...
000256378700020	60	59	A low cosmic microwave background variance in the Wilkinson Microwave Anisotropy Probe data
000259700200011	60	60	Consistency relations for non-Gaussianity

Table 5 lists the core documents of Cluster #5 of the first scenario with seven clusters as an example. The degrees given in the table also illustrates the role of core documents in the cluster: Core documents are by definition strongly interlinked with many other documents and therefore play a representative and central part in a network. And they are suited to depict the internal structure of the complete network, of a cluster or of parts of it. In this context Cluster #5 has not been chosen by chance. The core documents of this cluster form the centre of the structure. Links connecting core documents reveal the internal structure of both the field under study and the clusters as the links with other core documents of the same cluster as well as with those of other clusters are distinctly apparent. Beside this cluster, also cores documents of cluster 7 play a central part. This is shown in Figure 1. Core documents of cluster 5 are marked in pink, those of Cluster 7 in auburn.

By contrast, Figure 2 presents the concordance between the two scenarios. Indeed the two resolutions results in a different number of clusters as already have been shown in Tables 3 and 4. Now the question arises of whether the two approaches yield completely different structures or almost concordant hierarchic structures, where the choice of the resolution would go with merging and splitting clusters, respectively. The first case would, of course, be problematic and point to the possible inappropriateness of methodology, while latter case testifies consistency of the chosen method. Cluster concordance of the results of the two scenarios are visualised in Figure 2.

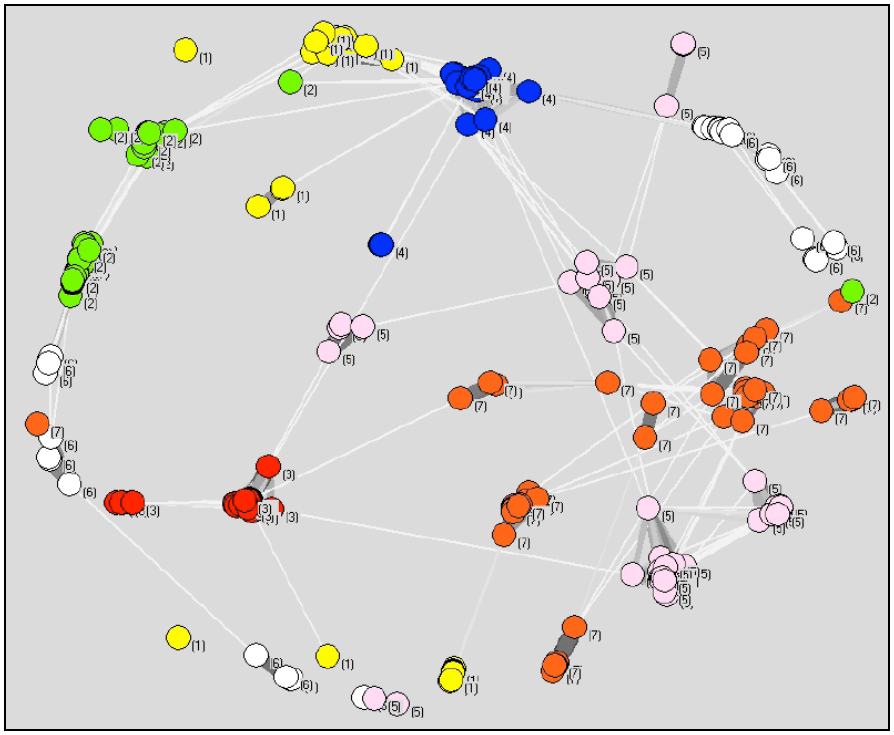


Figure 1. Structure of core documents in 7 clusters according to scenario 1 (Pajek with Fruchterman-Rheingold layout) [Data sourced from Thomson Reuters Web of Science Core Collection].

Hybrid Approach 13 Clusters \ Hybrid Approach 7 Clusters	1	2	3	4	5	6	7
1	94	0	0	5	0	1	0
2	96	1	0	2	0	1	0
3	1	95	3	1	0	0	0
4	1	0	91	0	0	7	0
5	1	0	0	86	2	1	10
6	0	0	0	100	0	0	0
7	0	0	0	2	98	0	0
8	2	0	0	4	0	94	0
9	0	0	0	0	97	0	2
10	1	3	1	1	42	51	0
11	1	2	0	1	0	95	0
12	0	0	0	1	2	1	95
13	5	61	24	0	2	0	7

Figure 2. Cluster concordance: scenario 1 – scenario 2 (overlap in %). [Data sourced from Thomson Reuters Web of Science Core Collection]

The document overlap in the corresponding clusters is expressed in per cents and, in order to facilitate interpretation, marked in different colours. Percentages sum up to 100% by rows. If one neglects the light-weight Cluster #13 in the second scenario, which actually represents just 0.4% of the total, one observes an almost perfect concordance of three clusters in scenarios 1 and 2 (#2 = #3, #3 = #4 and #7 = #12), one cluster splits up into two others (#4 = #5+#6) and finally two clusters split up into three clusters each, namely #5 = #7+#9+#10 and #6 = #8+#10+#11. Thus Cluster #10 in scenario 2 is the only one that breaches the strict hierarchy in the structures of the two scenarios. Its documents are almost equally distributed over Clusters #5 and #6 in scenario 1. The tiny one (#13) in the second

scenario can be considered a small sub-cluster of #2 in the first one, where it represents just slightly more than 2% of the documents of the total cluster.

Conclusions

Our main conclusions refer to two issues, firstly to the *clustering results* and secondly to the role of *core documents*. As to the clustering, both scenarios resulted in an almost perfect hierarchic structure. Cluster concordance and hierarchy was strong except for the cluster on 'Radio Pulsars' in the 13-cluster solution. This cluster was almost evenly spread over the clusters on 'Dark Energy' and 'Gamma Ray Burst' in the seven-cluster solution. Nevertheless, hierarchical assignment of 'Atmospheric Turbulence' in scenario 2 was also somewhat "fuzzy", but had a main concordance of more than 60% of documents with 'Coronal Loop' in the first scenario. In all other cases concordances were around or even above 90% document overlap.

The second group of remarkable observations refer to core documents. These documents represent the links across clusters as well as the internal topic structure of the clusters. In this context we have to repeat that core-document identification is in principle *independent* of clustering and thus does not require any cluster analysis or community detection, but it can be seamlessly integrated into clustering exercises, provided the same type of links, i.e., bibliographic coupling, co-citation, text similarity or hybrid, are used. Core documents reinforce the observation concerning centric results of the hybrid clustering. Core documents of the clusters on 'Dark Energy' and 'Neutrino' actually form the centre of the structure. The choice of the two resolution levels resulted in a hierarchic structure confirming the appropriateness of the applied method.

Acknowledgements

This work has been done as part of the international project 'Measuring the Diversity of Research' and in the framework a special workshop on the comparative analysis of algorithms for the identification of topics in science organised in Berlin in August 2014. The project and workshop series was jointly organised by the Humboldt Universität and Technische Universität Berlin. We would like to acknowledge their support of our study.

References

Batagelj, V. & Mrvar, A. (2003). *Pajek–Analysis and visualization of large networks*. In: M. Jünger & P. Mutzel (Eds.), Graph drawing software (pp. 77–103). Berlin: Springer.

Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008.

Boyack, K. W. & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404.

Glänzel, W. & Czerwon, H.J. (1996), A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, *37*(2), 195–221.

Glänzel, W. & Thijs, B. (2011), Using `core documents' for the representation of clusters and topics. *Scientometrics*, *88*(1), 297–309.

Glänzel, W. & Thijs, B. (2012), Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics*, *91*(2), 399–416.

Glänzel, W. (2012), The role of core documents in bibliometric network analysis and their relation with h-type indices. *Scientometrics*, *93*(1), 113–123.

Thijs, B., Schiebel, E. & Glänzel, W. (2013), Do second-order similarities provide added-value in a hybrid approach? *Scientometrics*, *96*(3), 667–677.

Mining Scientific Papers for Bibliometrics: a (very) Brief Survey of Methods and Tools

Iana Atanassova[1], Marc Bertin[2] and Philipp Mayr[3]

[1] *iana.atanassova@univ-fcomte.fr*
Centre Tesniere, University of Franche-Comte, (France)

[2] *bertin.marc@gmail.com*
Centre Interuniversitaire de Rercherche sur la Science et la Technologie (CIRST),
Université du Québec à Montréal (UQAM), (Canada)

[3] *philipp.mayr@gesis.org*
GESIS, Leibniz Institute for the Social Sciences (Germany)

Introduction

The Open Access movement in scientific publishing and search engines like Google Scholar have made scientific articles more broadly accessible. During the last decade, the availability of scientific papers in full text has become more and more widespread thanks to the growing number of publications on online platforms such as ArXiv and CiteSeer (Wu, 2014). The efforts to provide articles in machine-readable formats and the rise of Open Access publishing have resulted in a number of standardized formats for scientific papers (such as NLM-JATS, TEI, DocBook).

Corpora

Different projects have been carried out to respond to the need of full-text datasets for research experiments (PubMed, JSTOR, etc.) and corpora.
E.g. the *iSearch* dataset was designed to facilitate research and experimentation in information retrieval, and specifically in aspects of task-based and integrated (a.k.a. aggregated) search. Its compressed size is about 46GB of documents in English from the physics domain that were collected from public libraries and open archive resources.

Semantic Web and Information Retrieval

Scientific papers are highly structured texts and display specific properties related to their references but also argumentative and rhetorical structure. Recent research in this field has concentrated on the construction of ontologies for citations and scientific articles.
CiTO (Shotton, 2010), the Citation Typing Ontology, is an ontology for the characterization of citations, both factually and rhetorically. It is part of SPAR, a suite of Semantic Publishing and Referencing Ontologies. Other SPAR ontologies are described at http://purl.org/spar/.

Statistical Analysis of Textual Data

Text Mining in R

Temis, an R Commander plugin (Bastin, 2013) provides integrated tools for text mining. Corpora can be imported in raw text. Another package is IRaMuTeQ (Ratinaud, 2009), a python application which uses the R libraries.

Correspondence Analysis

Correspondence analysis is a technical description of contingency tables and is mainly used in the field of text mining (Morin, 2006).
These tools could be very useful on the perspectives for the development of new text analytics approaches for bibliometrics.

Natural Language Processing Tools

Research in the field of Natural Language Processing (NLP) has provided a number of open source tools for versatile text processing.
The Apache *OpenNLP* library (Baldridge, 2005) is a machine learning based toolkit for the processing of natural language text. Written in Java, it is open source and platform-independent.
Stanford *CoreNLP* (Manning, 2014) integrates many NLP tools, including a part-of-speech (POS) tagger, a named entity recognizer (NER), a parser, a coreference resolution system, a sentiment analysis tool, and bootstrapped pattern learning tools. Stanford CoreNLP is written in Java and licensed under the GNU General Public License.
MALLET (McCallum, 2002) is a Java-based package for statistical NLP, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text. It includes sophisticated tools for document classification: efficient routines for converting text to "features", a wide variety of algorithms (including Naïve Bayes, Maximum Entropy, and Decision Trees), and code for

evaluating classifier performance using several common metrics.

GATE (Cunningham, 2002) is open source free software for all types of computational tasks involving human language. It includes components for diverse NLP tasks, e.g. parsers, morphology, tagging, Information Retrieval tools, Information Extraction components for various languages.

CiteSpace (Chen, 2006) is a freely available Java application for visualizing and analyzing trends and patterns in scientific literature. It is designed to answer questions about a knowledge domain, which is a broadly defined concept that covers a scientific field, a research area, or a scientific discipline.

What is next?

Several studies examine the distribution of references in papers (Bertin, 2013). However, up to now full-text mining efforts are rarely used to provide data for bibliometric analyses. An example is the special issue on Combining Bibliometrics and Information Retrieval (Mayr, 2015). Novel approaches to full-text processing of scientific papers and linguistic analyses for Bibliometrics can provide insights into scientific writing and bring new perspectives to understand both the nature of citations and the nature of scientific articles. The possibility to enrich metadata by the full-text processing of papers offers new fields of application to bibliometrics studies like e.g. text reuse patterns in specific disciplines.

Working with full text allows us to go beyond metadata used in Bibliometrics. Full text offers a new field of investigation, where the major problems arise around the organization and structure of text, the extraction of information and its representation on the level of metadata. Unlike text-mining from titles and abstracts, full-text processing allows the extraction of rhetorical elements of scientific discourse, such as results, methodological descriptions, negative citations, discussions, etc. Scientific abstracts, by summarizing the text, provide only short, synthetic and thematic information.

Furthermore, the study of contexts around in-text citations offers new perspectives related to the semantic dimension of citations. The analyses of citation contexts and the semantic categorization of publications will allow us to rethink co-citation networks, bibliographic coupling and other bibliometric techniques.

Our aim is to stimulate research at the intersection of Bibliometrics and Computational Linguistics in order to study the ways Bibliometrics can benefit from large-scale text analytics and sense mining of scientific papers, thus exploring the interdisciplinarity of Bibliometrics and Natural Language Processing. Typical questions of this emerging field are: How can we enhance author network analysis and Bibliometrics using data obtained by text analytics? What insights can NLP provide on the structure of scientific writing, on citation networks, and on in-text citation analysis?

References

Baldridge, J. (2005). The Apache OpenNLP library. https://opennlp.apache.org/

Bastin, G., Bouchet-Valat, M. (2013). RcmdrPlugin. temis, a Graphical Integrated Text Mining Solution in R. *The R Journal 5*(1): 188–196

Bertin, M., Atanassova, I., Larivière, V., Gingras, Y. (2013). The Distribution of References in Scientific Papers: an Analysis of the IMRaD Structure. *Proceedings of the 14th ISSI Conference (ISSI-2013)*, Vienna

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JASIST, 57*(3): 359-377

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: an architecture for development of robust HLT applications. *Proceedings of the 40th Annual Meeting of the ACL*, pp. 168–175

Lykke, M., Larsen, B., Lund, H. & Ingwersen, P. (2010). Developing a Test Collection for the Evaluation of Integrated Search. Advances in Information Retrieval. *32nd European Conference on IR Research*, UK.

McCallum, A.-K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J. & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the ACL: System Demonstrations*, pp. 55-60

Mayr, P. & Scharnhorst, A. (2015). Scientometrics and Information Retrieval - weak-links revitalized. *Scientometrics, 102*(3): 2193-2199

Morin, A. (2006). Intensive use of factorial correspondence analysis for text mining: application with statistical education publications. *Statistics Educational Research Journal (SERJ)*

Ratinaud, P. (2009). IRaMuTeQ:Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires, http://www.iramuteq.org

Shotton, D. (2010). CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics, 1* (Suppl 1), S6.

Wu J., Williams K., Chen H.-H., Khabsa M., Caragea C., Ororbia A, Jordan D., Giles C. L. (2014). "CiteSeerX: AI in a Digital Library Search Engine," *Innovative Applications of AI, Proceedings of the 28th AAAI Conference*, pp. 2930-2937

A Multi-Agent Model of Individual Cognitive Structures and Collaboration in Sciences

Bulent Ozel

ozel@uji.es
Universitat Jaume I, Department of Economy, Castellon De La Plana (Spain)

Motivation

This research takes a multi agent perspective while simulating knowledge diffusion mechanism in science. Multi agent systems are systems that are composed of a large number of autonomous agents that are capable of interacting with each other. The autonomous agents are not controlled by a central mechanism, instead, their decision taking logics are part of their actions and they are decentralized, hence, they are able to make decisions in order to accomplish individual tasks (Wooldridge, 2009). In this research, a scientist who is situated within a co-authorship network is considered as an individual autonomous agent. Her decision process at picking another scientist to co-author a paper and outcome of such an interaction builds up our multi-agent system.

In a science network, if two scientists work on the same paper, then they are considered connected. The social interaction linkage between them is a possible channel for knowledge diffusion. In our model, each author is considered as an agent that is capable of working with other authors, choosing whom to work with and what subject to work on. In order to set-up initial environment of our multi-agent system we need to identify initial co-authorship network, as well as, we need to represent knowledge space of each individual author in the network. In order to capture a representation of an individual's expertise a set of keywords, which is driven from publications of the author is used to form the node set of the semantic network of that very individual. The semantic relations, namely the links, in between the keywords in the set are established by their co-occurrence on a published article.

There are a number of challenges at designing interaction and evolution of such multi agent system. The challenges are (i) being able to incorporate a dynamic social network perspective while modelling interactions in between agents, (ii) designing, simulating and examining various knowledge creation and diffusion mechanisms as the outcomes of agent-agent interactions.

The first challenge addresses a problem within multi-agent modelling research area. Computational simulation of social systems falls short at covering dense and multitude interactions in between actors. Majority of agent-agent interactions are implicitly and limitedly modelled via agent-agent interactions using environmental variables. This limitation is partly due to complexities at agent-agent interactions and mainly due to lack of empirically validated interaction mechanisms. In this work, we borrow and adopt models from social network literature. More specifically, we examine co-authorship networks and empirically validated interaction models within the field.

In the second challenge, we take a socio-cognitive approach. We model and exploit cognitive structure of each agent both at the incentives of individuals to select other agents to collaborate and at modelling the outcome of resulting interactions. Namely, agents purposefully interact to create and transfer new knowledge.

In addition to challenges mentioned above there are several implementation challenges to be addressed for the simulation model. First of all, not all agents in the population interact with each other at each run and preferences of interaction cannot be uniformly random. In the model, those ones who decide to collaborate compute the set of candidate collaborators autonomously. An agent's current knowledge space, and his/her ego network are taken into consideration at incentives to collaborate. For instance, literature suggests that repetition of joint collaborations follows a power law distribution (Morris & Goldstein, 2007) mimicking power law distribution of individual publication productivity. Likewise, propensity to collaborate with collaborator of an existing co-author is incorporated adopting transitivity property of social ties (Wellman, 1988). Another empirically validated model of social tie formation mechanism that is adopted is "preferential attachment". It is known that in a complex social network probability of a node to have a new connection is proportional to the connections it already has (Barabasi, 2002). At each round of the simulation each agent independently determines a candidate set of collaborators. This candidate set is formed employing above-mentioned mechanisms.

A second implementation challenge is how to incorporate knowledge of individual agents. Dynamic social network mechanism does not take actual knowledge space of individual into consideration. In other words, knowledge space of individuals does not play a direct role on the interactions. Besides, while social interaction mechanisms hint whom to pick to collaborate it

does not explain outcome of interactions. It is necessary to come up with empirically validated and sound models to represent what knowledge will be exchanged as the outcome of such social interactions.

Literature suggests that there are two competing social mechanisms, which may help to consider cognitive structure of individuals on the preferences of collaborators. They are 'cognitive distinctiveness' and 'cognitive similarity'. Cognitive distinctiveness or cognitive similarity of two agents is measured by comparing their knowledge bases. For a pair of agents when the distinctiveness is high then there are more possibilities for them to learn from each other. If their knowledge bases overlaps widely, the knowledge they can get from each other is limited (Carley, 1991). However, it is known that people, in some cases, tend to interact with people they are similar to; a tendency, which is known as homophily (McPherson et al., 2001). The experiments are devised to observe impact of these two competing models.

Implementation

As we have mentioned above, each author is represented as an agent. Each agent has its own individual memory, where its knowledge base and its co-authorship history is kept and updated throughout the simulation. Knowledge base of an agent is formed by set of keywords based on agent's publication records. This set of keywords is interrelated to each other. It is represented by a symmetric matrix. The matrix is a representation of cognitive structure of an agent. The entries of the matrix encode co-occurrence frequency of respective keywords. Co-authorship memory of an agent is a set of authors with whom the agent worked with on a publication.

Set of all the keywords that are gathered from all of the publications is represented as a weighted graph. If two keywords belong to the same publication, then they have a connection and weight of the connection is the number of the times they are used together. When entire set of publications for all agents is considered, then this graph is the cognitive structure of the entire network and it will be represented as an environmental component in the simulation.

It is certain that real agents learn from each other via collaboration, but this is not the only way of learning new things. They also learn from their readings, the workshops they attend and many other resources, etc. In order to represent all such various source of knowledge accumulation by agents, knowledge injection method is used. At each simulation time point, which is set as a year, a set of new keywords is added to the cognitive structure of entire population. A probabilistic model is adopted to update cognitive structures after injection of new keywords to the set. Betweenness centrality of existing keywords is used. The higher betweenness of a keyword, the higher chance it receives a new link.

Initial Findings and Future Work

Results from our initial experiments hint that in scenarios where agents are inclined to collaborate with cognitively dissimilar agents, then resulting collaboration structure rather mimics co-authorship relations seen within a research center. On the other hand, when cognitive similarity leads the incentives to pick a collaborator, then resulting co-authorship rather mimics network structures observed within domain of a journal in a field.

A large set of experiments is to be conducted to fully verify and validate our initial results, as well as, to discuss challenges addressed above.

There are a number of additional implementation challenges, which will be addressed and attempted as part of this ongoing research. They are (i) how to model when and in what circumstances multiple co-authorship occurs; (ii) at each run, not only new knowledge pieces but also new agents will be injected to the simulation. Knowledge base of those new agents will be composed of partially by a subset of keywords that is already in the current set and partially by new keywords that is not in the set. This approach will mimic arrival of new scientists in a field.

Bibliography

Barabasi, A. L., H. Jeong, Z. Neda, E. Ravasz, A. Schubert, & T. Vicsek. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications 311*(3-4), 590–614.

Carley, K. (1991). A theory of group stability. *American Sociological Review 56*, 331–354.

McPherson, M., Smith-Lovin, L. & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks, *Annual Review of Sociology 27*, 415-444.

Morris, S.A., & Goldstein, M.L. (2007). Manifestation of research teams in journal literature: a growth model of papers, authors, collaboration, coauthorship, weak ties, and Lotka's law. *JASIST*, *58*(12), 1764-1782.

Wellman, B. (1988). *Social Structures a Network Approach*, 19–61. Cambridge University Press.

Wooldridge, M. (2009). *Introduction to Multi-agent Systems*. John Wiley & Sons.

Hypothesis Generation for Joint Attention analysis on Autism

Jian Xu[1], Ying Ding[2], Chaomei Chen[3], Erjia Yan[3]

[1] *issxj@mail.sysu.edu.cn*
School of Information Management, Sun Yat-sen University, Guangzhou, Guangdong (China)

[2] *dingying@indiana.edu*
Department of Information and Library Science, Indiana University, Bloomington, Indiana (USA)

[3] *chaomei.chen@ drexel.edu and erjia.yan@drexel.edu*
College of Computing and Informatics, Drexel University, Philadelphia, Pennsylvania (USA)

Introduction

Every 20 minutes a new case of autism is diagnosed worldwide, which affects around 6% of the population of children. One of the major challenges in autism is how to reliably diagnose autism as early as possible so that early intervention can be imposed to dramatically change the whole situation, even lead to cure. Joint attention is among these early impairments that distinguish young kids with autism from normal kids. Joint attention is a transdisciplinary area which was studied in robotics, psychology, autism, and neuroscience. However, Due to the unaware of similar or related researches in different domains, researchers are unknowingly duplicating studies that have already been done elsewhere. On the other hand, due to the lack of domain knowledge in other domains, researchers can experience difficulties to understand the advances in other domains. To deal with this dilemma, generating hypotheses is considered a potentially effective way. It is a crucial initial step for scientific breakthroughs, and usually relies on prior knowledge, experience and deep thinking. Especially for transdisciplinary domains, generating hypothesis from literature in different but related disciplines can be exciting and highly demanded because it is no longer possible for domain experts in one domain to fully master the knowledge in another domain.

Although marked with several decades of research history, it is until recent years that hypotheses generating attracts more attention in transdisciplinary research domains. Swanson (1986) proposed ABC model to inference the literature-based hypotheses. Later on, Srinivasan (2004) presented open and closed text mining algorithms that are built within the discovery framework established by Swanson and Smallheiser. Their algorithms successfully generated ranked term lists where key terms representing novel relationships between topics are ranked high. Zhang et al. (2014) established the semantic Medline which biomedical entities and association are semantically annotated using concepts in UMLS. They assumed that the network motifs in the network can represent basic interrelationships among diseases, drugs and genes and reflect a framework in which novel associations can be derived as hypotheses to be further validated by domain experts. Spangler et al. (2014) presented a prototype system KnIT, which can mine the information contained in the scientific literature and represent it explicitly in a queriable network, and then further reason upon these data to generate novel and experimentally testable hypotheses. They applied their method to mine the publications related to p53 (a protein tumor suppressor) and are able to identify new protein kinases that phosphorylate p53. Malhotra et al. (2013) proposed a pattern matching approach for the detection of speculative statements in scientific text that uses a dictionary of speculative patterns to classify sentences as hypothetical. Their application on the domain of Alzheimer's disease showed that the automated approach captured a wide spectrum of scientific speculations and derived hypothetical knowledge leads to generation of a coherent overview on emerging knowledge niches. Song et al. (2007) constructed a Gene-Citation-Gene (GCG) network of gene pairs implicitly connected through citation and indicated that the GCG network can be useful for detecting gene interaction in an implicit manner. In this initiative, we use text mining approach to analyze related publications on joint attention from robotics, psychology, autism and neuroscience, to generate hypotheses which will be tested in the lab which collects eye contact and body movement sensor data. Here some preliminary results were reported and discussed.

Methodology

Due to the transdisciplinary character of "joint attention" research, we elaborately selected eight data sources (Wiley Online Library, ProQuest PsycINFO, Science Direct, Scopus, Web of Science, PubMed Central, Springer Link and Google Scholar) to maximize the coverage of the final dataset. The phrase "joint attention" is used to search separately on each data source.

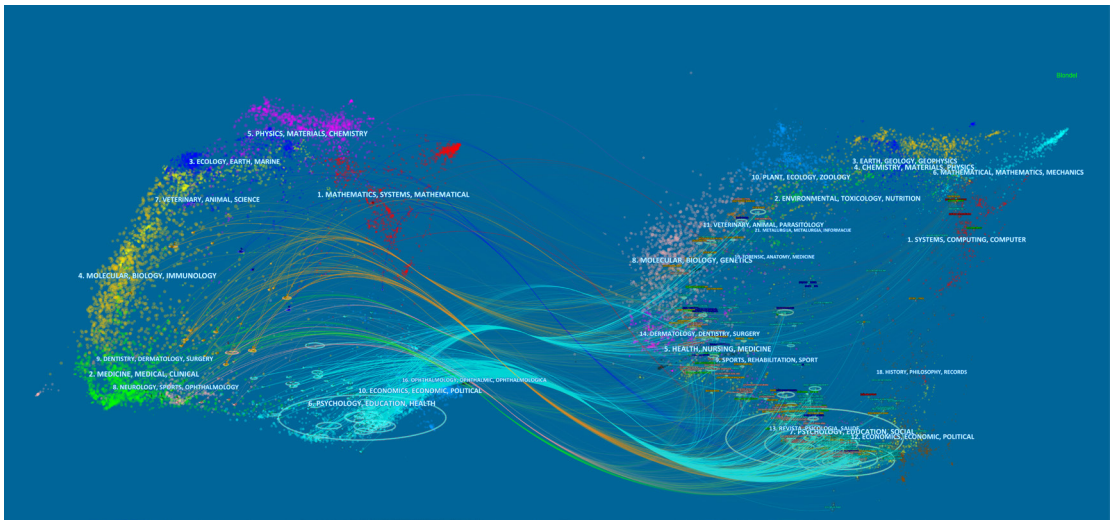


Figure 1. A dual-map overlay of "joint attention" search result from Web of Sciences.

Under the different download limitations, there are totally 39,845 records downloaded and 6,660 records left after remove duplicate records by the field "title". In the next step, keywords of each article in the dataset were extracted by using TF-IDF method. Then based on Keywords and other fields such as "journal name" and "citations", clustering were processed and relations among different clustering were analysed. By drawing the overall "research topic map", we can easily distinguish hot topics and their connections, and get to know their locations on the overall map. Then different dimensions (e.g., age, speech, language, and communication) were defined to analyse the distribution of current researches. Finally, from different dimension analysis aspects, research blind points were uncovered and new hypotheses were inferred, which will be tested in the lab.

Preliminary results

We tested a Web of Science query of "joint attention" (1,479 records) as a single dual-map overlay (Figure 1). Figure 1 shows the distribution of citing papers (left part) and cited papers (right part). Visualizations at this level are between journals, journal clusters, and overall maps. From the citation distribution and clustering results, we can identify the overall distribution of relevant sources and the most relevant targets (both ends with reference arcs). The label clustering result shows that the most popular domain discussing "joint attention" are Psychology, Education, Health, Medicine, Molecular, Economics, Mathematics, and Biology. It suggests that the Web of Science data is overwhelmingly dominated by a single journal *Journal of autism and developmental disorders*, with 169 papers. On the cited side, it is also the most cited journal in the dataset (6,640 citations). Other highly cited journals include *Child Development* (3,581 cites) and *Developmental Psychology* (2,328 cites).

Conclusions

This paper reports the ongoing effort on generating hypotheses in the transdisciplinary area of the joint attention research. We downloaded data from 8 separate data sources to maximize the coverage of "joint attention" related researches. Then text mining and visualization approaches were used to analyze related publications. Later stages of this research will generate hypotheses, which will be tested in the lab based on current research distributions on different predefined dimensions.

References

Swanson, DR. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med*, *30*(1), 7–18.

Srinivasan, P. (2004). Text mining: Generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, *55*(5), 396-413.

Zhang, Y., Tao, C., Jiang, G., Nair, A.A., Su, J., et al. (2014). Network-based analysis reveals distinct association patterns in a semantic Medline-based drug-disease-gene network. *Journal of Biomedical Semantics*, 5:33.

Spangler, S., Wilkins, A.D., Bachman, B.J., Nagarajan, M., Dayaram, T., et al. (2014). Automated hypothesis generation based on mining scientific literature. *20th ACM SIGKDD* (pp. 1877-1886). New York:ACM

Malhotra, A., Younesi, E., Gurulingappa, H., & Hofmann-Apitius, M. (2013) 'HypothesisFinder:' A Strategy for the detection of speculative statements in scientific text. *PLoS Comput Biol*, *9*(7): e1003117.

Song, M., Han, N.G., Kim, Y.H., Ding, Y., & Chambers, T. (2014) Correction: Discovering Implicit Entity Relation with the Gene-Citation-Gene Network. *PLoS ONE*, *9*(1).

Chen, C., & Leydesdorff, L. (2014) Patterns of connections and movements in dual-map overlays: A new method of publication portfolio analysis. *Journal of the American Society for Information Science and Technology*, *65*(2), 334-351.

"What Came First – *Wellbeing* or *Sustainability*?" A Systematic Analysis of the Multi-dimensional Literature Using Advanced Topic Modelling Methods

Mubashir Qasim[1] and Les Oxley[1]

[1] *mq21@students.waikato.ac.nz,* [1]*loxley@waikato.ac.nz*
Waikato Management School, University of Waikato, Hamilton, New Zealand

Introduction

Both sustainability and well-being (SaW) are inter-dependent, inter-disciplinary, multi-dimensional, and international subject areas. However, people tend to interpret the subjects significantly differently based on their professional affiliation, academic background, geographical location etc., (Brunn, 2014; Roberts et al., 2013). A search of the SaW literature, using any scholarly search engine, generates results ranging from the thousands to millions creating a challenge for the researcher in picking the right papers; constructing a reasonable structure and synthesizing the vast material in order to conduct a comprehensive review of the literature. The work presented here relates to the use of a sophisticated method to exploit the explanatory power of metadata, attached to the results of a search query, to identify hidden patterns in the universe of given articles. The methods and metadata used to conduct the systematic analysis are briefly discussed under following headings.

Components of systematic literature analysis

Acquisition of data

Our quest begins with the analysis of key characteristics of metadata obtained from JSTOR *Data for Research* (DFR), which enables exploration of >9.2 million articles. We collected and analysed the metadata for a sample of 68,817 papers from DFR which related to SaW for this exercise. Metadata were generated against four queries with different sets of keywords as listed in Table 1. Analysis of the metadata was conducted in three steps: Step 1., analysis of keywords, subject and subject groups, disciplines and discipline groups, journals, authors and trends of publications (as presented in a recent study by (Brunn, 2014) but with slightly different approach). In Step 2., we applied the Latent Dirichlet Allocation (LDA) to study language differentiation between SaW themes. The main aim of this exercise was to identify complex hidden patterns in the data and present them in easily understandable ways. In Step 3., we used a reference manager software package called *Qiqqa* to identify key themes in the personal library and to identify seminal and frontier studies within each theme using cross references in the collection.

Table 1: Detail of search queries.

Query	Results	Search keywords	Search in
A	4,903	wellbeing OR well-being	Abstract
B	57,681	sustainability OR sustainable development	Title
C	5,472	sustainability; sustainable development; wellbeing; well-being	Any
D	761	sustainability OR sustainable development; well-being OR wellbeing	Abstract

Analysis of keyterms

We sampled 300 top keywords appearing in the corpus of each query to represent the frequently used language patterns in the subjects of SaW. The results are presented in the form of word-clouds in which the terms with high frequencies of occurrence are represented by the larger size of the word. Each word in the cloud indicates a dimension or issue in a subject (Jaewoo & Woonsun, 2014). Broadly discussed dimensions in the well-being literature include income, health, relationships, family, child, psychology etc., are correctly identified in our word-clouds.

Type of journals and subject group

Inter-relatedness of the SaW literature is established by confirming the large number of journals shared by SaW papers as suggested by (Mimno, 2012). Here, we extracted the names of the top 20 journals by number of articles in each query. Our analysis validates the assumption that many journals include papers on both aspects of the SaW literature. The interdisciplinary nature of the SaW literature is further established by similar categorization of SaW papers with respect to different subject groups.

Trends in publications

Many modern databases are devoted to tracking publications e.g., as Google Scholar, ISI Web of Science, JSTOR, SCOPUS, etc., and enable

scholars to perform quick and broad browsing of the literature (Hood & Wilson, 2003). Their expansions or contractions over time can indicate the interest of scholars in an area and the evolution of novel approaches (Adam, 2002; Casagrandi & Guariso, 2009).

In our analysis, we find the first article related to Query A, appears in 1919 and the number of publications remains trivial until the 1970's. Thereafter, a huge influx of papers begins in the late 1970's with 30 papers per year, peaking at 311 papers in 2012. In contrast, papers related to sustainability in Query B started much earlier with the first paper published in 1800. This number reaches to 50 papers per year in the next 100 years and steadily increase thereafter for another 50 years to around 250 papers per year in 1950. Post-1950, the number of scholarly articles grew five fold over the next five decades and peaked in 2005 at 1304 papers per year. Articles related to both SaW in Query C emerge in the late 1970's and grow exponentially over the next 40 years. As Query D is a subset of Query C they exhibit similar trends. A comparison of these trends with the papers in the entire DRF corpus of 9.3 million articles indicates the level of interest of the scholars over different years.

Authors of publications and places

Another way to consider the SaW literature is to analyse the country of the main author(s) of an article in order to answer the key question "what countries are leading the SaW agenda?" We select the top 20 authors in each set of documents based on their number of publications. Their country is established from the place of their affiliation at the time of publication. Our results show 74 unique authors from 12 different countries wrote 1,869 SaW paper. Not unexpectedly, 9 of these countries are developed OECD countries with the United States the home of 61% of SaW authors and 29% of this literature is produced by people from Europe, Canada and South Africa and rest of them are from Australia, India and Botswana.

Differentiating language using LDA

Finally, we conducted probabilistic analysis of the SaW literature using Latent Dirichlet Allocation (LDA) in order to establish underlying topics within the corpus of documents in each query (a topic is a set of co-occurring words). Our analysis helps understanding what sort of language is used within and across disciplines; what clusters of words happen to occur together; and how the use of language changes overtime. Results are shown by java based interactive visuals made in the programing language R. Each topic provides a clear structure to build a paragraph in a literature review and the cluster of topics gives a clear indication of the categories/themes within each set of documents.

Identification of seminal and frontier studies

Most dominant papers in our set of documents are identified using in-bound references assuming that heavily cited and highly ranked articles are the key papers in each collection. Identification of these articles provides the best starting point to begin the traditional literature review with. We used network diagrams using a reference manager called Qiqqa to conduct this exercise.

Validation of results

The results are validated using the metadata from another widely used scholarly source called Web of Science. Most of our results exhibit the same characteristics as the results of DFR data.

References

Adam, D. (2002). Citation analysis: The counting house. *Nature*, *415*, 726–729. doi:10.1038/415726a

Brunn, S. D. (2014). Cyberspace Knowledge Gaps and Boundaries in Sustainability Science: Topics, Regions, Editorial Teams and Journals. *Sustainability*, *6*(10), 6576–6603. doi:10.3390/su6106576

Casagrandi, R. & Guariso, G. (2009). Impact of ICT in Environmental Sciences: A citation analysis 1990–2007. *Environmental Modelling & Software*, *24*(7), 865 – 871. doi:http://dx.doi.org/10.1016/j.envsoft.2008.11.013

Hood, W. & Wilson, C. (2003). Informetric studies using databases: Opportunities and challenges. *Scientometrics*, *58*(3), 587–608. Kluwer Academic Publishers. doi:10.1023/B:SCIE.0000006882.47115.c6

Mimno, D. (2012). Computational Historiography: Data Mining in a Century of Classics Journals. *J. Comput. Cult. Herit.*, *5*(1), 3:1–3:19. New York, NY, USA: ACM. doi:10.1145/2160165.2160168

Roberts, L., Brower, A., Kerr, G., Lambert, S., McWilliam, W., Moore, K., Quinn, J., et al. (2013). A Good Life: How nature's ecosystem services contribute to the wellbeing of New Zealand and New Zealanders. Department of Conservation.

Jaewoo, C. & Woonsun, K. (2014). Themes and Trends in Korean Educational Technology Research: A Social Network Analysis of Keywords . *Procedia - Social and Behavioral Sciences, 131*(0), 171 – 176. doi:http://dx.doi.org/10.1016/j.sbspro.2014.04.099

Multi-Label Propagation for Overlapping Community Detection Based on Connecting Degree

Xiaolan Wu[1] and Chengzhi Zhang[2]

[1]*wuxiaolananhui@163.com*, [2]*zhangchz@istic.ac.cn*
Dept. of Information Management, Nanjing University of Science and Technology, Nanjing 210094 (China)

Introduction

With the growth of social media, social network analysis draws a great attention and becomes a hot research topic in the field of complex network, web mining, information retrieval, etc. An important aspect of social networks analysis is community structure (Newman, 2003).

In general, community detection methods are classified into two categories: overlapping methods (and non-overlapping methods (Hofman & Wiggins, 2008)). The former allows communities overlap, while the latter assumes that a network only contains disjoint communities. In this paper, we focus on the overlapping community detection. To find overlapping community, researchers use a wide variety of techniques, such as Clique Percolation Method, COPRA (Gregory, 2010), etc. COPRA is very fast, but the result of COPRA is nondeterministic, so we propose an improved COPRA with high determinacy in this paper.

An Improved COPRA Algorithm Based on Connecting Degree

To eliminate the nondeterministic of COPRA, we use Connecting Degree as definition 1.

Definition 1: Let v be a node on the undirected Graph $G(V;E)$, C is the set of overlapped communities on Graph, the connecting degree between node v and community $c(c \in C)$, denoted $C(v,c)$, be computed by the following formula (Duanbing, Mingsheng, Xia, 2013).:

$$C(v,c) = \frac{\sum_{u \in c} w_{vu}}{k_v} \qquad (1)$$

Where k_v is the degree of node v, w_{vu}=1 if there is an edge between node v and node u, zero otherwise. *

Connecting Degree can reflect the community tendency for a node to its neighbour communities, so we proposed a COPRA Based on Connecting Degree, named COPRA-CD. COPRA-CD works as follows: 1) To start, all nodes are initialized with a unique community identifier and a belonging coefficent setting to 1; 2) Each node updates its community identifier by the union of its neighbours labels, the corresponding belonging coefficient is obtained by normalizing the sum of the belonging coefficients of the communities over all neighbours. Then, comparing all the belonging coefficients and the parameter v, if all the belonging coefficients are less than v, calculating the connecting degree between node and its neighbour community, then only retain neighbour community with greatest connecting degree, else keeping these belonging coefficients that are more than v, then renormalize these belonging coefficients of remaining communities so that they sum to 1. After several iterations, if the stop criteria proposed by Gregory is satisfied, the propagation procedure stops; 3) Remove communities that are totally contained by others; 4) Split disconnected communities.

Experimental Results and Discussion

Test networks

At first, we do experiments on four real-world networks, whose information are shown in Table 1.

Table 1. General information of real networks

Networks	Description	Node&Edge
Karate	Zachary's karate club (Zachary, 1977)	34 &78
Dolphin	Lusseau's Dolphins (Lusseau, 2003)	62 & 159
Books	Books about US politics	105 & 441
Football	American College football union (Girvan, Newman, 2002)	115 & 616

Then we also test the performance of COPRA-CD on six LFR synthetic networks with various mixing parameter μ ranging from 0.1 to 0.6, the other standard configuration of LFR synthetic network used in this experiment is: n=1000, t_1=2, t_2=1, k=10, $\max k$=30, $\min c$=10, $\max c$=50, O_n=100, O_m=2.

Test metrics

To measure overlapping communities detection, Q_{ov} was be proposed by Nicosia et al (2009). The formulation of Q_{ov} as following:

$$Q_{ov} = \frac{1}{m} \sum_{c \in C} \sum_{i,j \in V} \left[\beta_{l(i,j),c} A_{ij} - \frac{\beta^{out}_{l(i,j),c} k_i^{out} \beta^{in}_{l(i,j),c} k_j^{in}}{m} \right] \qquad (2)$$

Where A_{ij} is the adjacency matrix of Direct Graph $G(E,V)$, C is the set of overlapped

* Corr. author: C. Zhang, Tel: +86-25-84315963.

communities, $l(i,j)$ is a link which starts at node i and ends at node j. $\beta_{l(i,j),c}$ is the belonging coefficient of $l(i,j)$ for community c, $\beta^{out}_{l(i,j),c}$ is the expected belonging coefficient of any possible link $l(i,j)$ starting from a node into community c, $\beta^{in}_{l(i,j),c}$ is the expected belonging coefficient of any link $l(i,j)$ pointing to a node going into community c. k_i^{out} is the out degree of node i, while k_j^{in} is the in degree of node j.

Test results and discussion

In order to show its performance, we compare three multi-label propagation algorithms, i.e., COPRA, COPRA-CD, and RC-COPRA. RC-COPRA stands for the version of COPRA with initialization using RC proposed by Wu et al. (2012). In our test, we run each algorithm 100 times on each network for the same value of parameter v. The average modularity result on real-world network was shown in Table 2, and the comparison performance on LFR synthetic networks was shown in Figure 1.

Table 2. Test Results on real-world Networks.

Networks	**COPRA (v=2)**	**COPRA-CD (v=2)**	**RC_COPRA (v=2)**
Karate	0.428	0.745	0.703
Dolphins	0.645	0.759	0.761
Books	0.826	0.815	0.830
Football	0.684	0.661	0.668
Networks	**COPRA (v=3)**	**COPRA-CD (v=3)**	**RC_COPRA (v=3)**
Karate	0.408	0.717	0.725
Dolphins	0.652	0.710	0.713
Books	0.830	0.822	0.827
Football	0.677	0.665	0.670

From Table 2, we find the modularity of CORPA is lower than that of other algorithms at the same v. At v =3, RC_COPRA algorithm gives better average modularity for every network, but at v=2, the modularity of RC_COPRA algorithm on Karate network is not better than that of COPRA-CD.

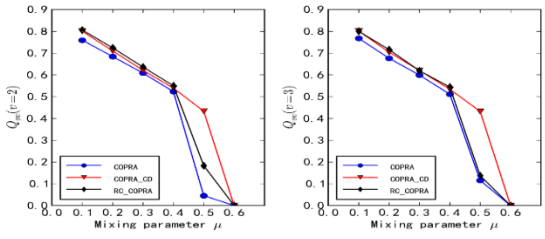


Figure 1. Experiment on synthetic networks.

As Figure 1 shows, when $\mu \leq 0.4$, all three algorithms show good performance. When $\mu =$ 0.5, LFR synthetic networks are very fuzzy, the overlapping community structure is not detected by COPRA and RC_COPRA, but detected by COPRA-CD, so we can conclude that for the given parameter, COPRA-CD is the most stable algorithm in these overlapping community detection algorithms.

Conclusions

In this paper, we propose COPRA-CD to uncover overlapping communities in social networks. Then we test it on four real-word networks and a group of synthetic networks. Experimental results show that both RC initialization and the connecting degree update strategy can bring improvements in quality, especially COPRA-CD has the best stability for fuzzy networks. In the future, COPRA-CD can be applied to analyze the community of co-author in paper.

Acknowledgments

This work is supported by the National Social Science Fund Project (grant number 14BTQ033).

References

Duanbing, C., Mingsheng, S., & Xia, L. (2013). Two-phase strategy on overlapping communities detection. *Computer Science, 40*(1), 225-228.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *PNAS, 99*(12), 7821-7826.

Gregory, S. (2010). Finding overlapping communities in networks by label propagation. *New Journal of Physics*, *12*(10), 103018.

Hofman, J. M., & Wiggins, C. H. (2008). Bayesian approach to network modularity. *Physical review letters, 100*(25), 258701.

Krebs, V. (2008). A network of co-purchased books about US politics, www.orgnet.com.

Lusseau, D. (2003). The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270 (Suppl 2), S186-S188.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167-256.

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical review E, 69*(6), 066133.

Nicosia, V., Mangioni, G., Carchiolo, V., & Malgeri, M. (2009). Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment,* (03), P03024.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*(7043), 814-818.

Wu, Z.-H., Lin, Y.-F., Gregory, S., Wan, H.-Y., & Tian, S.-F. (2012). Balanced multi-label propagation for overlapping community detection in social networks. *Journal of Computer Science and Technology, 27*(3), 468-479.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research,* 452-473.

Reproducibility, consensus and reliability in bibliometrics

Raul I. Mendez-Vasquez[1] and Eduard Suñen-Pinyol[2]

[1]raul.mendez@fundaciorecerca.cat, [2]eduard.sunen@fundaciorecerca.cat
Fundació Catalana per a la Recerca i la Innovació (FCRI), Bibliometrics. Pg. Lluís Companys, 23 E-08010 Barcelona (Spain)

Introduction

Bibliometrics, and scientometrics in general, have been enjoying what seems to be an endless party. Far from stopping, the demand for bibliometric indicators from governmental bodies, administrators and researchers, is continuously growing. During this "give me the indicators" phase several solutions have been provided by the community, let say new and more sophisticated indicators, which in turn geared the transition to the present "give me the indicators, but really?" phase. The impressive penetration of bibliometric indicators in decision making processes, some of which are crucial in the development of researchers' careers, has also brought the necessity for credibility on bibliometrics, and more specifically, on how it is practiced. Examples of improper use of bibliometric indicators have raised skepticism among users of bibliometric reports[1].

As a scientific discipline, bibliometrics is subject to the principle of replication and corroboration of results, just like any other discipline. Precisely, the credibility of scientists goes hand in hand with the reproducibility of their results.

The objective of this contribution is to bring attention to the importance of the reproducibility of the number of publications as an indicator of the quality of bibliometric reports.

Methods

We compared the numbers of publications estimated by three units following this schema: CTWS vs. BAC (us) and SCIMAGO vs. BAC. Sixteen universities reported in the CTWS Leiden Ranking 2011/2012, and 20 universities reported in the Iberoamerican Ranking SIR 2012 produced by SCIMAGO were selected for the study. Source, type of document, language and period were matched in each comparison. The numbers of publications produced by the BAC were sourced with the National Citation Report for Spain (NCR), an *ad hoc* database built in July 2012 as a live extraction from the Web of Science that compiles all the publications between 1970 and 2011, with at least one address in Spain. The unification was performed by hand based solely on the information contained in the address field of the NCR. Hierarchy relationships such as university campuses and institutes, affiliated hospitals, etc, were reconstructed in the system. All the addresses were also located to a specific administrative unit (a city in the majority of cases). Both, the information on the organizational hierarchy and location of the addresses were used to unify the name variants of subunits whenever mother organizations were not present in the addresses. Changes in the structure of the organizations within the analyzed period were recorded in the system. The unification terminated when a precision higher than 97% was achieved.

Results

A simple examination of the number of publications of a small set of universities revealed important reproducibility issues, even when controlling for source dataset, period of time and the document type (Table 1. several rows and columns were removed). A positive and statistically significant correlation ($p<0.01$) was observed between the numbers of publications produced by the three units (CTWS & BAC, *rho* 0.785; SCIMAGO & BAC, *rho* 0.860). The dispersion around the regression line was smaller in the comparison between SCIMAGO & BAC, than between CTWS & BAC, suggesting the presence of an outlier observation, whose removal increased the correlation between CTWS and BAC (*rho* 0.975, $p<0,001$). The concordance between the rankings produced by the three units was also positive and high, (CTWS & BAC, *tau* 0.733, $p<0.001$; SCIMAGO & BAC, *tau* 0.705, $p<0.001$). Removing the mentioned outlier observation increased the concordance between the CTWS and BAC (*tau* 0.905, $p<0.001$)

Discussion

These technical issues may explain the observed variability in the number of publications.

1) Completeness of the unification. The CTWS unit selected the universities with at least 500 publications per year and extended the unification to the name variants occurring at least five times in the source dataset. The BAC unit aims at attributing all variants to corresponding universities. However, mistakenly attributed name variants and non-identified variants were allowed to a maximum of 3%. The CTWS unit attributed the publications

[1]The title of a number of articles published in Nature in 2010 reflect this position: "Assessing assessment", "Do metrics matter?", "How to improve the use of metrics", "Let's make science metrics more scientific". Available at: http://www.nature.com/news/specials/metrics/index.html.

based on author names, a procedure not performed by the BAC. SCIMAGO provides no information on the unification in the website of the report.

Table 1. Differences in the number of publications produced by three units.

	(A)	(B)	A-B	(A-B)/A	(C)	(D)	C-D	(C-D)/C
UB	7,672	11,804	-4,132	-53,86	15,290	16,222	-932	-6,10
UAB	5,992	9,319	-3,327	-55,52	13,262	13,200	62	0,47
UCM	6,616	8,863	-2,247	-33,96	13,240	12,160	1,080	8,16
UPM	2,323	8,813	-6,490	-189,2	7,458	11,096	-3,638	-48,78
UAM	5,236	8,034	-2,798	-53,44	10,591	10,873	-282	-2,66
UV	5,077	7,892	-2,815	-55,45	11,191	10,458	733	6,55
UGR	3,966	5,918	-1,952	-49,22	9,128	8,117	1,011	11,08
USC	3,589	5,181	-1,592	-44,36	7,132	6,854	278	3,90
US	3,848	4,909	-1,061	-27,57	7,933	6,366	1,567	19,75
UPC	3,067	4,900	-1,833	-59,77	11,068	6,502	4,566	41,25
UZAR	3,394	4,612	-1,218	-35,89	7,607	6,102	1,505	19,78
EHU	3,047	4,536	-1,489	-48,87	7,520	6,535	985	13,10
n			**16**	**16**			**20**	**20**
Avg[1]			**-2,165**	**-51,40**			**659**	**7,30**
SDev.[2]			**1,508**	**-39,37**			**1,722**	**19,56**
CI[3]			**-739**	**-19,29**			**755**	**8,57**

A, data reported in the Leiden Ranking 2011/2012; **B**, number of publications estimated by BAC; **A-B**, magnitude of the difference between CTWS and BAC; (**A-B)/A**, percentage of change between CTWS and BAC; **C**, data reported in the Iberoamerican Ranking SIR 2012; **D**, number of publications estimated by BAC applying SCIMAGO criteria, but sourcing the analysis with the WOS; **C-D**; magnitude of the difference between SCIMAGO and BAC; (**C-D)/C**, percentage of change between SCIMAGO and BAC. 1; average; 2, standard deviation; 3, 95% confidence interval of the average. Acronyms: UB, Univ. de Barcelona; UAB, Univ. Autònoma de Barcelona; UCM, Univ. Complutense de Madrid; UPM, Univ. Politécnica de Madrid; UAM, Univ. Autónoma de Madrid; UV), Univ. de València; UGR, Univ. de Granada; USC, Univ. de Santiago de Compostela; US, Univ. de Sevilla; UPC, Univ. Politècnica de Catalunya; UZAR, Univ. de Zaragoza; EHU, Univ. del País Vasco.

2) Exactness of the unification. The CTWS unit estimated a 5% of false negative cases, while the BAC ensures a maximum percentage of error of 3%. SCIMAGO provides no information on this regard.

3) Proximity to the units under analysis. Two observations support the notion that local knowledge may explain a substantial part of the observed discrepancies: 1) the difference between SCIMAGO & BAC was smaller than between CTWS & BAC, and 2), SCIMAGO attributed more publications to their neighboring universities (UGR & US) than BAC, and vice versa in the case of the UB & UAB). A comparison of the number of publications of the Dutch universities between CTWS and BAC may shed some light on the effect that local knowledge or "regional peculiarities" (Moed, 1996) have on this indicator.

4) Delineation of the universities. The CTWS unit took into account "important university institutes" and changes in the structure of universities, while BAC took into account institutes, but also faculties, technical schools, locations, and structural changes. Failing to aggregate the publications of subunits could also explain the observed differences (de Mesnard, 2012).

5) Completeness and accuracy of the database (location of addresses). There is a difference between the sources used by the CTWS unit and BAC. The NCR may compile fewer records than the WOS, as addresses have to be located to Spain and errors are likely to happen during this process. This inconsistency may also play a lesser role in the comparison between CTWS and BAC.

Final considerations

Discrepancies in the number of publications of universities in the order of 10^2 or 10^3 are irrelevant when comparing the figures produced by different units. However, the magnitude of the difference might represent half of the output in some cases. Fortunately, the numbers of publications produced by the three units correlated pretty well, and the rankings were concordant. Technical issues can no longer be used as arguments to explain divergences of this magnitude, as none of the factors presented here are completely dependent on the technical capacity of a unit, rather than on procedural decisions: 1) completeness and 2) exactness of the unification, 3) knowledge of the surrounding environment, 4) completeness and accuracy of the source or 5) the type of document and period of time. The findings suggest that a consensus addressing these factors would do more in reaching a methodological "greatest common denominator" between the different units enabling improving the reproducibility of the indicators.

References

de Bruin R.E. & Moed H.F. (1990). The unification of addresses in scientific publications. In: Egghe L., Rousseau R. (editors.), *Informetrics*, 65-78.

Butler L. (1999). Who 'Owns' this Publication?, *Proc. ISSI*, 87-96.

de Mesnard L. (2012). On some flaws of university rankings: The example of the SCImago report. *The Journal of Socio-Economics 41*(5), 495–9

Moed H.F. (1996). Differences in the construction of SCI based bibliometric indicators among various producers: A first overview. *Scientometrics, 35*(2):177-191 DOI: 10.1007/BF02018476

Van Raan A. F. J. (2005a). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, *62*(1), 133-143.

Semantometrics: Fulltext-based Measures for Analyzing Research Collaboration

Drahomira Herrmannova[1] and Petr Knoth[2]

[1] *d.herrmannova@open.ac.uk,* [2]*petr.knoth@open.ac.uk*
Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes (United Kingdom)

Introduction

The aim of this article is to demonstrate some of the possible uses of a novel set of metrics called *Semantometrics* in relation to the role of "bridges" in scholarly publication networks. In contrast to the existing metrics such as Bibliometrics, Altmetrics or Webometrics, which are based on measuring the number of interactions in the scholarly network, Semantometrics build on the premise that full-text is needed to understand scholarly publication networks and the value of publications.

Up to date many studies of scientific citation, collaboration and coauthorship networks have focused on the concept of cross-community ties (Shi et al., 2010; Guimerà et al., 2005; Silva et al., 2014). It has been observed that in citation networks, bridging or cross-community citation patterns are characteristic for high impact papers (Shi et al., 2010). This is likely due to the fact that such patterns have the potential of linking knowledge and people from different disciplines. Likewise, in collaboration and coauthorship networks, it has been shown that newcomers in a group of collaborators can increase the impact of the group (Guimerà et al., 2005).

The studies up to date have been focusing on analysing citation and collaboration networks without considering the content of the analysed publications. Our work has focused on analysing scholarly networks using semantic distance of the publications in order to gain insight into the characteristics of collaboration and communication within communities. Our hypothesis states that the information about the semantic distance of the communities will allow us to better understand the importance and the types of the cross-community ties (bridges).

More specifically, in order to gain insight into the type of collaboration between authors we are currently investigating the possibility of utilising semantic distance in a coauthorship network together with the concept of *research endogamy*. In social sciences, endogamy is the practice or tendency of marrying within a social group. This concept can be transferred to research as collaboration with the same authors or collaboration among a group of authors. The concept of research endogamy has been previously used to evaluate conferences (Montolio et al., 2013) as well as journals and patents (Silva et al., 2014).

Furthermore, in (Knoth & Herrmannova, 2014) we have introduced and tested the first Semantometric measure which we call *contribution(p)* and which can be used to estimate research publication contribution. Our results suggested that measuring semantic similarity of publications can be utilised to provide meaningful information about the value of a research publication, which is not captured by traditional bibliometric measures.

Types of research collaboration in a coauthorship network

We are currently investigating the possibility of combining semantic distance and research endogamy in the publication's collaboration network. The rationale behind this approach is based on how research collaboration happens. In case the authors of a publication come from different disciplines, their research is likely to link the two disciplines and to build a bridge between them. This bridge can help to provide vision and ideas otherwise unseen and help to transfer knowledge between the disciplines.

We propose to measure the semantic distance of coauthors of a publication based on semantic distance of all pairs of the coauthors, where the distance of a pair of authors can be expressed similarly as the *contribution(p)* measure (Knoth & Herrmannova, 2014). This situation is depicted in Figure 1, where the sets A and B correspond to the publication records of the two authors.

Table 1. Types of research collaboration based on semantic distance and research endogamy.

	High endogamy	**Low endogamy**
High distance	Established interdisciplinary collaboration	New interdisciplinary collaboration
Low distance	Expert group	New expert collaboration

In order to distinguish between emerging, short-term and established research collaboration, we propose to combine the semantic distance with research endogamy value of the publication as defined in (Silva et al., 2014). We assume that based on the combination of semantic distance and research endogamy the types of research

collaboration can be divided into four groups (Table 1).

We believe this classification is a useful tool in characterising the types of research collaboration that goes beyond the traditional understanding of the concept of bridges as used in scholarly communication networks. While semantic distance allows distinguishing between inter- and intra-disciplinary collaboration, research endogamy allows differentiating between emerging and established research collaborations.

Using semantic distance to measure research contribution in a citation network

A similar Semantometric approach based on the concept of semantic distance can be applied in citation networks. We have used this approach in (Knoth & Herrmannova, 2014) to develop a measure which we call *contribution(p)*. This measure is based on a hypothesis, which states that the added value of publication *p* can be estimated based on the semantic distance from the publications cited by *p* to the publications citing *p*. This situation is depicted in Figure 1.

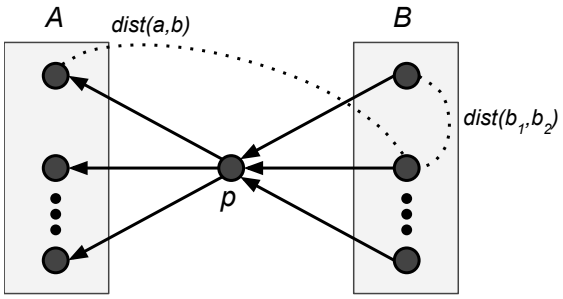


Figure 1. Explanation of contribution(p) calculation.

This hypothesis is based on the process of how research builds on the existing knowledge in order to create new knowledge on which others can build. A publication, which in this way creates a bridge between existing knowledge and something new, which will be developed based on this knowledge, brings a contribution to science. A publication has a high contribution if it connects more distant areas of science. Building on these ideas, we have developed a formula, which can be used for assessing research contribution of a publication. In order to adjust the contribution value to a particular domain and publication type, the metric uses a normalisation factor, which is based on the semantic distance of publications within the set of publications citing *p* and the publications cited by *p*. The measure and our experiments are in detail described in (Knoth & Herrmannova, 2014).

Conclusion

In this paper we proposed to apply the Semantometric idea of using full-texts to recognise types of scholarly collaboration in research coauthorship networks. We have applied semantic distance combined with research endogamy to classify research collaboration into four broad classes. This classification can be useful in research evaluation studies and analytics, e.g. to identify emerging research collaborations or established expert groups. Furthermore, we have presented another Semantometric measure, which we call *contribution(p)* and which is based on the idea of the importance of bridges in a citation network.

While bridges have been the concern of many research studies, their identification has been limited to the structure of the interaction networks. In contrast to these approaches, our approach takes into account both the interaction network (coauthorship, citations) as well as the semantic distance between research papers or communities. This provides additional qualitative information about the collaboration, which hasn't been previously considered.

References

Bornmann, L. & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1).

Guimerà, R., Uzzi, B., Spiro, J. & Nunes Amaral, L. A. (2005). Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, *308*(April), 697–702.

Knoth, P. & Herrmannova, D. (2014). Towards Semantometrics: A new Semantic Similarity Based Measure for Assessing a Research Publication's Contribution. *D-Lib Magazine*, *20*(11).

Montolio, S. L., Dominguez-Sal, D. & Larriba-Pey, J. L. (2013). Research Endogamy as an Indicator of Conference Quality. *SIGMOD Record*, *42*(2), 11–16.

Priem, J. & Hemminger, B. M. (2010). Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. *First Monday*, *15*(7).

Seglen, P. O. (1992). The Skewness of Science. *Journal of the American Society for Information Science*, *43*(9), 628–638.

Shi, X., Leskovec, J. & Mcfarland, D. A. (2010). Citing for High Impact. Proceedings of the 10th Annual Joint Conference on Digital Libraries - *JCDL '10* (p. 49). New York, New York, USA.

Silva, T. H. P., Moro, M. M., Silva, A. P. C., Meira Jr., W. & Laender, A. H. F. (2014). Community-based Endogamy as an Influence Indicator. *Digital Libraries 2014 Proceedings.* London, United Kingdom.

Uncovering the Mechanisms of Co-authorship Network Evolution by Multirelations-based Link Prediction

Jinzhu Zhang, Chengzhi Zhang, Bikun Chen

{zhangjinzhu, zhangcz,chenbikun}@njust.edu.cn
Nanjing University of Science and Technology, Dept of Information Management, Xiaolinwei Str 200, Nanjing (China)

Introduction and literature review

Co-authorship network, a proxy of research collaboration, reveals the collaboration patterns and the determining factors through social network analysis perspective, with nodes representing authors and links representing co-authorships (Ortega, 2014; Yan & Ding, 2009). If we know what mechanisms push the evolution of co-authorship network, we could predict which authors may collaborate in future.

Most of the studies correlate co-authorship evolution mechanisms to similarity indicators which quantitatively compared by link prediction in homogeneous network (Lu & Zhou, 2010). In order to integrate multirelations between authors, path-based similarity indicators are proposed for co-authorship prediction in DBLP heterogeneous network (Sun et al., 2011; Sun & Han, 2013). However, what is the role of each mechanism plays and how to combine multiple mechanisms to suit the co-authorship network evolution need to be clarified, moreover, the method need to be verified in different domains.

Therefore, we integrate similarity indicators based on multirelations in heterogeneous network and quantitatively evaluate them by link prediction justly, to uncover and infer the mechanisms of co-authorship network evolution. Firstly, similarities between authors are represented by a matrix where the rows are multirelations and the columns are multirelations' measures. Secondly, the evaluation of similarities is processed based on link prediction, to reveal the importance of each mechanism which is the weight for combining multiple mechanisms. Finally, experiments are presented in the domain of Library and Information Science (LIS), which reveals the best appropriate mechanism, the significance of each mechanism and the combination strategy of different mechanisms.

Data and method

Data

We collect the data from the SCIE (Science Citation Index Expanded) databases in Thomson Reuters' Web of Science, using journal publications on subject category of LIS across 2000 to 2009.

We choose the authors that the frequency greater than or equal to five as the experiment data, which includes 669 authors, 3,948 articles, 6,476 keywords, 14 subject categories, 29 journals and 79,717 references.

We eliminate the subject categories because of too small numbers and references because of computing complexity. The co-author network has 1052 edges that indicate co-authorship, where we randomly choose 946 (90%) edges as training set and the remaining 106 edges as the testing set.

Multirelations-based link prediction

(1) Representation of co-authorships via multi-relations: Co-authorships via multirelations are systematically represented and extracted in a heterogeneous bibliographic network shown in Figure 1. Part of multirelations between authors could be represented in Table 1.

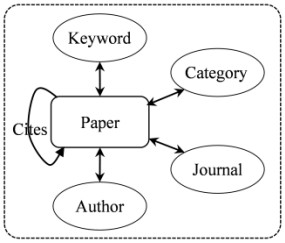


Figure 1. The nodes and relations in heterogeneous bibliographic network.

Table 1. Multirelations between authors.

Relations	*Description*
A-P-A-P-A	Common neighbours
A-P-A-P-A-P-A	Common neighbours' neighbours
A-P-J-P-A	Publish paper at the same journal
A-P-K-P-A	Authors have the same keyword
A-P-K-P-K-P-A	Authors' keywords co-word in same paper
A-P→P-A	Author x cite author y
A-P←P-A	Author x is cited by author y
A-P→P←P-A	Authors x and y cite the same paper
A-P←P→P-A	Authors x and y co-cited by same paper
A-P→P→P-A	Author x cite the paper that cite author y
A-P←P←P-A	The reverse relation of the above

(2) Measures of each relation: The four measures are the follows: path count (PC) is the number of

shortest path between two authors, normalized path count (NPC) is to discount PC by their overall connectivity, random walk (RW) and symmetric random walk (SRW) (Sun & Han, 2013).
(3) Evaluation of similarities based on link prediction: The relations and their measures combine the similarities, so there are 44 similarity indicators combined by 11 relations with four measures. We evaluate all the similarity indicators based on link prediction with precision and area under the curve (AUC).

Results

The three comparison perspectives are: (1) from the horizontal axis, compare which relation is best appropriate to the mechanism. (2) From the longitudinal axis, compare which measure is best to describe the mechanism. (3) Comparison between combined-relations-based and single-relation-based mechanisms.

The evolution mechanisms based on single-relation-based similarities

In Table 2 and Table 3, the entries emphasized in bold and italic corresponding to the highest accuracies from the horizontal axis.
In precision, the APAPA with NPC is the best appropriate and important mechanism in LIS where NPC plays the best in four measures, yet the APJPA with RW plays the worst. In AUC, the APAPA with SRW is the best mostly with little differences. There is lots of information loss in the projection from heterogeneous network to homogeneous network compared with CNs.

Table 2. The precision/AUC of single-relation-based similarities.

Relations	*PC(%)*	*NPC(%)*	*RW(%)*	*SRW(%)*
APAPA	38.4/87.5	***42.5***/87.5	31.7/87.7	41.4/***87.9***
APAPAPA	24.0/***86.2***	***32.9***/86	21.1/***86.2***	29.4/85.8
APJPA	3.2/76.8	***3.9***/77.2	0.9/76.7	2.6/***77.4***
APKPA	7.6/81.4	***20.4***/82.1	9.4/81.8	16.3/***82.3***
APKPKPA	2.2/70.8	***4.9/72.5***	2.5/70.9	4.3/72
CNs	23.4/84.1			

Comparison between combined-relations-based and single-relation-based mechanisms

The paper designs five combination strategies for comparison: (1) CR1: Combination of all relations without weights. (2) CR2: Combine all relations except APJPA. (3) CR3: Combination of all relations with weights denote by precision in Table 2. (4) CR4: the combination formed via just authors which is APAPA+APAPAPA. (5) CR5: the combination formed via just keywords, which is APKPA+APKPKPA. The precision and AUC are listed in Table 3.
In precision, the CR3 with NPC is the most appropriate and important mechanism in LIS where NPC plays the best in four measures, yet the CR5 with PC plays the worst. The AUC is consistent with the precision result mostly and others with little differences. The CR2 and CR3 with each measure are all outperformed the single-relation-based mechanisms. The CR4 performs much better than CR5 proves that in co-authorship formation the author is more important than research interest.

Table 3. The precision/AUC of different combinations of relations.

Relations	*PC(%)*	*NPC(%)*	*RW(%)*	*SRW(%)*
CR1	28.6/86.4	**40.8/88.6**	26.3/88.4	36/88.3
CR2	38.6/84.8	**43.7/87.4**	32.4/86.4	43.6/86.8
CR3	45.1/89.1	**49.2**/89.3	39.8/89.0	47.2/**89.5**
CR4	24.2/86	**38.6/86.4**	27.1/86.2	35.3/86.1
CR5	2.2/80.6	**16.7**/82.8	6.6/**83.1**	12/82.7

Conclusion and discussion

This paper uncovers the mechanisms of co-authorship network evolution by multirelations-based link prediction in LIS. In the next, we will consider other factors that influence research collaborations, all relations especially related to references to enhance the accuracy and validation in two or more different areas with different article types (e.g., journal and conference).

Acknowledgments

Our work is supported by the Ministry of Education of China Project of Humanities and Social Sciences (Grant No. 14YJC870025), the Fundamental Research Funds for the Central Universities (Grant No. 30915013101) and the National Natural Science Foundation of China (Grant No. 71173211).

References

Lu, L. & Zhou, T. (2010). Link prediction in complex networks: A survey. *Arxiv preprint arXiv:1010.0725*.

Ortega, J. L. (2014). Influence of co-authorship networks in the research impact: Ego network analyses from Microsoft Academic Search. *Journal of Informetrics, 8*(3), 728-737.

Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C. & Han, J. (2011). Co-author Relationship Prediction in Heterogeneous Bibliographic Networks. *Proc. ASONAM*.

Sun, Y. & Han, J. (2013). Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD Explorations Newsletter, 14*(2), 20-28.

Yan, E. & Ding, Y. (2009). Applying Centrality Measures to Impact Analysis: A Coauthorship Network Analysis. *Journal of the American Society for Information Science and Technology, 60*(10), 2107-2118.

JOURNALS, DATABASES AND ELECTRONIC PUBLICATIONS

DATA ACCURACY AND DISAMBIGUATION

MAPPING AND VISUALIZATION

Citing e-prints on arXiv
A study of cited references in WoS-indexed journals from 1991-2013

Valeria Aman[1]

[1] *aman@forschungsinfo.de*
iFQ - Institute for Research Information and Quality Assurance, Schützenstraße 6a, 10117 Berlin (Germany)

Abstract

This study deals with the analysis of cited references in Web of Science (WoS) to e-prints on arXiv. Created in 1991, arXiv accelerated the scholarly communication and developed into a well-established e-print repository that functions as an essential access point to the latest research in physics, astrophysics, mathematics, computer science and related fields. Authors evidently rely on arXiv full texts and refer to them in their own research papers. These cited references to arXiv that represent the acceptance of e-prints in journals and series indexed in WoS are tackled in this paper. A total of 900,000 cited references to arXiv have been identified for the 1991-2013 period. Object of investigation is on the one hand the set of cited references to arXiv, and on the other hand the set of papers in WoS that cite arXiv. Among other things, the paper illustrates that citations to arXiv peak in the year after submission and drop rapidly. The geographical distribution of authorship citing arXiv in their papers shows that authors from the US, Germany, GB, France and Italy rely heavily on arXiv. The paper identifies "arXiv-friendly" journals where the majority of articles refer to arXiv.

Conference Topic
Journals, databases and electronic publications

Introduction

The arXiv is a convenient vehicle to disseminate research results prior to the publication of peer-reviewed articles. It is also common to submit postprints for reasons of wide availability and archiving. There is no doubt that e-prints are read by a wide community and are regarded to be of good quality. Thus, it is of interest to learn more about the perception of arXiv as a source of relevant information that supports researchers' ideas and discoveries. The study sets out to answer the following questions: 1) Do authors publishing in journals covered by Web of Science (WoS) cite e-prints on arXiv? 2) What characteristics in citations can be observed? 3) In which countries are authors situated that rely on e-prints in arXiv? 4) What are the journals that include the highest rate of articles with cited references to arXiv?

Background

The rise of preprints, e-prints and arXiv

There are several definitions for the term "preprint". Lim (1996) defines a "preprint" as a manuscript that has been reviewed and accepted for publication, a manuscript that has been submitted for publication, but for which a decision to publish has not been made yet, or a manuscript that is intended for publication, but is being circulated for comments among peers prior to journal submission. Electronic prints (e-prints) refer both to preprints and post-prints (peer-reviewed published papers), and other documents that are made available on the Internet. The "preprint culture" dates back to the 1960ies, when high-energy physicists were eager to disseminate their results by printing and mailing copies of their manuscripts simultaneously to journal submission (Goldschmidt-Clermont, 1965). The time consuming process of peer-review was hence effectively bypassed. With the advent of the World Wide Web in the early 1990ies, the emergence of new methods of scientific discourse were encouraged, altering the traditional channels of scholarly communication (Brown, 2001).

In summer 1991, Paul Ginsparg conceived the repository arXiv at the Los Alamos National Laboratoy (LANL) in New Mexico. Ginsparg (1994, p.157) stated that "the realization of arXiv was facilitated by a pre-existing 'preprint culture', in which the irrelevance of refereed journals to ongoing research has long been recognized". Ginsparg (1994, p.159) designed arXiv (formerly xxx.lanl.org) as a fully automated system, where users could maintain a database to disseminate information without outside intervention.

Originally, arXiv was intended for the High-Energy Physics (HEP) community, but expanded rapidly to cover all of Physics, Astrophysics, Mathematics and Computer Science. Since September 2003 arXiv covers Quantitative Biology. In April 2007 Statistics was included, followed by Quantitative Finance in December 2008. Today, arXiv is hosted at Cornell University in New York with seven mirror sites all over the world. It contains more than 1,000,000 full-text e-prints, receiving about 9,000 new submissions each month.[1] Researchers can check arXiv for new information, search for relevant papers, post their own papers and cite references by arXiv ID. It is a self-organizing publication mode that costs the users nothing (Langer, 2000). Another reason for arXiv's popularity is its democracy, because scientists "can post their research results without being hassled by grumpy editors and referees" (ibid., p.35). According to Ginsparg (1994, p.157) physicists have learned to determine from the author, title and abstract whether to read a paper "rather than rely on the alleged verification of overworked or otherwise careless referees".

Nowadays, researchers still regard it as valuable to publish their work in peer-reviewed journals. Prior to formal publication, the findings may be spread as conference proceedings, reports, working papers or preprints. As Heuer, Holtkamp and Mele (2008, p.2) point out "scientists expect unrestricted access to comprehensive scientific information in their field, state-of-the-art information venues to optimize their research workflow and quality assurance at the parallel existence of traditional peer-review and the immediacy of dissemination and feedback". A publication delay of several months between the completion of a work and its appearance in a peer-reviewed journal is simply a "negative phenomenon in scientific information dissemination" (Amat, 2008, p.379). Amat (ibid.) found that the publication delay depends primarily on the peer-review process (see also Luwel, 1998). ArXiv serves to overcome this delay and helps to circulate results upon realization.

Previous work

The citation behaviour of e-prints available through arXiv has been studied extensively. Youngen (1998) identified the growing importance of e-prints in the published literature. He found that e-prints became the first choice among physicists and astronomers for finding current research and keeping up with colleagues and competitors at other institutions. Brown (2001) studied citations of e-prints on arXiv in astronomy and physics journals from 1998 to 1999. The citation analysis showed that the peak of citations to e-prints is reached after three years, which is comparable to papers in print journals. Garner, Horwood & Sullivan (2001) determined the place of e-prints in the scholarly information delivery, concluding that rapid dissemination of results in form of preprints establishes priority and enables rapid feedback. Brown (2003) asked for the opinion of chemists about citing e-prints in the articles they author. Fifty-two percent said they would cite e-prints whenever possible, whereas 48% stated that they would not. Reasons for avoiding to cite the Chemistry Preprint Server (CPS) are the lack of relevant articles, the lack of customary to cite, and the lacking awareness of CPS (ibid., p.365). The study of infiltration of CPS e-prints into the literature of chemistry revealed that "no citations to e-prints were found in the journal literature using ISI's Web of Science from 2000 to 2001" (ibid., p.366). Prakasan & Kalyane (2004) focused on the

[1] http://arxiv.org/stats/monthly_submissions / [Last visited January 06, 2015]

citations in Science Citation Index to e-prints on arXiv, submitted under the four categories hep-ex, hep-lat, hep-ph and hep-th[2], providing a broad insight into citation habits.

Several studies focused on the citation impact of e-prints on arXiv, also within the Open Access debate (see Harnad & Brody, 2004; Antelman, 2004). Schwarz & Kennicutt (2004) analyzed articles published in the Astrophysical Journal in 1999 and 2002 and reported that papers posted to the astro-ph-section on arXiv were cited more than twice as often as those without a version on arXiv. In accordance, Metcalfes (2005) findings show that astronomy papers in the highly-cited journals Science and Nature received higher citation rates when their authors posted their papers on arXiv's astro-ph. Metcalfe (2006) studied the field of solar physics with the result that papers posted to arXiv are on average 2.6 times as often cited as papers not being posted. He concludes that higher citation rates are not a result of self-selection of outstanding papers, since conference proceedings reveal the same result. Moed (2007) analyzed how the citation impact of articles deposited in the Condensed Matter section in arXiv and subsequently published in a journal compares to that of articles not deposited on arXiv. He concluded that arXiv accelerates citations, because it makes papers earlier available. Davis & Fromerth (2007) examined whether mathematics journals from 1997 to 2005 with a previous preprint version on arXiv receive more citations than non-deposited. Their findings show that articles in arXiv receive on average 35% more citations, which translates to 1.1 citations per article. They explain the citation advantage with the Open Access, the Early View, and the Quality postulates, which are non-exclusive.

Henneken et al. (2007) analyzed whether e-prints on arXiv are preferred over the journal articles in four core journals in astrophysics. They found that as soon as an article is published, the community prefers to read and cite it, so that the usage in the NASA Astrophysics Data System (e-print system) drops to zero. They also showed that the half-life (the time at which the use of an article is half the use of a newly published article) for an e-print is shorter than for a journal article. Gentil-Beccot, Mele & Brooks (2009) investigate whether HEP scientists still read journals or rather prefer digital repositories. Their citation analysis shows that free and immediate dissemination of preprints results in a citation advantage for HEP journals. Furthermore, their analysis of clickstreams reveals that high-energy physicists prefer preprints and seldom read journals.

Some of the studies suggest that articles with a previous preprint on arXiv receive more citations than articles without. Other studies report no such effect. Gentil-Beccot, Mele & Brooks (2009) did not detect any citation advantage from publishing in Open Access HEP journals. Their finding is similar to that of Moed (2007) in Condensed Matter, Davis (2007) in Mathematics and Kurtz & Henneken in Astrophysics (2007).

Brody, Harnad & Carr (2006) examined the correlation of the number of article downloads and the number of citations. On the basis of arXiv they show that the short-term Web usage impact of e-prints predicts a medium-term citation impact of the final article. Haque and Ginsparg (2009; 2010) found that e-prints posted to arXiv at the beginning and end of a day reach a wider readership and receive higher citation rates over the course of ensuing years than posting in the middle of day. Shuai, Pepe & Bollen (2012) analyzed the online response to preprint publications on arXiv, studying the delay of article downloads and Twitter mentions following submission.

Larivière et al. (2014) analyzed the proportion of papers across all disciplines on arXiv for the 1991-2012 period, just as the proportion of arXiv papers that are published in WoS-indexed journals. They determine the time between arXiv submission and journal publication, ageing characteristics and impact of arXiv e-prints and their published alter ego. They also focus on

[2] High energy physics - experiment (hep-ex), high energy physics - lattice (hep-lat), high energy physics - phenomenology (hep-ph), and high energy physics - theory (hep-th).

the proportion of cited references in WoS to arXiv e-prints by discipline. Working with percentages, they quantify that journals in nuclear and particle physics have 6.6% of their references to arXiv e-prints, whereas in mathematics this share is below 1.5% (ibid., p.1163). Stimulated by the work of Larivière et al. (2014), this study sets out to quantify the number of cited references in WoS to arXiv manuscripts, and to provide a broader view on characteristics of cited references and the papers that include them.

Data and methods

Database

The study builds upon the bibliometric database at the "Competence Center for Bibliometrics for the German Science System" that is hosted at the iFQ.[3] It consists of data from Thomson Reuter's Web of Science. Peer-reviewed journal articles are the primary mode of communication of scientific research. Researchers write reviews or articles with discoveries, theories and results. To relate their work they cite other articles if they know the article and believe it to be relevant to their own work. They might also provide negative citations in order to disagree or to say that a paper has flaws (see Brody, Harnad & Carr, 2006). Citations can be therefore used as a measure of influence and importance of preceding articles.

The identification of references to arXiv depends on the quality of the bibliographic information (e.g. the presence of the reference to arXiv) and the extent to which WoS was able to parse the references of the citing articles. Identifying cited references to arXiv can lead to false positives, when a reference looks like an arXiv identifier but is actually not, or where authors make mistakes. A linking by bibliographic data is more precise as it builds upon author names, journal title, volume, page number, year of publication etc.

Data collection

Different from Youngen (1998), who analyzed those cited references that state explicitly "preprint" in ISI's SciSearch (p.451), this study also includes postprints. Hence, all manuscripts on arXiv are in the following referred to as "e-prints". The e-print identifier assigned by arXiv provides a standardized number that allows each e-print to be uniquely identified. This uniqueness is required for correct citing of the work. ArXiv has established a subject grouping and numbering system for submitted e-prints. Examples are Astrophysics (astro-ph), Condensed Matter (cond-mat), High-Energy Physics-Theory (hep-th) or Nuclear-Experiment (nucl-ex), followed by a numerical string, indicating the year and month of submission, and an increasing accession number. A typical example is quant-ph/95002, where quant-ph stands for Quantum Physics, "95" for the year 1995 and "002" for the accession number. Up to March 2007 this ID enabled a broad subject categorization. In April 2007, the arXiv-ID was changed and no longer contains subject categories. It consists of eight digits, of which the first four represent the year and month of submission. Divided by a period, they are followed by a four-digit long accession number, e.g.: arXiv: 0705.0002. We can infer that this e-print was loaded in May 2007. Since the accession number will soon reach its capacity, the length of the accession number has been extended by one digit in January 2015.[4]

The search for arXiv e-prints in the cited reference field in WoS was approached in several steps. E-prints up to 2007 were identified on the basis of an alphanumeric string that contains the subject category followed by the year of submission and the accession number.[5] E-prints published in 2007 or later were identified by the string "arXiv" followed by a numerical string. This led to an overall satisfying result, since the string "arXiv" is unique and causes

[3] http://www.bibliometrie.info/ [Last visited January 06, 2015]
[4] http://arxiv.org/new#dec19_2014 [Last visited January 06, 2015]
[5] The categories in bold print were used for the matching: http://arxiv.org/ [Last visited January 06, 2015]

almost no confusion. A low number of false positives cited references were deleted manually. Only one in four cited references had a publication year assigned, which is indeed not necessary, since it is part of the arXiv ID. With the application of Regular Expressions in SQL the year of e-print publication was deduced for more than 99% of cited references. A publication year was not deducible, where authors cited arXiv simply in this fashion: "arXiv". The search strategy may not include citations to works that technically have to be considered as arXiv e-prints. According to Youngen (1998, p.451) authors may have cited preprints as "submitted to...", "to be published in...", "in press" or "unpublished", depending on their state in the publication cycle. Thus, in reality, the number of citations to e-prints on arXiv may be much higher than presented here.

Data corpus[6]

With the search strategy described, 892,867 cited references to arXiv were identified for the 1991-2013 period, of which 357,557 have a distinct character string. Due to multiple subject categorizations in arXiv, author typos, or erroneous data parsing in WoS, one and the same e-print can be referred to in different spelling variants. Hence, the actual number of arXiv e-prints cited in the 1991-2013 period by papers in WoS is lower. At the same time 289,145 distinct papers were identified in WoS that constitute these 892,867 cited references. To relate these figures, Brown (2001) found 35,928 citations to arXiv e-prints (posted between 1991 and 1999) in astronomy and physics journals published in 1998-1999. In the following, analyses are based on the cited references to arXiv and the WoS-papers that include them.

Results and discussion

Figure 1 provides an overview of the data collected. The number of e-prints submitted to arXiv has been gradually rising from 303 in 1991 to 92,641 in 2013.[7] The number of papers in WoS citing at least one e-print on arXiv has steadily increased and comprises around 28,000 papers in 2013. In addition, we can see the number of cited references to e-prints on arXiv with the publication year of the citing paper as indicated on the x-axis. We can derive that a paper citing arXiv includes on average more than one citation to e-prints on arXiv. Most of the citations to e-prints were provided in 2012 (ca. 76,000).

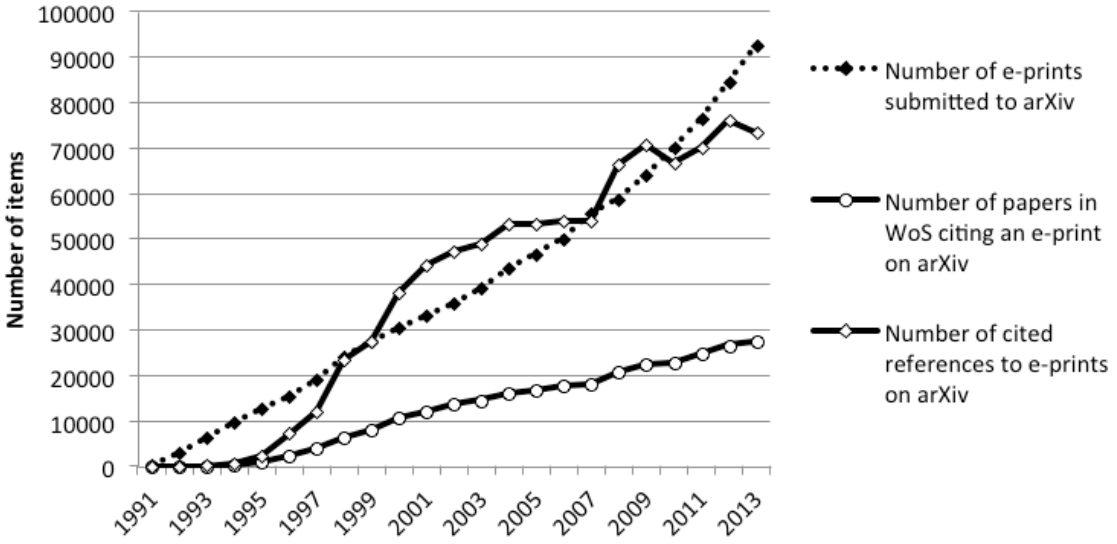


Figure 1. Overview of the yearly growth of submissions to arXiv, the number of papers in WoS citing arXiv e-prints according to their publication year, and the number of cited references.

[6] The data corpus can be requested on demand.
[7] http://arxiv.org/stats/monthly_submissions [Last visited January 06, 2015]

The analysis of document types shows that articles rank first with 96.0% of all WoS documents from 1991-2013 that cite arXiv. Reviews (3.2%) refer to arXiv as well, in order to provide a broad or up-to-date state of research. Editorials, Letters, Corrections and Notes also reference arXiv.

In the following, it does make a difference whether cited references are analysed or the WoS-papers that include those. Due to different citation habits, even within a broad field such as physics, it appears more suitable to consider primarily the citing papers. Table 1 provides an overview of the subject areas that constitute most of the citations to arXiv. The first column lists the Subject Categories[8] (SC) in WoS in a descendant order, regarding the number of arXiv citing papers assigned to this SC. We can see that Particle Physics ranks first (21%), followed by Astronomy and Astrophysics. In total, these 12 SC cover more than 90% of all citing papers that refer to arXiv between 1991 and 2013. The percentages and order of the SC changes when we have a look on the number of cited references to arXiv. Particle Physics still ranks first, claiming almost one-third of all cited references to arXiv. The results suggests that papers in Particle Physics have on average a higher number of cited references to arXiv than those in other SC.

Table 1. Overview of Subject Categories in WoS that contribute to the majority of papers that cite arXiv and their number of cited references. The data is based on 289,145 arXiv-citing papers in WoS that provide 892,867 cited references in 1991-2013.

Subject Category	No. of papers citing arXiv	Share in %	No. of cited references	Share in %
Physics, Particles & Fields	88,757	21.0	398,022	30.5
Physics, Multidisciplinary	70,383	16.7	248,091	19.0
Astronomy & Astrophysics	68,805	16.3	225,326	17.3
Physics, Mathematical	28,073	6.7	82,490	6.3
Physics, Condensed Matter	25,658	6.1	49,852	3.8
Mathematics	23,894	5.7	46,952	3.6
Physics, Nuclear	22,838	5.4	83,712	6.4
Optics	13,602	3.2	27,414	2.1
Physics, Atomic, Molecular & Chemical	12,754	3.0	25,625	2.0
Mathematics, Applied	10,976	2.6	20,169	1.5
Physics, Applied	9,223	2.2	17,099	1.3
Physics, Fluids & Plasmas	5,704	1.4	9,488	0.7

This leads us to the analysis of the distribution of cited references among the papers in WoS that cite arXiv. Table 2 illustrates the frequency of citing papers in WoS that include as many cited references as stated in the left column. We can see that six papers in WoS have more than 200 references to arXiv in their list of references. Every eleventh paper, out of the set of arXiv citing papers, includes 6 to 10 references to arXiv. Nevertheless, around 46% of citing papers provide a single reference to arXiv. A closer look on the paper with the highest number of cited references to arXiv shows that it is a review article from 2000 on String Theory and Gravity, where a link to arXiv was set additionally to the journal article reference. This brings us to the analysis of characteristics in citations to arXiv. Are e-prints on arXiv immediately cited when there is no corresponding journal article or are they also used in future and even preferred over the corresponding journal article?

[8] The 260 SC in WoS are assigned to journals on the basis of their scope and citation links.

Table 2. Distribution of cited references among WoS-papers that cite e-prints on arXiv.

Number of references to arXiv in a single paper	Number of papers citing arXiv	%
more than 200	6	0.00
151 to 200	8	0.00
101 to 150	29	0.01
51 to 100	222	0.08
21 to 50	2,567	0.89
11 to 20	9,375	3.24
6 to 10	25,859	8.94
5	12,544	4.34
4	18,939	6.55
3	30,969	10.71
2	56,204	19.44
1	132,423	45.80
Total	289,145	100.00

Figure 2 shows on the one hand the line graph of all citations to e-prints on arXiv up to 2013. Different from Figure 1 the x-axis signifies the year of e-print publication. Thus, the sudden decrease of cited e-prints from 2008 on is due to the fact that they had less time to be referenced than those posted in earlier years. In addition, Figure 2 provides bars indicating the years in which these e-prints were cited by WoS papers. Each bar represents the number of cited references to arXiv in the same year as the e-print was published, the subsequent year and two and three years respectively after publication of the e-print. The space between the line graph and the bars represents the cited references to e-prints that were provided more than three years after e-print publication. Since e-prints from recent years did not have much time to be cited, the bars coincide with the line graph of the total number of cited e-prints.

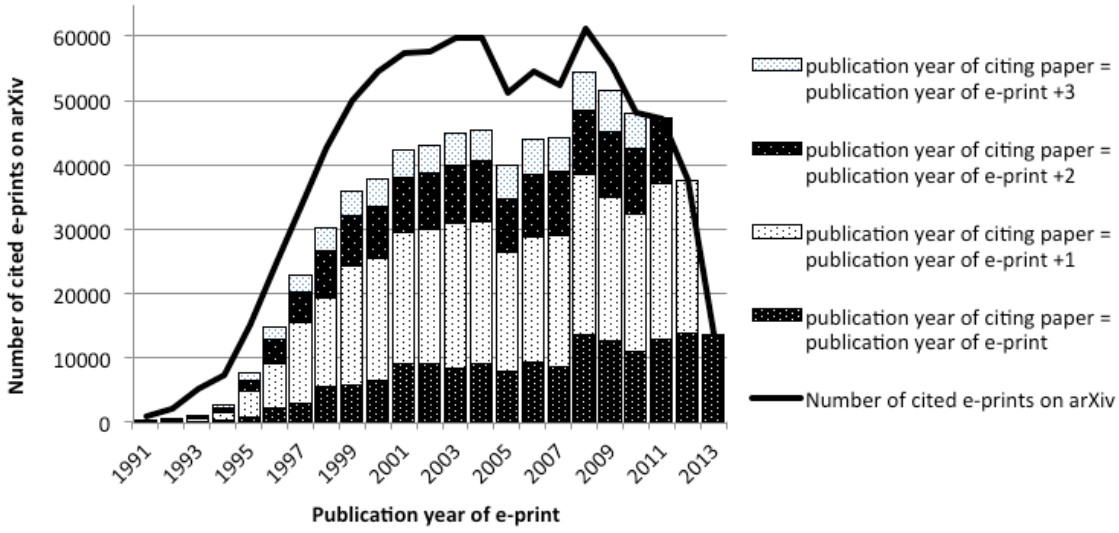


Figure 2. Time series of citation distribution. Illustrated are citations that equal the year of e-print submission, citations to e-prints that are one year old, up to the age of three years. The line graph signifies the total number of e-prints cited, published in the year as indicated.

It becomes evident that e-prints on arXiv are mostly cited in the subsequent year of e-print post. Almost half of all cited references in a year relate to e-prints that were placed on arXiv the preceding year. This is in accordance with Larivière et al. (2014, p.1166), who found that citations to e-prints on arXiv peak the year following submission. The figure also indicates that e-prints are cited immediately in the same year of posting. Only a small share of cited

references points to three-year old e-prints. On the contrary, Brown's (2011) analysis in astronomy and physics showed that the peak of citations to e-prints is reached after three years. The results in Figure 2 are in little accordance with Henneken et al. (2007, p.19) who showed that the usage of e-prints drops to zero as soon as the journal article has appeared, suggesting that authors have access to subscribed journals and prefer to cite the refereed version. Garner, Horwood & Sullivan (2001, p.251) quantified that 90% of papers on arXiv are later published in journals so that a corresponding article can be found and cited properly.

Nevertheless, there are many reasons that underscore the high citation rates of e-prints. Davis & Fromerth (2007) write that the arXiv copy is sufficient for the purpose of citing it in one's own work. They found that articles that are also accessible on arXiv receive 23% fewer downloads from the publisher's web site two years after publication (ibid., p.23). Gentil-Beccot, Mele & Brooks (2009) found that citations start before publication, because scientists in HEP do not wait for an article to be published. Even in the first few months after journal publication authors read and cite the preprint (ibid., p.6). According to Moed (2007) colleagues start to read a paper and cite it in their own articles earlier if it is deposited on arXiv. The following Figure 3 illustrates the relation between the publication year of a WoS-paper citing arXiv, and the publication year of the cited e-print. The whole bar in each year (y-axis) represents the total number of cited references to e-prints on arXiv from this year (cf. Figure 1). The cited references from each year are grouped by the publication year of the cited e-print. Each bar indicates the share of e-prints, according to their year of publication. For the year 2013 we can see that 13,000 cited references (top black part of the 2013-bar) refer to e-prints published in the same year. The lion's share of cited references in 2013 (24,000) is to e-prints published in 2012. In general, we can conclude from Figure 3 that the majority of references in each year points to e-prints published in the preceding year.

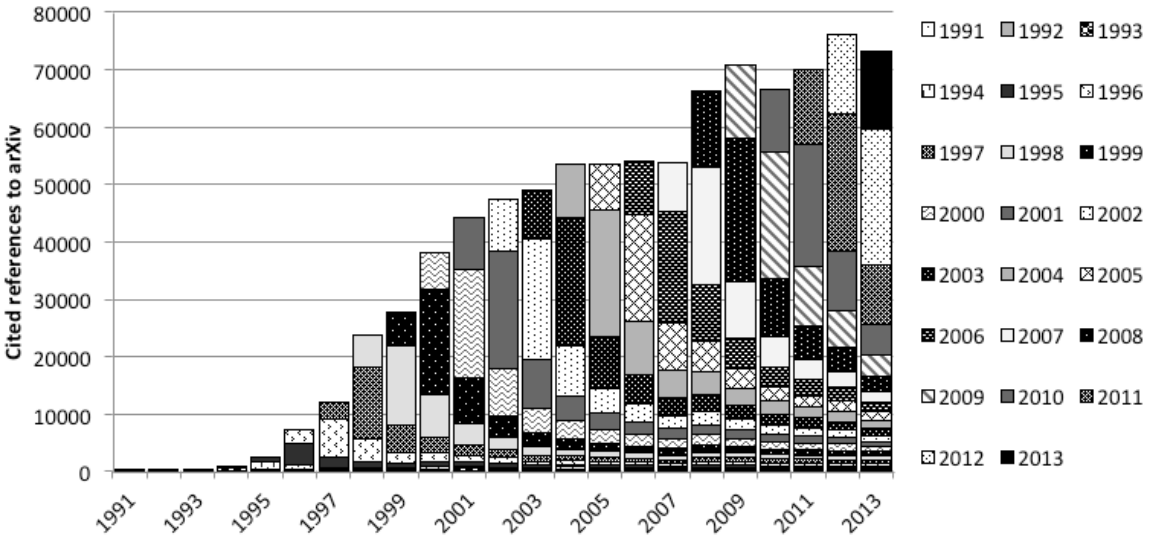


Figure 3. Time series of cited references to e-prints on arXiv. The x-axis represents the publication years of WoS-paper citing an e-print, whereas each bar represents the share of the years a cited e-print was published in.

To see where the authors that frequently cite arXiv are from, Table **3** provides a ranking of countries according to the highest number of papers in WoS with at least one cited reference to arXiv. USA rank first with one-third of all papers that cite arXiv. They are followed by Germany and Great Britain. Note that the percentages do not add up to 100, since co-authored papers can be attributed to multiple countries.

Table 3. Overview of countries that most frequently cite arXiv e-prints. The percentages are calculated on the basis of the total number of citing papers (289,145).

Rank	Country	No. of WoS-papers citing e-prints	%	Rank	Country	No. of WoS-papers citing e-prints	%
1	USA	97,085	33.6	11	Switzerland	14,489	5.0
2	Germany	45,842	15.9	12	India	11,764	4.1
3	GB	30,776	10.6	13	Poland	9,332	3.2
4	France	28,159	9.7	14	Brazil	9,004	3.1
5	Italy	27,896	9.6	15	Netherlands	8,361	2.9
6	China	25,467	8.8	16	South Korea	8,271	2.8
7	Japan	25,196	8.7	17	Australia	7,296	2.5
8	Russia	22,772	7.9	18	Israel	7,019	2.4
9	Spain	15,902	5.5	19	Sweden	5,402	1.9
10	Canada	14,879	5.1	20	Belgium	4,709	1.6

The journals whose articles most often cite e-prints on arXiv are identified in Table 4. On the left of the table, journals are ranked according to their number of citing papers in the 1991-2013 period. On the right of the table journals are ranked according to their number of cited references to arXiv. Evidently, most of the journals carry a majority of HEP content. Among these are *Physical Review D*, *Journal of High Energy Physics* (JHEP), *Physics Letters B* and *Nuclear Physics B*. Striking are also the astrophysical journals, among which we can find the *Astrophysical Journal, Monthly Notices of the Royal Astronomical Society* and *Journal of Cosmology and Astrophysical Physics*.

Table 4. Overview of journals in WoS with the highest number of papers citing arXiv and journals with most of the cited references to arXiv in the 1991-2013 period.

Journal	Citing papers	%	Journal	Cited ref.	%
Physical Review D	30,287	10.5	Physical Review D	112,261	12.6
Physical Review B	15,080	5.2	Journal of High Energy Physics	77,431	8.7
Journal of High Energy Physics	14,881	5.1	Physical Review B	66,750	7.5
Physical Review Letters	13,816	4.8	Nuclear Physics B	50,757	5.7
Physics Letters B	13,707	4.7	Physics Letters B	29,195	3.3
Physical Review A	9,599	3.3	Physical Review Letters	28,873	3.2
Astrophysical Journal	8,428	2.9	Classical and Quantum Gravity	22,969	2.6
Nuclear Physics B	8,033	2.8	Physical Review A	20,480	2.3
Monthly Notices of the Royal Astronomical Society	6,256	2.2	Journal of Cosmology and Astrophysical Physics	19,559	2.2
Physical Review E	5,081	1.8	International Journal of Modern Physics A	18,685	2.1
Sum	125,168	43.3	Sum	446,960	50.1

Youngen (1998) could not find firm rules for citing preprints, with the exception of the Astrophysical Journal, which stated that "References to private communications, papers in preparation, preprints, or other sources generally not available to readers should be avoided" (p.453). Nevertheless, it ranks seventh among the most active journals citing e-prints on arXiv. This restriction must have been eased over the years, as can be seen in Figure 4. Depicted are time series of percentages of papers in a journal that cite arXiv, for the ten

journals with the highest number of arXiv-citing papers (see Table 4). We can observe that up to 1997 the *Astrophysical Journal* had less than 10% of their papers citing e-prints on arXiv. This share was growing in the following years to reach approx. 25%.

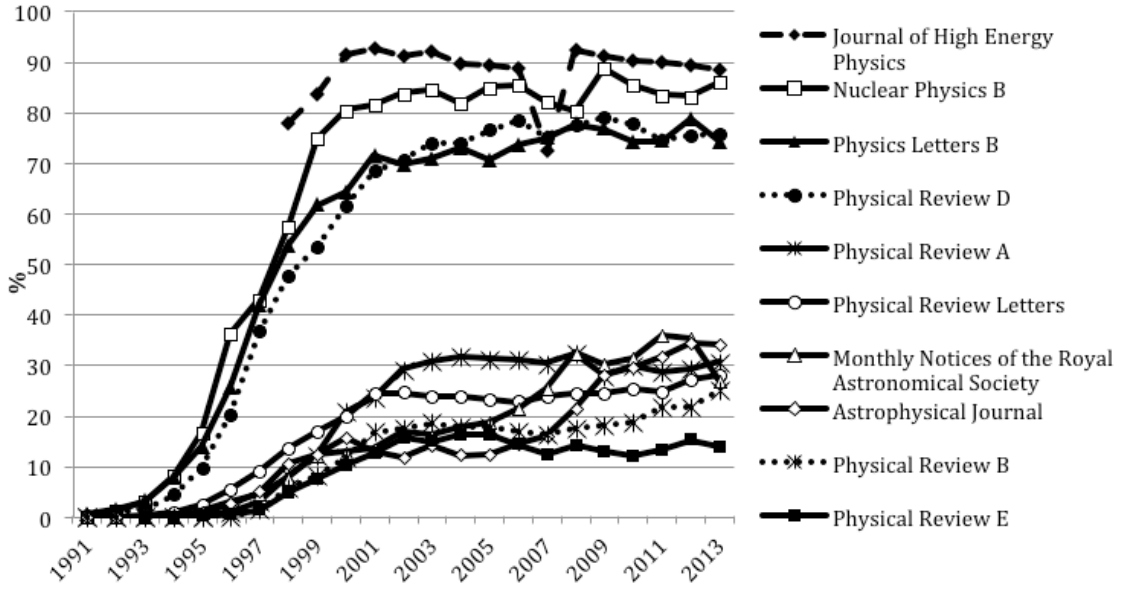


Figure 4: Time series of the percentages of papers in a journal that cite arXiv. Displayed are the 10 journals that most actively cite arXiv.

Striking is the decline of the share of papers in JHEP with references to arXiv in 2007, for which no explanation can be given. Overall, the shape of the line graphs suggests a rapid growth of arXiv's acceptance in the 1990ies and a constant reliance on arXiv in the past 15 years. The following table identifies other "arXiv-friendly" journals, where the majority of papers rely on arXiv. Since the number of papers published in a journal can differ immensely, Table 5 indicates percentages of the number of a journal's papers that cite arXiv. To provide an up-to-date view, only papers published between 2004 and 2013 are considered.

Table 5: Journals in WoS with the highest share of papers citing arXiv. Analyzed are only citing papers that were published between 2004 and 2013.

Journal	%	Journal	%
Journal of Cosmology and Astroparticle Physics	89.9	Journal of Physics G-Nuclear and Particle Physics	59.0
Advances in Theoretical and Mathematical Physics	81.7	International Journal of Modern Physics A	59.0
Annual Review of Nuclear and Particle Science	80.7	International Journal of Modern Physics D	57.5
Communications in Number Theory and Physics	79.8	Progress of Theoretical and Experimental Physics	56.2
European Physical Journal C	70.9	Physics Reports-Review Section of Physics Letters	55.5
Fortschritte der Physik-Progress of Physics	70.4	General Relativity and Gravitation	54.0
Quantum Information & Computation	69.3	Gravitation & Cosmology	54.0

Modern Physics Letters A	62.3	Journal of Sympletic Geometry	53.6
Progress in Particle and Nuclear Physics	61.5	Reviews of Modern Physics	52.8
Acta Physica Hungarica A-Heavy Ion Physics	60.4	Algebraic and Geometric Topology	52.2
Geometry & Topology	60.3	Progress of Theoretical Physics	51.7
Classical and Quantum Gravity	60.0	Astroparticle Physics	51.2

Ranking the journals on the basis of percentages instead of absolute numbers enables us to spot mathematics journals. The 24 journals listed prove that the circle of users coincides with the target group of arXiv that consists mainly of high-energy physicists. In HEP it is usual practice to submit papers to arXiv prior to journal submission. According to Gentil-Beccot, Mele & Brooks (2009) the arXiv often presents a version very similar to the published one. Finally, the arXiv version is freely available, while the journal versions require subscription.

Conclusions

The rapid dissemination of research results enabled by arXiv has accelerated the read-and-cite process (see Brody, Harnad & Carr, 2006). The identified number of cited references to arXiv and the rapid citation of e-prints in WoS-indexed journals indicate that e-prints are accepted within certain communities as well as among journal editors. Taking citation counts as a proxy for quality, e-prints on arXiv can be regarded as of good quality. They are valued, read and used within the scientific community, mainly because they present results upon finalization, circumventing the publication delay. To refer to these most up-to-date findings, authors evidently do not hesitate to cite arXiv e-prints in their research papers. The high number of cited references presented in this study suggests the usage of e-prints over the journal articles, as it was also found by Davis & Fromerth (2007). One reason for the preference of arXiv e-prints is the free availability of full text, especially if readers do not have access to the journal. Besides, the arXiv version is often similar to the formal journal article and can be easily cited by ID. An obvious reason to cite arXiv full texts even years after publication might be simply that the e-print does not have a published alter ego to be cited. Furthermore, the results showed that citations to e-prints peak in the year after publication and drop rapidly in the following years. Authors may still rely on the e-print but cite the formal publication, so the decline in citations does not necessarily indicate a decline in use. This could be proved in a future study with download data of arXiv e-prints over time.

Whereas this initial study is mostly exploratory, future work will link arXiv data to the data in WoS to examine, whether the cited e-prints have a journal version or not. So far, Larivière et al. (2014, p.1161) found that 64% of all arXiv e-prints are published in a WoS-indexed journal. An improved unification in our bibliometric database of institution names will allow analysing reasons why certain institutions rely on arXiv. Is it due to the presence of large physics departments, research centres, outstanding and highly-active researchers, collaboration or cutting-edge research? Moreover, a qualitative study of authors and their reasons to cite arXiv instead of the journal article would provide valuable information on the recent scholarly communication process.

References

Amat, C. B. (2008) Editorial and publication delay of papers submitted to 14 selected Food Research journals. Influence of online posting. *Scientometrics* 74, 3, 379-389.

Antelman, K. (2004). Do open-access articles have a greater research impact? *College and Research Libraries*, 65 (2004) 372 – 382.

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science & Technology*, 57(8), 1060–1072.

Brown, C.M. (2001). The E-volution of preprints in the scholarly communication of physicists and astronomers. *Journal of the American Society for Information Science and Technology*, 52(3), 187–200.

Brown, C. (2003). The role of electronic preprints in chemical communication: Analysis of citation, usage, and acceptance in the journal literature. *Journal of the American Society for Information Science & Technology*, 54(5), 362–371.

Davis, P.M., & Fromerth, M.J. (2007). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics*, 71(2), 203–215.

Garner, J., Horwood, L., & Sullivan, S. (2001). The place of eprints in scholarly information delivery. *Online Information Review*, 25(4), 250–253.

Gentil-Beccot et al. (2009). Information Resources in High-Energy Physics: Surveying the Present Landscape and Charting the Future Course. *Journal of the American Society for Information Science and Technology*, 60 (2009) 150–160.

Gentil-Beccot, A., Mele, S., & Brooks, T.C. (2009). Citing and reading behaviours in high-energy physics. How a community stopped worrying about journals and learned to love repositories. Retrieved January 6, 2015 from: arXiv:0906.5418.

Ginsparg, P. (1994). First steps towards electronic research communication. *Los Alamos Science,* 22, 156-165.

Goldschmidt-Clermont, L. (1965). Communication Patterns in High-Energy Physics. Retrieved January 6, 2015 from: http://eprints.rclis.org/archive/00000445/02/communication_patterns.pdf

Haque, A., & Ginsparg, P. (2009). Positional effects on citation and readership in arXiv. *Journal of the American Society for Information Science and Technology*, 60(11), 2203–2218

Haque, A., & Ginsparg, P. (2010). Last but not least: Additional positional effects on citation and readership in arXiv. *Journal of the American Society for Information Science & Technology*, 61(12), 2381-2388.

Harnad, S. & Brody, T (2004). Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals*, D-Lib Magazine* 10.

Henneken, E.A. et al. (2007). E-prints and journal articles in astronomy: A productive co-existence. *Learned Publishing*, 20, 16-22.

Heuer, R.-D., Holtkamp, A., Mele, S. (2008). Innovation in Scholarly Communication: Vision and Projects from High-Energy Physics. pp.1-15. DESY-08-054. Retrieved January 6, 2015 from: https://bib-pubdb1.desy.de/record/86123/files/getfulltext.pdf

Kurtz, M & Henneken, E. (2007). Open Access does not increase citations for research articles from The Astrophysical Journal, Retrieved January 6, 2015 from: arXiv:0709.0896

Langer, James. (2000). "Physicists in the new era of electronic publishing." *Physics Today Online*, 53(8):35-38.

Larivière, V., Sugimoto, C.R, Macaluso, B., Milojevic', S., Cronin, B. & Thelwall, M. (2014). arXiv E-Prints and the Journal of Record: An analysis of Roles and Relationships. *Journal of the American Society for Information Science & Technology*, 65(6):1157–1169.

Lim, D. (1996). Preprint Servers: A New Model for Scholarly Publishing? *Australian Academic and Research Libraries (AARL)* 27 (1), 21–30.

Luwel, M. (1998). Publication delays in the science field and their relationship to the ageing of scientific literature. *Scientometrics*, 41, 29-40.

Metcalfe, T.S. (2005). The rise and citation impact of astro-ph in major journals. *Bulletin of the American Astronomical Society,* 37, 555–557. Retrieved January 6, 2015 from: http://arXiv.org/abs/astro-ph/0503519

Metcalfe, T.S. (2006). The citation impact of digital preprint archives for solar physics papers. *Solar Physics,* 239, 549-553.

Moed, H.F. (2007). The effect of 'Open Access' on citation impact: An analysis of arXiv's condensed matter section. *Journal of the American Society for Information Science & Technology*, 58(13), 2047–2054.

Prakasan, E.R. & Kalyane, V.L. (2004). Citation analysis of lanl high energy physics e-prints through Science Citation Index (1991-2002). Retrieved January 6, 2015 from: http://eprints.rclis.org/archive/00002200/

Schwarz, G.J., Kennicutt, R. C. J. (2004). Demographic and citation trends in astrophysical journal papers and preprints. *Bulletin of the American Astronomical Society,* 36 (2004), 1654–1663.

Shuai, X., Pepe, A. & Bollen, J. (2012). How the Scientific Community Reacts to Newly Submitted Preprints: Article Downloads, Twitter Mentions, and Citations. *PLoS ONE* 7(11): e47523.

Youngen, G.K. (1998). Citation patterns to traditional and electronic preprints in the published literature. *College & Research Libraries*, 59(5), 448–456.

Evolutionary Analysis of Collaboration Networks in *Scientometrics*

Yuehua Zhao[1], Rongying Zhao[2]

[1]*yuehua@uwm.edu*
University of Wisconsin-Milwaukee Milwaukee, School of Information Studies, P. O. Box 413 Milwaukee, WI 53201 (United States)

[2]*zhaorongying@126.com*
Wuhan University, School of Information Management, Research Center for China Science Evaluation, The Center for the Studies of Information Resources, Wuhan, Hubei 430072 (China)

Abstract
The research area of scientometrics began during the second half of the 19th century. After decades of growth, the international field of scientometrics has become increasingly mature. The present study intends to understand the evolution of the collaboration network in *Scientometrics*. The growth of the discipline is divided into three stages: the first time period (1978-1990), the second period (1991-2002), and the third period (2003-2014). Both macro-level and micro-level network measures between the studied time periods were compared. Macro-level analyses show that the degree distribution of the collaboration in each timespan are consistent with power-law, and both the average degree and average distance steadily increase with time. Micro-level structure analyses illustrate the authors with high performance in raw degree measure, degree centrality measure, and betweenness measure are dynamic in different timespans. From three dimensions (raw degree, degree centrality, and betweenness centrality), the collaboration dominators are identified in each time span. In addition, the visualization methods are applied to display the evolution of the collaboration networks for each of the three stages of scientometrics' development.

Conference Topic
Journals, databases and electronic publications

Introduction
Scientometrics is an interdisciplinary field that uses mathematical, statistical, and data-analytical methods and techniques to perform a variety of quantitative studies of science and technology (Chen, Börner, & Fang, 2013). In short, it can be defined as the science of science. The term "Scientometrics" has been first used as a translation of the Russian term "naukometriya" (measurement of science) coined by Nalimov and Mulchenko (1969). The research area of scientometrics began during the second half of the 19th century. This paper proposed a macro- and micro-level overview of the author collaboration patterns in journal *Scientometrics* to study the evolution of the field of scientometrics. The present study intends to understand the evolution of the collaboration network in *Scientometrics*. In this study, social network analysis methods are employed to describe the evolution of scientometrics over nearly 40 years after entering the development stage of this field. Both macro-level and micro-level network measures between the studied time periods were compared. Then, visualization methods were applied to display the evolution of the collaboration networks in three periods: the first time period (1978-1990), the second period (1991-2002), and the third period (2003-2014).

Related Works and Research Questions
Scientometrics has been studied for more than 100 years. Over the past years, scientists' studies of scientometrics shifted from the unconscious to consciousness, from qualitative research to quantitative research, and from external description to detailed study revealing the inherent properties of scientific production. Previous scholars (Pang, 2002; Yuan, 2010) tend to divide the development of scientometrics into three stages: embryonic period (from the

second half of the 19th century to early 20th century), the founding period (from the beginning of the 20th century to the 1960s), and development period (after the 1970s). In order to study the development period of scientometrics, Schubert (2002) indicated that as the representative communication channel of its field, the journal *Scientometrics* reflects the characteristic trends and patterns of the past decades in scientometric research. Therefore, in this study, we employed the publications in *Scientometrics* over the past 37 years to detect the evolution of the scientific collaboration networks in this field.

Previous research has provided some insight into the author collaboration network analysis in different disciplines. Barabasi et al. (2002) investigated the collaboration network in mathematics and neuroscience articles published between 1991 and 1998. Newman (2001) compared the co-authorship networks of in physics, biomedical research, and computer science, and found the differences of the collaboration networks between experimental and theoretical disciplines. By using the bibliometric methods, Ardanuy (2012) analyzed the level of co-authorship of Spanish research in Library and Information Science (LIS) until 2009, and found a significant increase in international collaboration. Given the advanced visualization techniques, Franceschet (2011) represented a collaboration picture of computer science collaboration including all papers published in the field since 1936.

These studies have investigated the collaboration networks in different disciplines and compared their differences. However, few studies investigated the field of scientometrics over the past 37 years. There is a need for researchers to identify and compare both the macro-level and micro-level characteristics of the scientific collaboration network in *Scientometrics* through different time periods.

This paper intended to address the following two research questions:

RQ1. What are the macro-level features of the collaboration networks in *Scientometrics* in each time period?

RQ2. What are the micro-level features of the collaboration networks in *Scientometrics* in each time period?

Method

Data collection

For the development period of scientometrics, the foundation of the journal *Scientometrics* (in September, 1978) is a landmark event. Following some of the predecessors (Schoepflin & Glänzel, 2001; Hou, 2006), this study used the journal as a representative model of scientometrics research. The research data involves 3627 documents published in *Scientometrics* during 1987 to 2014 retrieved from the Web of Science on December 10th, 2014, and the other 347 articles published from 1978 to 1986 retrieved on April 20th, 2013. The total of 37 years were divided into three periods: the first time period (1978-1990), the second period (1991-2002), and the third period (2003-2014).

The raw data extracted from Web of Science database that consisted of the bibliometric information of each paper. Microsoft Excel was applied to build the 2-mode author-to-paper matrices for each time period. In order to produce the collaboration networks, the 2-mode author-to-paper matrices were transferred to 1-mode author-to-author matrices based on the formula proposed by Breiger (1974): $P=A(A^T)$. In this case, the matrix A was the 2-mode author-to-paper matrix and the matrix AT was the transposition of the matrix A, and the 1-mode author-to-author matrix was generated by multiplying these two 2-mode matrices. In the produced author-to-author matrix, each row and column represented an author, the intersection cells contained the cumulative number of the co-authored papers by two authors, and the diagonal cells demonstrated the total number of papers written by each author.

Data analysis

Two social network analysis software packages (Ucinet and Netdraw) (Borgatti, Everett, & Freeman, 2002) were adopted in the data analysis to calculate the network measures and draw the networks. Ucinet is a software package which mainly deals with the social network analysis, and Netdraw, the network visualization tool, can be used to display the networks generated by Ucinet.

Results and Discussion

An overview

Over the 37 years, a total of 4,211 authors published 3,974 papers in *Scientometrics*. Figure 1 indicates the distribution of the number of articles and the number of scholars in each time period. In Figure 1, the *X*-axial represented the 3 time periods, and the *Y*-axial represented the frequencies, and the 2 bars in each period showed the number of authors and articles separately, and the line showed the trend of the differences between the two bars. Separately, 626 papers were contributed to by 435 authors from 1978 to 1990, 1,106 papers were published by 1,029 authors from 1997 to 2005, and 2,242 papers were written by 3,102 authors from 2006 to 2014. Based on Figure 1, both the number of articles and the number of authors increased over the three time spans. When we compared the two frequencies in each period, the number of articles was greater than the number of authors at the first two stages, but the number of authors boomed at the third stage which resulted in the number of authors being much greater than that of the authors. The increases of the total number of articles and authors suggested the rises of the collaboration opportunities through the three time periods.

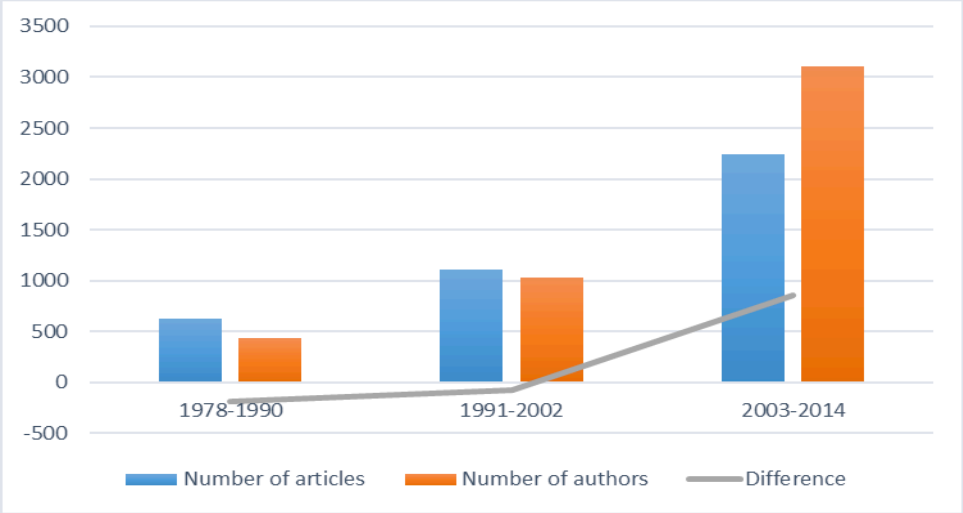


Figure 1. Distribution of the number of articles and authors in three time periods.

Macro-level structure analysis

In order to study the evolution of the scientific collaborations through three time periods, three 1-mode author-to-author matrices were plugged in Ucinet to calculate a variety of network measurements. There are a number of measures which can be used to evaluate the structure of a network. In this study, we will mainly focus on four elements to approach: degree distribution, average degree, average distance, and cluster coefficient.

The number of collaborators that each author has in a collaboration network is the degree of a node (Ding, Rousseau, & Wolfram, 2014). In Figure 2, three lines illustrated the distributions

of the node degree in each time span, respectively. The *X*-axial represented the number of authors, and the *Y*-axial represented the degree of the authors. From Figure 2, it can be seen that most authors held the low degree in all three periods. Based on the locations of three distribution lines, more authors tended to join more collaborations from 1978 to 2014 with the increase of the number of total authors published on the journal.

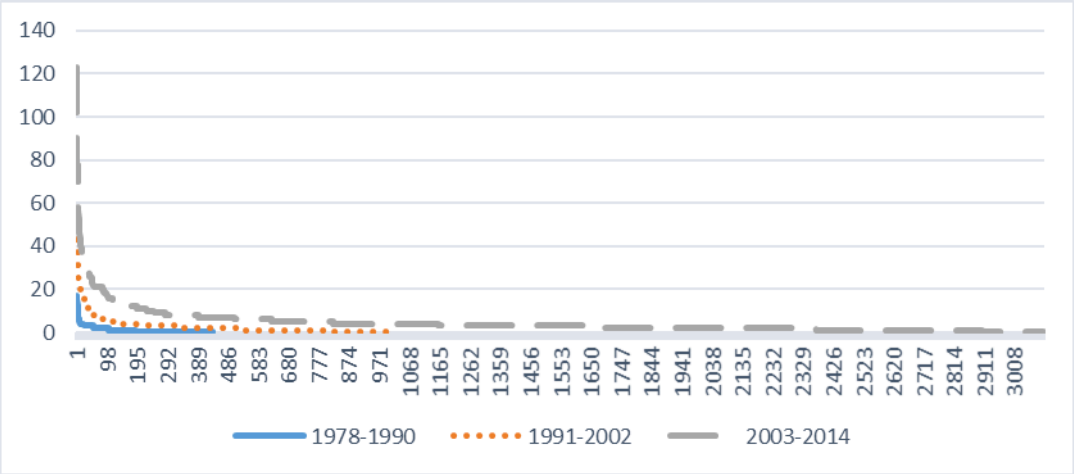
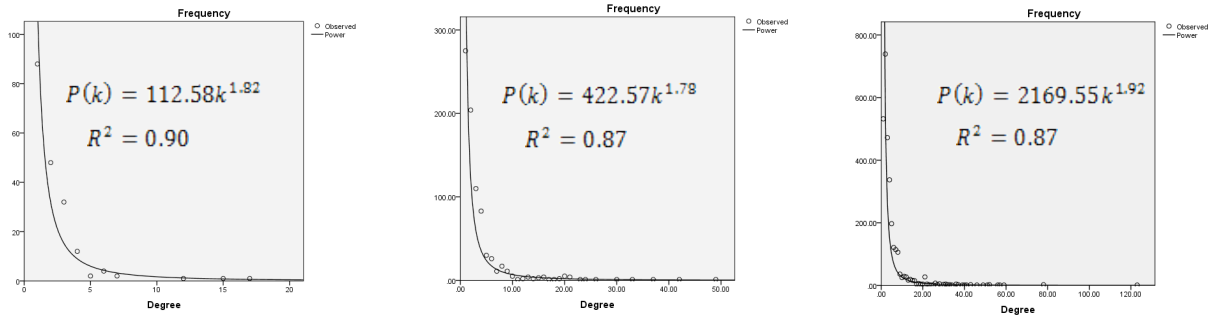


Figure 2. Degree distribution for authors in three time periods.

The degree distribution characterizes the spread of the edges each node has in a network. Although the degree distribution of a random graph is a Poisson distribution, Albert and Barabási (2002) have discovered that, for most large networks, the degree distribution has a power-law tail: $P(k) \sim k^{-\gamma}$, where $P(k)$ is the distribution function. In this study, the distributions of the collaboration network in each period were calculated and drawn in Figure 3. Power-law regression model was used to detect the degree distribution patterns in different timespans (Albert & Barabási 2002). Figure 3 illustrated the modeling results for the three periods, and the x-axis plots low degree nodes on the left and high degree nodes on the right; the y-axis indicates their probability. In both cases, power-law model performed the good fits to the observed data. In relationship between the degree of the authors and the corresponding frequencies can be estimated by: $P(k) = 112.58k^{1.82}$ with $R^2 = 0.90$ in 1978-1990, $P(k) = 422.57k^{1.78}$ with $R^2 = 0.87$ in 1991-2002, and $P(k) = 2169.55k^{1.92}$ with $R^2 = 0.87$ in 2003-2014. As discussed by Albert and Barabási (2002), the degree distribution of the collaboration network of high-energy physicists reach the almost perfect power-law



1978-1990 1991-2002 2003-2014

Figure 3. Degree distribution plots for collaboration networks.

with an exponent of 1.2, while the collaboration networks of mathematicians and neuroscientists between 1991 and 1998 held the degree exponents 2.1 and 2.5 (Barabasi et al., 2002). Comparing with those previous studies in different disciplines, the degree distribution of the collaboration of *Sicentometrics* in each timespan were consistent with power-law with degree exponents 1.82, 1.78, and 1.92, respectively. In addition to degree distribution, previous studies proved that there were several other useful indicators to feature a social network. Table 1 represented the four key measures for each time periods. Figure 3 describes the changes of each measure between 1978 and 2014.

Table 1. Four key measures of the collaboration networks in each time periods.

	1978-1990	*1991-2002*	*2003-2014*
Average Degree	0.794	2.101	3.435
Average Distance	1.412	4.673	7.106
Clustering Coefficient	0.941	0.873	9.014
Components	309	420	701
Diameter	4	11	19

Average degree is calculated by counting the average number of links per author (Barabasi et al., 2002). In the collaboration network, the average degree characterizes the interconnectedness between authors. Yin, Kretschmer, Hanneman, and Liu (2006) identified that the higher the average degree, the tighter the network. From Table 1, we can see that the average degree steadily increased with time, which demonstrated that authors cooperated more often. This results confirmed Barabasi et al.'s (2002) observations in Mathematics and Neuroscience. One possible reason might be the sharp increase of the total number of authors led to more possible connections between the new authors and also between the new authors and the existing authors.

The distance between two nodes is measured by the length of the shortest path between those two nodes. Average distance in a network is calculated by the average length of the geodesic paths between all reachable pairs of nodes (Borgatti, Everett, & Freeman, 2002). From Table 1, the average distance of the collaboration networks started form 1.412 (in 1978-1990), grew to 4.673 (in 1991-2002), and finally reached 7.106 (in 2003-2014). Watts and Strogatz (1998) examined that many social networks show a “small world” phenomenon that have small characteristic path lengths. According to Yin et al. (2006), short average distance allows authors to share information more rapidly. In this case, the average distance of the collaboration network enlarged with time, but actors were still able to reach the others within short paths in all periods. The cluster coefficient for the co-authorship network in *Scientometrics* appeared to have increased sharply: rising from 0.941 in 1978-1990 to 9.014 in 2003-2014.

Micro-level structure analysis

Micro-level structure analysis was adopted to measure the individual authors. One of the main purpose of social network analysis is to identify the core actors in a network. We applied four measures (raw degree, degree centrality, betweenness centrality, and closeness centrality) to investigate the structural characteristics of each author in each timespan.

Table 2 summarized the top 10 authors with highest degrees in each time period. Freeman (1978) defined the degree of a point as the number of other points to which a given point is adjacent. In the collaboration networks, the degree of an author represents the number of authors a given author co-authored with before. Schubert A held the highest degree with 17 in the first period, which showed he cooperated with 17 authors between 1978 and 1990. In both

second and third timespan, Glänzel W. achieved the first place with 49 and 123 collaborators in 1991-2002 and 2003-2014, respectively.

Table 2. Raw degree (top 10 authors) in each time period.

1978-1990		*1991-2002*		*2003-2014*	
Schubert, A	17	Glänzel, W	49	Glänzel, W	123
Braun, T	15	Schubert, A	42	Chen, DZ	78
Zsindely, S	12	Braun, T	37	Huang, MH	78
Moed, HF	7	Moed, HF	33	Debackere, K	59
Vanraan, AFJ	7	Gupta, BM	30	Zhang, X	57
Burger, WJM	6	Gomez, I	26	Rousseau, R	56
Courtial, JP	6	Courtial, JP	24	Gorraiz, J	52
Frankfort, JG	6	Rivas, AL	23	Thijs, B	52
Lepair, C	6	Dore, JC	21	Abramo, G	51
Lancaster, FW	5	Miquel, JF	21	D'Angelo, CA	49

Apart from the raw degree of the actors, the centrality is one of the most important structural attributes of social networks (Freeman, 1978). Over the past years, a number of centrality measures have been proposed by sociologists. In the case of co-authorship network, each centrality measure demonstrate special characteristics of the author cooperation. The centrality indicators are designed to identify the "core" authors from different perspectives. The degree centrality can be seen as an index of its potential communication activity. For the co-authorship network, the authors with high degree centrality may result in the status of "elite" (Yin et al., 2006). Freeman's (1978) betweenness centrality is based upon the frequency with which a point falls between pairs of other points on the shortest or geodesic paths connecting them. Regarding to the collaboration, betweenness centrality can be used to assess the potential of an author for control of communication in the knowledge flow network. Tables 3 and 4 summarized the top 10 authors with the highest degree and betweenness centralities in each time period, respectively.

From Table 3, we can see that authors with high degree centrality were dynamic in different timespans. New authors arrived in a field and gathered more collaborations, whereas the existing authors decayed, to some extent, with time. No author ranked in the top 10 in all three time periods. From the perspective of potential communication ability, the "star" of the collaboration networks changed over time. When it comes to the betweenness centrality, Glänzel W was no doubt the core author in both the second and third time periods. Interestingly, from both dimensions (degree centrality and betweenness centrality), Glänzel W occupied the genuine dominator (or "star") position from 2003 to 2014, which suggests that he possesses potential communication ability as well as the possible ability to control the communication between other authors in recent years.

Collaboration network visualization

Figures 4 to 6 present the evolution of the collaboration network in the three stages. Clearly, both the number of the authors and the collaborations boosted, which also illustrated the expansion of this field. With the time advanced, the collaborations between authors were strengthened. To highlight the changes in collaboration, we removed removed isolated nodes in the network in both Figures and displayed only the collaborating authors and their connections. The size of both the nodes and the labels indicated the degree of the authors. The strength of the collaboration was shown by the thickness of the ties between nodes. The authors with high degree in Table 2 were outstanding in the networks.

Table 3. Degree centrality (top 10 authors) in each time period.

1978-1990		*1991-2002*		*2003-2014*	
Courtial, JP	1.379	Moed, HF	1.846	Glänzel, W	1.419
Lepair, C	1.379	Courtial, JP	1.652	Rousseau, R	1.387
Lancaster, FW	1.149	Gupta, BM	1.458	De Moya-Anegon, F	0.967
Braun, T	0.92	Rousseau, R	1.458	Ho, YS	0.935
Dobrov, GM	0.92	Tijssen, RJW	1.458	Borner, K	0.903
Krebs, M	0.92	Glänzel, W	1.361	Park, HW	0.838
Nagy, JI	0.92	Gomez, I	1.263	Thelwall, M	0.838
Plagenz, K	0.92	Rivas, AL	1.263	Chen, DZ	0.838
Porta, MA	0.92	Deshler, JD	1.166	Wu, YS	0.806
Schubert, A	0.92	Gonzalez, RN	1.069	Debackere, K	0.806

Table 4. Betweenness centrality (top 10 authors) in each time period.

1978-1990		*1991-2002*		*2003-2014*	
Braun, T	0.017	Glänzel, W	1.408	Glänzel, W	5.478
Nagy, JI	0.016	Kretschmer, H	1.1	Rousseau, R	3.918
Courtial, JP	0.012	Moed, HF	1.017	Park, HW	2.17
Lepair, C	0.01	Gupta, BM	0.855	Leydesdorff, L	1.661
Schubert, A	0.007	Rousseau, R	0.489	Kretschmer, H	1.478
Dobrov, GM	0.005	Tijssen, RJW	0.397	Ho, YS	1.423
Inhaber, H	0.005	Gomez, I	0.351	Chen, J	1.374
Narin, F	0.005	Luwel, M	0.262	Meyer, M	1.284
Lancaster, FW	0.004	Braun, T	0.261	Huang, JS	1.219
Studer, KE	0.004	Schubert, A	0.259	Aguillo, IF	1.218

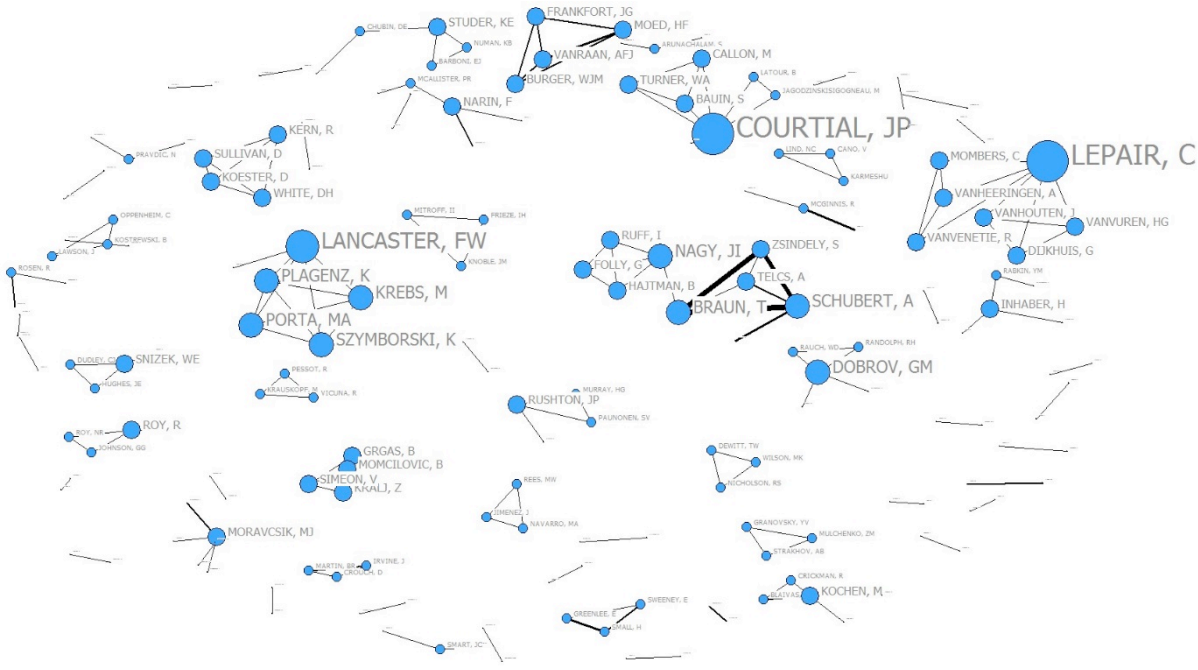


Figure 4. The collaboration networks in 1978-1990.

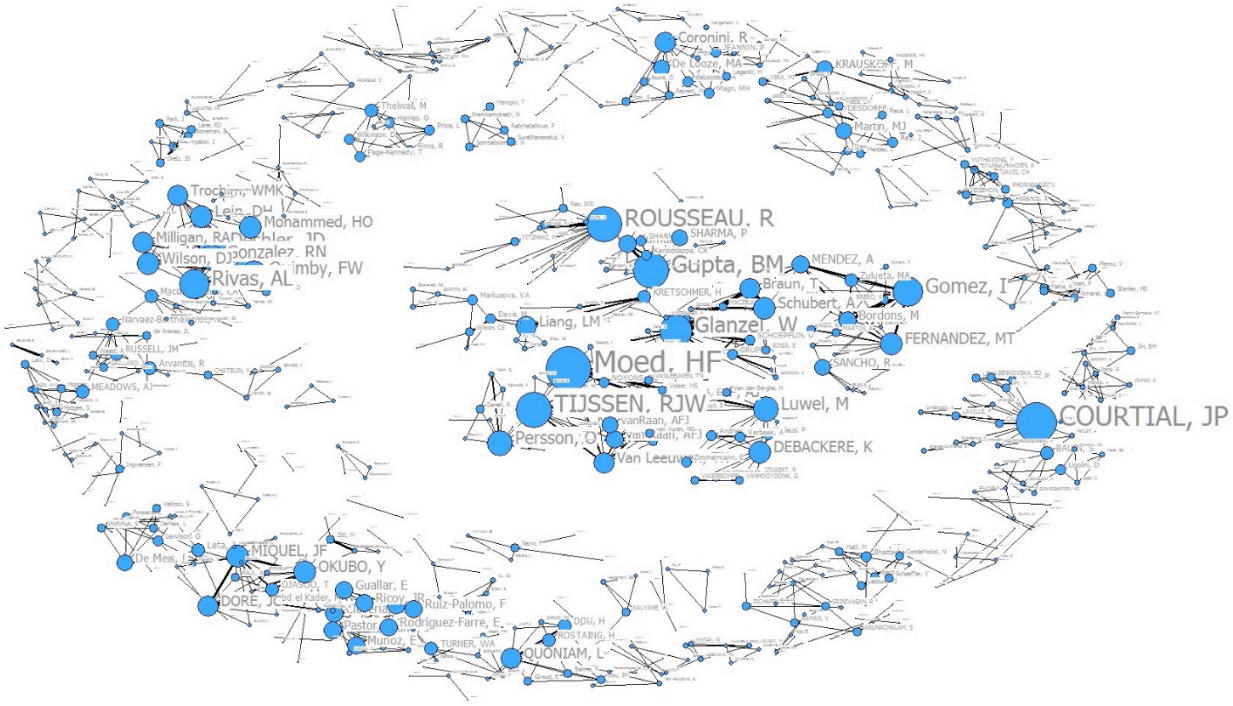


Figure 5. The collaboration networks in 1991-2002.

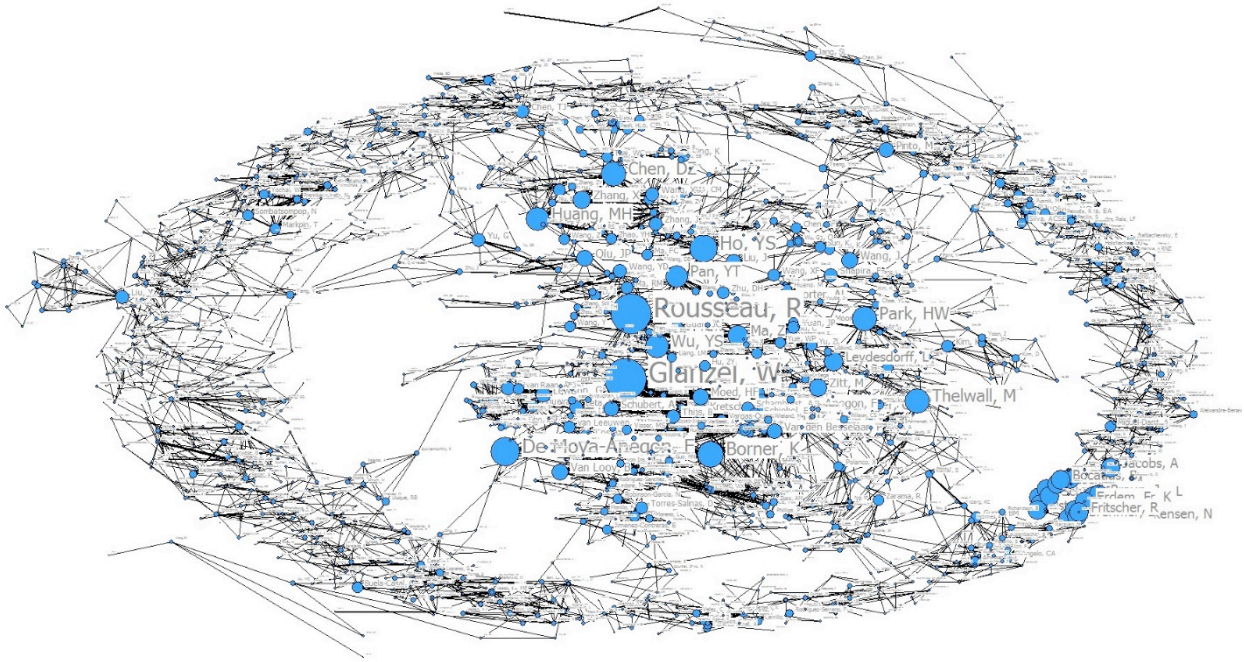


Figure 6. The collaboration networks in 2003-2014.

Conclusion

This paper approached the evolution of the scientific collaboration networks of scientometrics based on the publications in *Scientometrics*. The past 37 years were divided into three timespans: the first time period (1978-1990), the second period (1991-2002), and the third period (2003-2014). Based on the macro-level structure analyses, the degree distribution of the collaboration of *Scientometrics* in each timespan were consistent with power-law, and both the average degree and average distance steadily increased with time, which

demonstrated that the cooperation between authors was getting more frequent. Micro-level structure analyses illustrated the authors with high performance in raw degree measure, degree centrality measure, and betweenness measure were dynamic in different timespans. Interestingly, on each dimension, Glänzel W became the genuine dominator (or "star") in the most recent period: 2003-2014. Finally, the visualization of the evolution of the collaboration network in three stages was presented, and the boosts of the number of authors and their collaborators were displayed in the network graphs.

References

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97. doi:10.1103/RevModPhys.74.47

Ardanuy, J. (2012). Scientific collaboration in Library and Information Science viewed through the Web of Knowledge: the Spanish case. *Scientometrics*, *90*(3), 877–890.

Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and Its Applications*, *311*(3-4), 590–614.

Borgatti, S.P., Everett, M.G. & Freeman, L.C. (2002). *Ucinet 6 for Windows: Software for Social Network Analysis*. Harvard, MA: Analytic Technologies.

Breiger, R. L. (1974). The duality of persons and groups. *Social Forces*, *53*(2), 181–190.

Chen, Y., Börner, K., & Fang, S. (2013). Evolving collaboration networks in Scientometrics in 1978–2010: a micro–macro analysis. *Scientometrics*, *95*(3), 1051–1070.

Ding, Y., Rousseau, R., & Wolfram, D. (Eds.). (2014). *Measuring Scholarly Impact: Methods and Practice*. New York: Springer.

Franceschet, M. (2011). Collaboration in computer science: A network science approach. *Journal of the American Society for Information Science and Technology*, *62*(10), 1992–2012. doi:10.1002/asi.21614

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239.

Hou, H. (2006). Study on the evolution of scientometrics based on the scientific map. Retrieved from http://www.cnki.net/

Nalimov, V.V. & Mulchenko, B.M. (1969). *Scientometrics*. Moscow: Nauka (in Russian).

Newman, M. (2001). Scientific collaboration networks. Network construction and fundamental results. *Physical Review E*, *64*(1), 016131.

Pang, J. (2002). *The Research Methodology of Scientometrics*. Beijing, China: Scientific and Technical Documentation Press.

Schoepflin, U., & Glänzel, W. (2001). Little Scientometrics, Big Scientometrics…and Beyond? *Scientometrics*, 30: 375-384.

Schubert, A. (2002). The Web of Scientometrics: A statistical overview of the first 50 volumes of the journal. *Scientometics*, 53(1):3-20.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*(6684), 440–442.

Yin, L., Kretschmer, H., Hanneman, R. A., & Liu, Z. (2006). Connection and stratification in research collaboration: An analysis of the COLLNET network. *Information Processing & Management*, *42*(6), 1599–1613.

Yuan, J. (2010). *The Advanced Tutorial of Scientometrics*. Beijing, China: Scientific and Technical Documentation Press.

Open Access Publishing and Citation Impact - An International Study

Thed van Leeuwen, Clifford Tatum, and Paul Wouters

leeuwen@cwts.nl, c.c.tatum@cwts.leidenuniv.nl, p.f.wouters@cwts.leidenuniv.nl
CWTS, Leiden University, Wassenaarseweg 62a, Leiden (the Netherlands)

Abstract
This paper describes the analysis of open access (OA) publishing in the Netherlands in an international comparison. As OA publishing is now actively stimulated by Dutch science policy, similar to the UK, a bibliometric baseline measurement is conducted to assess the current situation, to be able to measure developments over time. For the study we collected data from various sources, and for three different smaller European countries (the Netherlands, Denmark, and Switzerland). Not all of the analyses for this baseline measurement are included here; the analysis presented in this paper mainly focuses on the various ways OA can be defined while using Web of Science, and the problems with interpreting these results. From the data we collected, we can conclude that the way OA is currently registered in various electronic bibliographic databases is quite unclear, and various methods applied deliver results that are different, although the impact scores point in the same direction.

Conference Topic
Journals, databases, and electronic publications

Introduction
Acceleration of open access goals in the Netherlands coincides with implementation of new current research information systems (CRIS) at Dutch universities and research institutes. This deployment of institutional CRIS systems provides an opportunity for national level tracking of open access through coordinated metadata schemes and common registration practices. As open access is notoriously difficult to measure, contemporary analyses often employ random sampling techniques (Archambault et al., 2014; Björk et al., 2010). All publication records in a given sample are tested to determine the proportion of full texts that are open access publications. National level coordination of research information provides an opportunity for improved, more precise assessment of open access publishing. In this study we use bibliographic data to establish a baseline analysis of the proportion of open access publishing in the Netherlands.

Assessment of open access publishing is complicated by a growing diversity of what counts as open access, the copyright restrictions for when a publication can be made openly accessible, and the lack of clear and consistent identification of open access publications in bibliographic data. To examine these challenges we begin with a definition from the Budapest open access Initiative (BOAI):

> Free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited. (BOAI 2002)

This definition highlights two distinct channels of access: (1) human access to read, download, and reuse the full text of published articles; and (2) machine access to crawl, index, or analyze the content of articles. The BOAI also proposes two operational paths to access through open access journals and self-archiving in repositories, subsequently referred to as

Gold open access and Green open access (Bailey, 2005). Hybrid open access generally refers to the situation whereby authors can pay to make their articles in subscription journals openly accessible on the Web (Björk, 2012).

In addition to the broad categories of Gold, Green, and Hybrid modes of open access, multiple versions of a manuscript may exist due to variations in publishers' licensing agreements. These agreements typically specify how, when, and under which conditions a manuscript may be openly accessible on the web. For example, a publisher may allow Green open access through self-archiving in an institutional repository. However, publishers' copyright restrictions differ on the stage of manuscript development that may be openly accessible, thus assigning different rights to different versions of the text. Commonly specified version types include the submitted manuscript (before peer review), the accepted manuscript (peer-reviewed but not formatted), and an exact copy of the published manuscript (Björk et al., 2013). This creates the possibility that the open access version of a manuscript is substantively different from the published version. In such instances, it is unclear whether the open access version has been sufficiently validated through the quality control measures such as peer review.

Another variation is delayed access, which is applied as an embargo period, after which a copy of the publication may be self-archived or the publisher may remove access restrictions on the journal website. Embargo periods are generally specified as a delay of 6, 12, 18, or 24 months after publication, with 12 months being the most common embargo period (Laakso & Björk, 2013). For Green open access, it is thus left to authors and institutions to track and manage a variety of self-archiving policies, which in itself has been shown to be a barrier to open access (Davis & Connolly, 2007). However, this kind of administrative overhead is largely absent from subscription journals that convert articles to open access after a specified delay (e.g. 12 months). In addition, a bibliometric analysis of 'delayed access' journals found journal and article impact factors higher than comparable averages from both subscription journals and direct (no delay) open access journals (Laakso & Björk, 2013).

A common refrain among proponents of open access is that open access publishing yields increased citation impact. While there are conflicting reports regarding an open access citation advantage (OACA), heightened attention to this issue has increased our understanding about citation behaviour more generally. Numerous bibliometric studies claim that open access publishing results in a significant increase in citations. In these studies the size of advantage varies widely based on a variety of issues, such as disciplinary differences, methodological approaches, variation in how open access is defined, and difficulty in determining when an article is made openly accessible (Swan, 2010). In addition, a number of confounding factors have been shown to influence citation frequency such as early exposure to draft versions of a manuscript (Moed, 2007), self-selection bias whereby an author may choose open access for only her best publications (Kurtz et al., 2007), the availability at multiple access points (Xia, Myers & Wilhoite, 2011), and physical proximity of researchers (Lee et al., 2010).

To control for these factors, Davis et al. (2008) employ randomized controlled trial methods, whereby randomly selected articles in subscription based journals are switched to open access. The resulting configuration is similar to hybrid open access, such that the article is made to be openly accessible and is listed among the non-open access articles on the journal's website. In the Davis et al. (2008) study a citation advantage was not present. However, the research design used to control for confounding variables (randomized controlled trial) also limited applicability of the findings to the hybrid model of open access. More recently, Archambault et al. (2014) show variation in the accumulation of citations associated with the different modes of open access. The authors find a citation *advantage* most prominently associated with the self-archiving mode of open access (Green OA) and a citation

disadvantage associated with full and immediate open access journals (Gold OA). This study also establishes a general ranking of citation accumulation on the bases of open access, listed in order of most to least: Green OA, Other OA, Not OA, and Gold OA." (Archambault et al., 2014, pp. 20, 24)

To address the variability of circumstances associated with open access publishing, recent studies invert the research design from top-down queries of bibliometric datasets to bottom-up testing whether a publication is an open access publication. This approach involves random sampling of a given publishing domain, harvesting full-texts from the Internet, and analysis of available metadata from harvested manuscripts (Björk et al., 2010). While this approach circumvents much of the variability noted above, it is nevertheless dependent on the presence and quality of metadata. (The potential for improved metadata practices is addressed in the discussion section below.)

The objective of our analysis is to show the challenges of bibliometrically analysing OA publications and associated impact scores. We use Web of Science (WoS) data, either directly retrieved from the database, or combined with article-level data extracted from journals listed in the Directory of Open Access Journals (DOAJ). As both data sources are incomplete with respect to open access publications, the analysis is focused on comparison of relative output and relative impact among three European countries of similar size and scientific production: the Netherlands, Denmark, and Switzerland, in order to show developments in time, as well as differences resulting from both approaches. It is important to note that Green OA articles are excluded from our analysis. While the Netherlands maintains a robust national repository for Green OA (NARCIS), there is not yet a reliable system of identifying the self-archived state of publications within bibliometric datasets. As such, the proportion of open access and associated impact comparisons are limited to the available data on Gold OA.

Data collection

In the study we make use of data from various sources. The Web of Science (WoS) database is used in its internet version, available to most Dutch researchers. We also used the CWTS version of the WoS, a tailor-made database based upon state-of-the-art bibliometric techniques and indicators. In this version, the functionality to search for OA output is not yet available. Finally, we make use of the journals and the publications listed in the Directory of Open Access Journals (DOAJ). From this data source, we will further focus on the digital object identifiers (DOIs), while leaving out other elements (such as the license types, as this information is unclearly defined as well as unclearly linked to the publications).

Method I: The first way of data collection from WoS starts from the desktop interface of the WoS database. The functionality to collect this information is not yet available in the in-house WoS database at CWTS, so therefore we had to collect these data from the internet version directly. This approach involved the following steps:

1) Collect the output of one of the selected countries for a particular year;
2) Within that set, further distinguish the OA part of that selected output;
3) Download these publications from the WoS database (including the so-called UT-code, a unique identifier within WoS that allows for linking to the CWTS WoS database);
4) Select within the CWTS database the output for the three countries;
5) Match the selected output from the Internet version of the WoS with the in-house CWTS version;
6) Create two sets within the CWTS database, an OA formatted set of publications, and a non OA formatted set of publications.

These steps were taken for all three countries, collecting publications from 2000-2013.

The definition of how the publications were defined as OA is based upon the following statement on the WoS database' website: "The Thomson Reuters Links open access Journal

Title List includes free journal content that are available for linking from the Web of Science."

Method II: The second method started from the Directory of Open Access Journals (DOAJ). This list contains journals that have implemented the Gold open access business model. CWTS has downloaded the complete list, and all publications published in the journals on the DOAJ list. By making use of this dataset, we could use a second approach to the OA output of the three countries taking the following steps:

1) First select within the CWTS database the output for the three countries;
2) Collect their Digital Object Identifiers (doi);
3) Match these with the doi's of the publications downloaded from the DOAJ list;
4) Create two sets within the CWTS database, an OA formatted set of publications, and a non OA formatted set of publications.

We focused on articles, letters and reviews only, excluding other types of documents such as editorials, meeting abstracts, book reviews, etc. The choice for these types is based upon the importance of these three types in communicating scientific findings among peers, and their relative homogeneity within the system.

Methods

In the study we present a number of indicators. In cases we present numbers of publications, this is indicated with a P. In case citation data are presented, we use MNCS (Mean Normalized Citation Score), as well as the MNJS, the field normalized journal impact indicator, to indicate the normalized impact scores in the study (Waltman et al., 2011a; Waltman et al., 2011b). While the output indicator can be used for the various electronic systems we use in the study, and P can relate to various document types analysed, the citation impact indicators are used only within the context of the WoS database. In case of the impact indicators, the length of the citation window is one year longer than the presented year block (so in case of the last block, 2009-2012, the citation impact is measured up until 2013, currently the last year fully covered in the CWTS WoS database).

Results

First we present the results from Method I, described above. The output numbers of the three countries according to the methodology I are found in Table 1 along with the two separate parts of the output, distinguished by openness. The analysis covers the period 2000 up until 2012 for publication data, and up until 2013 for citation impact data. In this analysis we use moving publication year windows, in order to create more solid and stable trend lines, as we are more interested in the trends than in variation from year to year.

The data presented in Table 1 clearly show that OA publishing is becoming increasingly important, in all three selected countries. The Netherlands is lagging somewhat behind Denmark and Switzerland, albeit with only a small part of the total output.

In Figure 1, we have distinguished between the open access format output of the three countries (indicated by the 'Ex OA' label to the country names). What we observe are increasing trends for the parts of the output not published in OA format, which is also visible for the OA format of the output of these three countries, and as shown above in Table 1, increases somewhat faster for Denmark and Switzerland as compared to the Netherlands.

Table 1. Output (P) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non-OA output, 2000-2012.

	NL Ex OA	NL OA	Share OA	DK Ex OA	DK OA	Share OA	CH Ex OA	CH OA	Share OA
2000 - 2003	75607	712	1%	30616	452	1%	53283	995	2%
2001 - 2004	78087	858	1%	31262	557	2%	54793	1220	2%
2002 - 2005	81849	1180	1%	31972	728	2%	56982	1836	3%
2003 - 2006	85386	1663	2%	33024	949	3%	60319	2217	4%
2004 - 2007	88745	2349	3%	34082	1244	4%	63205	2790	4%
2005 - 2008	92349	3265	4%	35273	1631	5%	65920	3517	5%
2006 - 2009	96278	4269	4%	36672	1997	5%	69518	3912	6%
2007 - 2010	101270	5587	6%	38726	2554	7%	72687	4981	7%
2008 - 2011	106560	7299	7%	41417	3264	8%	76658	6354	8%
2009 - 2012	111990	9504	8%	44264	4420	10%	80786	7990	10%

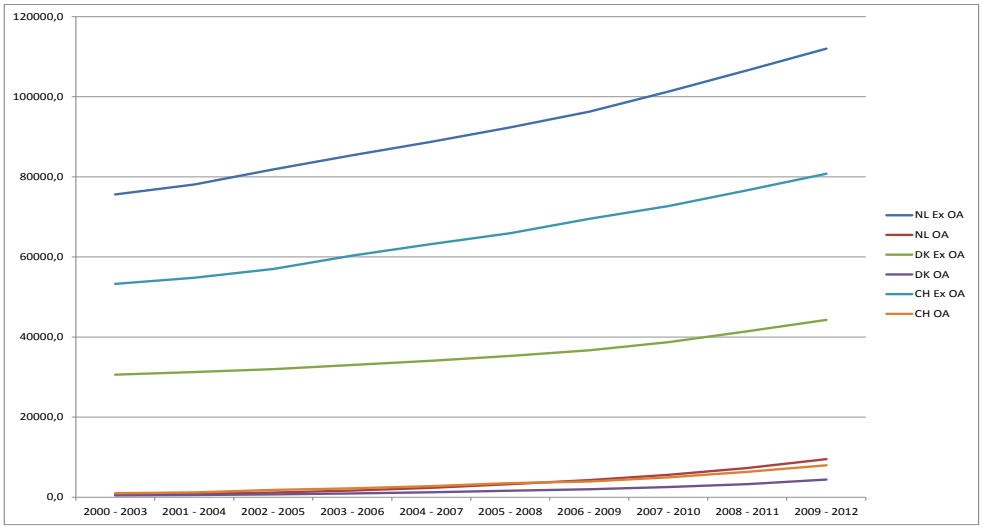


Figure 1. Output development (P) of Denmark, the Netherlands, and Switzerland, 2000-2012/2013.

In Table 2, we present the citation impact scores as represented by the MNCS indicator, the field normalized impact of the outputs of the three countries, again separated by the two types of publication output: open access and non-open access publications.

Figure 2 shows that for all three countries the non-OA part of the output has a citation impact well above world average, with Switzerland topping the other two countries, which have a nearly equal field normalized impact score. The impact of OA publications is lower for all three countries. The impact of the OA part of the national outputs of Denmark and Switzerland were initially well above world average. This is also the case for Swiss publications, as the OA format published output is lower on MNCS only from 2007-2010/2011 onwards. In case of Denmark, this drop started somewhat earlier, while in the case of the Netherlands, the OA output never got an impact higher than that of the non-OA format output. Another interesting phenomenon is the increase of the gap between the impact of OA and non-OA output. This is particularly the case for Switzerland and Denmark, where we observe a clear drop of the impact of OA format output compared to their non-OA formatted output, and to a lesser extent for the Netherlands, where the two impact lines are more slowly diverging.

Table 2. Citation impact (MNCS) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non-OA output, 2000-2012.

	NL Ex OA	NL OA	DK Ex OA	DK OA	CH Ex OA	CH OA
2000 – 2003	1,29	0,99	1,30	1,03	1,37	1,11
2001 - 2004	1,30	0,95	1,29	1,31	1,35	1,21
2002 - 2005	1,30	0,99	1,29	1,39	1,36	1,36
2003 - 2006	1,31	1,07	1,31	1,34	1,36	1,46
2004 - 2007	1,30	1,12	1,31	1,30	1,38	1,47
2005 - 2008	1,31	1,13	1,32	1,30	1,39	1,48
2006 - 2009	1,35	1,15	1,34	1,26	1,39	1,39
2007 - 2010	1,38	1,17	1,37	1,26	1,42	1,37
2008 - 2011	1,40	1,18	1,40	1,25	1,46	1,36
2009 - 2012	1,44	1,18	1,44	1,18	1,50	1,33

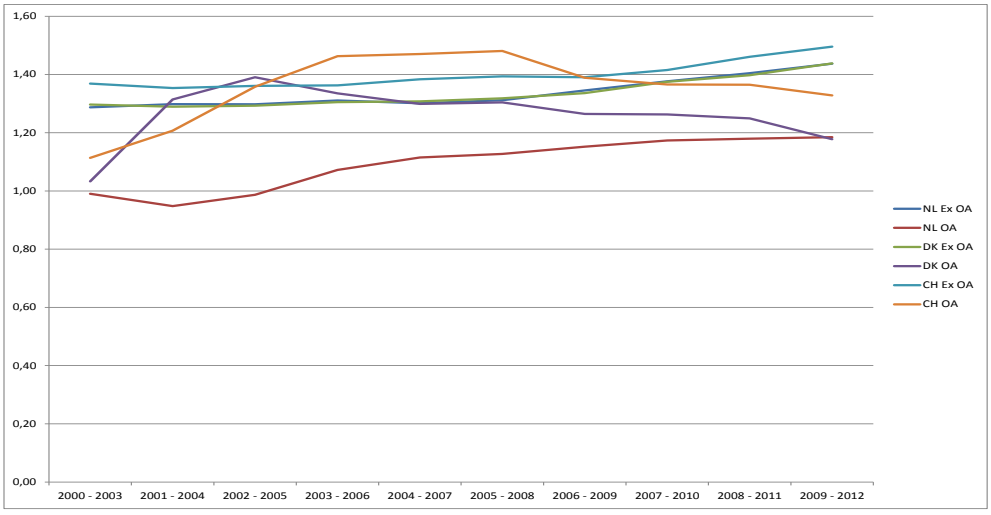


Figure 2. Impact development (MNCS) of Denmark, the Netherlands, and Switzerland, 2000-2012/2013.

If we shift our focus towards the journal impact analysis (see Table 3 and Figure 3), for which we use the indicator MNJS, we see an even more interesting phenomenon. While the output in non-OA format published journals shows a choice for journals with increasing impact scores, the OA format published outputs end up in journals with decreasing field normalized impact scores. We even notice a diverging trend in these two clusters of trend lines: non-OA format published journals tend to show increasing impact scores, while OA format published journals show decreasing impact trends. This is striking since these are three of the 'scientifically stronger' nations, as far as can be measured with bibliometric instruments.

Here we start with the results from methodology II. The results of the output analysis are shown in Table 4, which again covers a similar distinction between OA and non-OA format output, but now according to the definition described above under Method II. We combined the DOIs of journals on the DOAJ list with the DOIs available in the WoS. From the total set of 787,611 DOIs in the DOAJ list, we matched 226,641 publications in WoS on the basis of available DOIs. The reason for this seemingly low recall is twofold. In the first place, not all journals covered by the DOAJ list are processed for the WoS database, and secondly, not all publications in journals covered in WoS do contain DOIs. This means that for some journals that are both covered in the DOAJ list as well as in WoS, a match is impossible, particularly

for the earlier years in the analysis. Like the first methodology we followed, we separated the OA format published output from the Netherlands, Denmark, and Switzerland from the total set of publications for the three countries under study.

Table 3. Journal-to-field citation impact (MNJS) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non-OA output, 2000-2012

	NL Ex OA	NL OA	DK Ex OA	DK OA	CH Ex OA	CH OA
2000 - 2003	1,18	0,95	1,15	0,84	1,19	1,06
2001 - 2004	1,19	0,97	1,16	1,02	1,20	1,03
2002 - 2005	1,19	1,00	1,16	1,08	1,20	1,19
2003 - 2006	1,20	1,06	1,16	1,11	1,20	1,20
2004 - 2007	1,22	1,09	1,18	1,12	1,22	1,11
2005 - 2008	1,24	1,09	1,20	1,10	1,24	1,14
2006 - 2009	1,26	1,11	1,22	1,07	1,26	1,11
2007 - 2010	1,29	1,11	1,25	1,06	1,29	1,11
2008 - 2011	1,30	1,10	1,26	1,05	1,31	1,11
2009 - 2012	1,32	1,09	1,28	1,00	1,33	1,09

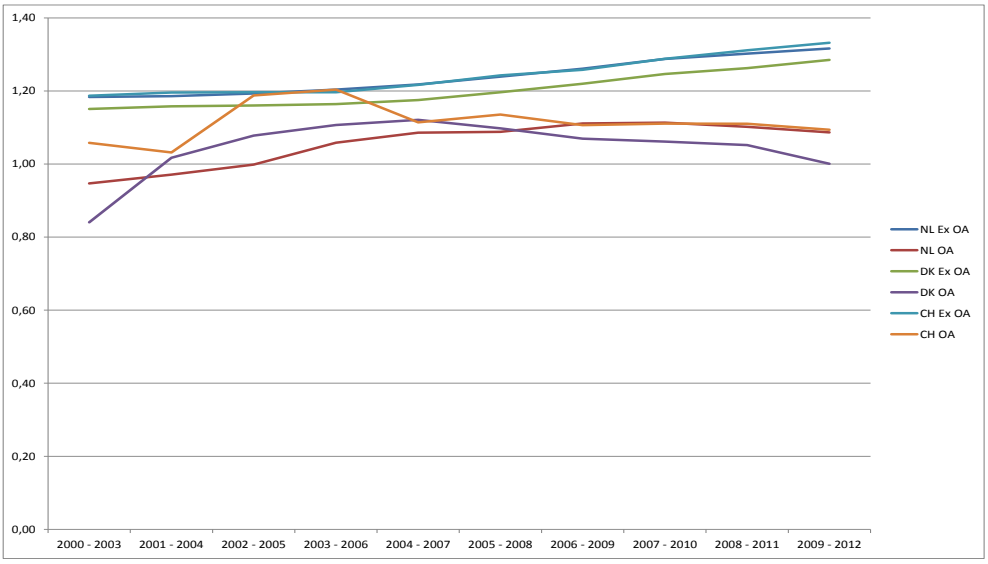


Figure 3: Journal impact development (MNJS) of Denmark, the Netherlands, and Switzerland, 2000-2012/2013.

First of all, we observe that the overlap between the DOAJ list/WoS combinations with Dutch/Danish/Swiss publications in WoS is much smaller compared to the previous analysis on Dutch/Danish/Swiss output in OA format, which is most likely the result of the missing DOIs in the WoS database. If we compare the results of Table 1 with those presented in Table 4, we find much lower shares of OA output compared to the overall output of the three countries. This is further underlined by Figure 4, in which the OA format output of the three countries is at the low end of the graph, while we simultaneously observe a strong increase in the output of the non-OA format output of the three countries.

Table 4. Output (P) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non-OA output (based on DOI-matching), 2000-2012

	NL Ex OA	NL OA	Share OA	DK Ex OA	DK OA	Share OA	CH Ex OA	CH OA	Share OA
2000 - 2003	75607	10	0%	30616	4	0%	53283	2	0%
2001 - 2004	78087	35	0%	31262	25	0%	54793	30	0%
2002 - 2005	81849	136	0%	31972	83	0%	56982	97	0%
2003 - 2006	85386	344	0%	33024	170	1%	60319	232	0%
2004 - 2007	88745	648	1%	34082	312	1%	63205	420	1%
2005 - 2008	92349	1068	1%	35273	486	1%	65920	690	1%
2006 - 2009	96278	1531	2%	36672	664	2%	69518	972	1%
2007 - 2010	101270	2207	2%	38726	924	2%	72687	1461	2%
2008 - 2011	106560	3036	3%	41417	1231	3%	76658	2062	3%
2009 - 2012	111990	3896	3%	44264	1595	4%	80786	2608	3%

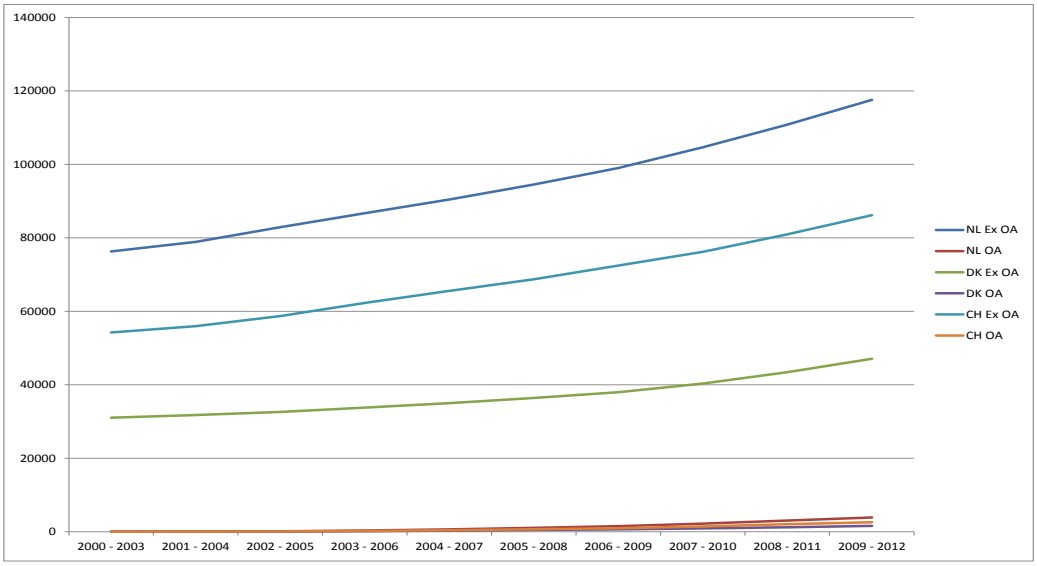


Figure 4. Output development (P) of Denmark, the Netherlands, and Switzerland, based on matching of DOI's, 2000-2012/2013.

In Table 5, we present the impact scores of the three countries, again distinguishing OA format output and non-OA format output. Again we observe lower impact scores for the OA format output of the three countries, except for the starting block of the analysis (please note that the output numbers are extremely low in this part of the analysis for the Netherlands and Denmark, respectively 10 and 4 papers). From the second year block onwards, we observe increasing trends in the impact of the OA format of the three countries, although we must stress that this is also the case for the non-OA format output of the three countries.

Figure 5 shows this stable development of both sets of publications in time, whereby the impact scores are increasing on both sets, although the 'difference' remains more or less the same between the two sets of scores.

In Table 6 we present the outcomes of the analysis on the journal impact scores, based upon methodology II. Here we observe, similar to the previous outcomes, fluctuations in the initials years of the analysis for the OA format output, followed by a more stable situation from 2005-2008 onwards. This finding is even more visible in the graphical representation of Table 6, as in Figure 6.

Table 5. Citation impact (MNCS) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non-OA output (based on DOI-matching), 2000-2012

	NL ex OA	NL OA	DK ex OA	DK OA	CH ex OA	CH OA
2000 - 2003	1,28	1,65	1,29	1,32	1,36	
2001 - 2004	1,29	0,87	1,29	0,91	1,35	1,03
2002 - 2005	1,29	0,87	1,30	0,98	1,36	1,18
2003 - 2006	1,31	0,87	1,31	0,78	1,37	0,95
2004 - 2007	1,30	0,75	1,31	0,72	1,39	0,96
2005 - 2008	1,31	0,83	1,32	0,86	1,40	0,91
2006 - 2009	1,35	0,85	1,34	0,89	1,40	0,92
2007 - 2010	1,38	0,90	1,38	0,96	1,42	0,97
2008 - 2011	1,40	0,97	1,40	1,00	1,46	1,07
2009 - 2012	1,43	1,03	1,43	0,96	1,49	1,06

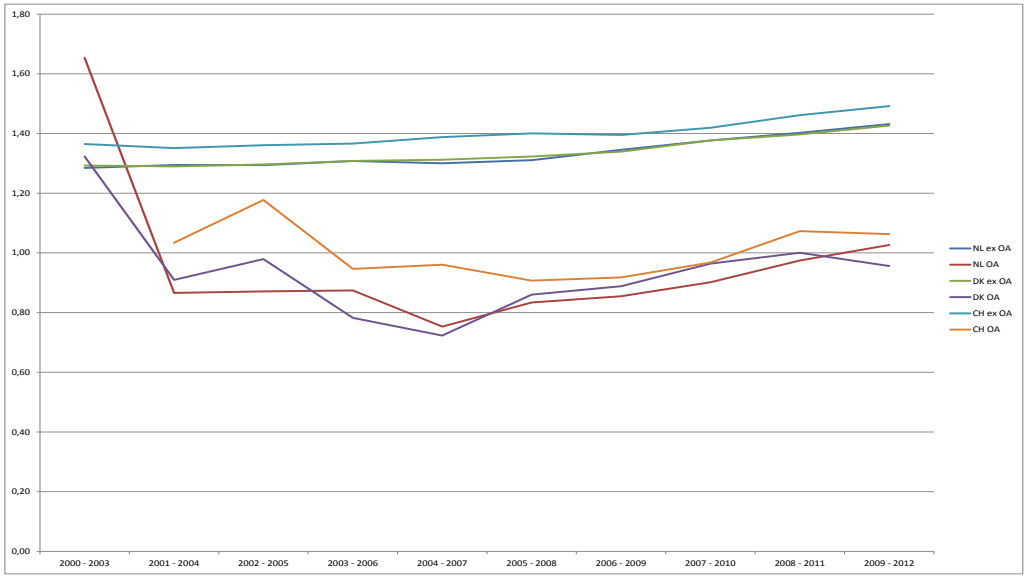


Figure 5. Impact development (MNCS) of Denmark, the Netherlands, and Switzerland, based on matching of DOIs, 2000-2012/2013.

Table 6. Journal-to-field citation impact (MNJS) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non-OA output (based on DOI-matching), 2000-2012

	NL ex OA	NL OA	DKex OA	DK OA	CH ex OA	CH OA
2000 - 2003	1,18	0,54	1,15	1,28	1,19	0,24
2001 - 2004	1,18	0,84	1,16	0,92	1,19	1,22
2002 - 2005	1,19	0,77	1,16	0,84	1,20	1,00
2003 - 2006	1,20	0,84	1,16	0,79	1,20	0,90
2004 - 2007	1,22	0,86	1,18	0,83	1,22	0,88
2005 - 2008	1,24	0,88	1,20	0,86	1,24	0,86
2006 - 2009	1,26	0,90	1,22	0,87	1,26	0,87
2007 - 2010	1,29	0,94	1,24	0,91	1,29	0,91
2008 - 2011	1,30	0,97	1,26	0,93	1,31	0,96
2009 - 2012	1,31	0,97	1,27	0,92	1,32	0,97

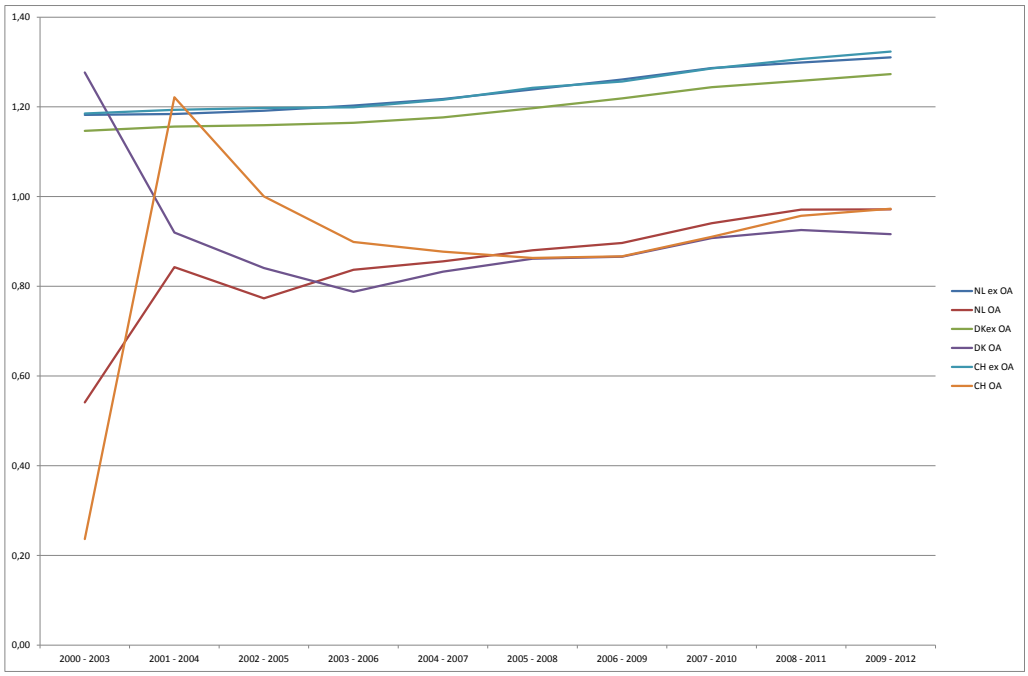


Figure 6: Journal impact development (MNJS) of Denmark, the Netherlands, and Switzerland, based on matching of DOI's, 2000-2012/2013

Conclusion and Discussion

In this final part of the paper, we will summarize the main bibliometric findings, and then move towards limitations in the ways OA is now disclosed in electronic systems supporting bibliometric analyses. Finally, we will discuss the need to improve identification of open access publications and the use of bibliometric techniques to measure OA.

Please note that our conclusions are mainly related to the domains in which journal publishing is the dominant way of communication (the natural, life and medical sciences, and to a lesser extent the social sciences and humanities (van Leeuwen, 2013). We observe for the three countries that the share in output in OA journals is lagging behind as compared to the journals that maintain the non-OA format. We observe a divergence in the development of citation impact for (Gold) OA and non-OA publications with consistently lower impact for the OA publications.

Second, we observe that OA journals have lower journal impact scores than non-OA journals. This may mean that they still struggle to find their position within the total 'reputational hierarchy' of the domain, and as such also within the WoS database. This is a common problem for new journals, and OA journals are no exception. It should be noted however, that our findings associated with OA impact are consistent with what others have found: Gold OA is associated with no citation advantage or a disadvantage (e.g. Archambault et al., 2014). With the inclusion of the various forms of Green OA, we would expect to find a larger proportion of open access articles and a more nuanced outcome related to impact. That Green OA has been found to have increased accumulation of citations (Archambault et al., 2014), may be associated with the circumstances identified above as confounding factors (e.g. early exposure, multiple access points, and proximity of researchers).

Third, we may need to worry about the role of peer review in the journals that are part of the expansion of the WoS database in the last couple of years, many of which are in the OA segment of the database. The Institute for Scientific Information, the predecessor of the current owner of the WoS database Thomson Reuters, always clearly indicated that a properly functioning peer review system within a journal was one of the conditions for a journal to be included in the system (next to other criteria, such as international focus, regular appearance,

preferably in the English language, etc.). We do not know whether this is still such a strong criterion, particularly given the fact that so many new journals appeared around the OA development.

A fourth conclusion relates to the messy situation around the various manners by which open access is defined in electronic databases. The two different ways open access can be operationalized within the world of WoS is an example of this unclear and somewhat messy situation. The fact that the Scopus database did not have the functionality to clearly define open access for users of the system is another instance of the situation around open access. Further examples of this lack of clarity are the various ways open access is operationalized by the publishing industry. There is no clear way of operationalizing in the larger databases of the various business models (such as Gold, Green, and Hybrid open access). Yet another example relates to the various license types related to open access.

A recently published metadata standard for open access holds some promise for improving both human and machine identification of open access publications (Carpenter, 2013). Here, too, stakeholders involved in the new standard were unable to agree on a precise definition of open access. Instead, the standard specifies metadata elements for *free to read* and *license reference*, the latter of which should point to copyright information publicly accessible on the Web (NISO 2015). Increased attention to national research assessment and increased use of institutional CRIS systems together provide a potentially welcoming context for implementing new metadata practices. This would ideally include the possibility of tracking open access among the diversity of research outputs maintained by CRIS systems and considered in assessment events. In this context, it becomes important to assign openly accessible, persistent identifiers to all research objects (Tatum & Wouters 2014). This would increase the potential use of institutional research information for tracking open access as part of regular research assessment practices, rather than relying solely on estimation derived from random sampling of commercial datasets.

References

Archambault, E., Amyot, D., Deschamps, Nicol, A., Provencher, F., Rebout, L. & Roberge, G. (2014). Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels—1996–2013. *Rapport, Commission Européenne DG Recherche & Innovation; RTD-B6-PP-2011-2: Study to Develop a Set of Indicators to Measure Open Access.*

Björk, B.-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T. & Guðnason, G. (2010). Open Access to the Scientific Journal Literature: Situation 2009. *PLoS ONE,* 5 (6): e11273. doi:10.1371/journal.pone.0011273.

Björk, B.-C. 2012. The Hybrid Model for Open Access Publication of Scholarly Articles: A Failed Experiment? *Journal of the American Society for Information Science and Technology,* 63 (8): 1496–1504. doi:10.1002/asi.22709.

Björk, B-C., Laakso, M., Welling, P. & Paetau, P. (2013). Anatomy of Green Open Access. *Journal of the Association for Information Science and Technology,* 65 (2): 237–50. doi:10.1002/asi.22963.

BOAI. (2002). Budapest Open Access Initiative. *The Open Society Foundations*. http://www.opensocietyfoundations.org/openaccess.

Carpenter, T. (2013). Progress Toward Open Access Metadata. *Serials Review,* 39 (1): 1–2. doi:10.1016/j.serrev.2013.02.001.

Davis, P.M., & Connolly, M.J. L. (2007 March). Institutional Repositories: Evaluating the Reasons for Non-use of Cornell University's Installation of DSpace. http://hdl.handle.net/1813/5195.

Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E. & Murray, S.S. (2005). The Effect of Use and Access on Citations. *Information Processing & Management*, Special Issue on Infometrics, 41 (6): 1395–1402. doi:10.1016/j.ipm.2005.03.010.

Laakso, M., & Björk, B.-C. (2013). Delayed Open Access: An Overlooked High-impact Category of Openly Available Scientific Literature. *Journal of the American Society for Information Science and Technology,* 64 (7): 1323–29. doi:10.1002/asi.22856.

Lee, K., Brownstein, J.S., Mills, R.G. & Kohane, I.S. (2010). Does Collocation Inform the Impact of Collaboration? *PLoS ONE,* 5 (12): e14279. doi:10.1371/journal.pone.0014279.

van Leeuwen, T.N. (2013). Bibliometric research evaluations, Web of Science and the Social Sciences and Humanities: a problematic relationship? *Bibliometrie - Praxis und Forschung*, 1-18. http://www.bibliometrie-pf.de/article/viewFile/173/215

Moed, H.F. (2007). The effect of "Open access" on citation impact: An analysis of ArXiv's condensed matter section. *Journal of the American Society of Information Science & Technology*, 58 (13), 2047-2054

NISO. 2015. *Access License and Indicators NISO RP-22-2015*. National Information Standards Organization.

Swan, A. (2010). The Open Access Citation Advantage: Studies and Results to Date. Technical Report. http://eprints.ecs.soton.ac.uk/18516/.

Tatum, C. & Wouters, P.F. (2014). Next Generation Research Evaluation: The ACUMEN Portfolio and Web Based Information Tools. *OpenAIRE-COAR Conference: Open Access Movement to Reality - Putting the Pieces Together*. Athens. doi:10.6084/m9.figshare.1033681.

Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S. & van Raan, A.F.J. (2011a). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, 5(1), 37-47.

Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S. & van Raan, A.F.J. (2011b). Towards a new crown indicator: An empirical analysis. *Scientometrics*, 87 (3), 467-481.

Xia, J., Myers, R.L. & Wilhoite. S.K. (2011). Multiple Open Access Availability and Citation Impact. *Journal of Information Science,* 37 (1): 19–28. doi:10.1177/0165551510389358.

Measuring the Competitive Pressure of Academic Journals and the Competitive Intensity within Subjects

Ma Zheng[1], Pan Yuntao[2], Wu Yishan[2], Yu Zhenglu[2] and Su Cheng[2]

[1] *mazheng@istic.ac.cn*
Institute of Scientific and Technical Information of China (ISTIC), 15 Fuxing Rd. 100038 Beijing (China); and Nanjing University, School of information Management, 22 Hankou Rd. 210093 Nanjing (China)

[2] *panyt@istic.ac.cn, wuyishan@istic.ac.cn*, *luluyu@istic.ac.cn, sucheng@istic.ac.cn*
Institute of Scientific and Technical Information of China (ISTIC), 15 Fuxing Rd. 100038 Beijing (China)

Abstract
A journal's impact and similarity with rivals is closely related to its competitive intensity. A subject area can be considered as an ecological system of journals, and can then be measured using the competitive intensity concept from plant systems. Based on Journal Citation Reports data from 1997, 2000, 2005, 2010, and 2013, we calculated the mutual citation, cosine similarity, and competitive relationship matrices for mycology journals. We derived the mutual citation network for mycology according to Journal Citation Reports data from 2013. We calculated each journal's competitive pressure, and the competitive intensity for the subject. We found that competitive pressures are very variable among journals. Differences between a journal's absolute and relative influence are related to the competitive pressure. A more powerful journal has lower competitive pressure. New journals have more competitive pressure. If there are no other influences, the competition intensity of a subject will continue to increase. Furthermore, we found that if a subject has more journals, its competitive intensity decreases.

Conference Topic
Journals, databases, electronic publications

Introduction
Scientific and technical (S&T) journals have an important role in science and knowledge dissemination. Journals that are focussed on the same subject are at competition with each other. We must build a favourable competitive environment to realize the optimal allocation of limited resources. At the same time, the "survival of the fittest" mechanism boosts the development of S&T journals.

To build a sustainable environment and competition mechanism, we must analyse and measure the present environment of S&T journals, especially in terms of competition. Many researchers have investigated the competitive environment of S&T journals.

Reaching a consensus on the relationship between the journal environment and competition

Scholars began to study the competitive relationship of journals in the 1920s. Competition is mainly related to the resources of subeditors, editors, and authors. Studies found that competitive power is related to a journals' impact factor (IF) (Campanario 1996). Zhu (1999) discussed the relationship between an S&T journal's quality and competitive spirit. A few years later, scholars proposed that competition is a basic attribute of science and noted the differences between different journals' abilities to secure resources. Powerful journals typically attract more attention, which results in a Matthew effect on the journal's development. Scholars have attempted to measure competition between journals using quantitative indexes (Manfred & Scharnhorst, 2001). Researchers have generally accepted that S&T journals develop within a competitive environment. They have explored definitions of the competition between S&T journals (Cai, 2003), how to increase a journal's core competitive strength (Chen 2005), and how to take advantage of market competition (Gao,

2004). Recently, Leydesdorff, Wagner and Bornmann (2014) focused on competition between highly cited journals dependent on the proportions of most-frequently cited publications in the European Union, China, and the United States, which are represented differently because they use different databases.

Determining the competitive relationship between journals using quantitative methods

Leydesdorff noted that Pearson correlations could be used as similarity measures for citation patterns based on bi-connected graphs (Leydesdorff, 2004). He then used principal component analysis and factor analysis to design indicators for the position of the cited journals in the dimensions of the database (Leydesdorff, 2006). Yang analysed the relationship between a journal's value chain and competitive edge using value chain theory (Yang, 2006). As a whole, these ideas and methods for quantitatively measuring a journal's competitive relationship have not been generally accepted, and are not fully developed.

Applying research ideas from ecological competition

Recently, ideas related to competition and competitive intensity in ecology have been applied to research related to S&T journals. Scholars such as Tao, Daoping and Gaoming (2007) have attempted to consider the survival and development of S&T journals from an ecological perspective. Xinyan (2008) researched the concentration ratio of an S&T journal's market share and its competition. She also analysed the index model of competitive intensity in ecology, and applied it to measure a journal's competitive intensity (CI). This was a meaningful exploration, but did not result in a proper index for measuring a journal's distance in terms of the ecological system of S&T journals (Xinyan, 2008).

The competitive environment of S&T journals has been extensively analysed. Progress has been made in terms of the quantitative analysis. Although the CI concept from ecology is useful, we do not know how to define and measure the "distance" between journals. The institute of Scientific and Technical Information of China has measured journal similarity using the mutual citation matrix and cosine similarity method since 2011 (ISTIC, 2011). This provides a measurement of the distance between journals.

In this study, we considered a journal's absolute impact value and similarity as parameters based on the *Journal Citation Reports*. We measured the competitive pressures of mycology journals and the CI for the entire subject using scientometrics and the CI.

Methodology

In this study, we used the concept of CI from the field of ecological research to define the "competitive pressure" among S&T journals. The following design scheme illustrates how we calculate the relevant values.

Main factors that influence the competitive relationship between S&T journals

In a relatively closed ecological environment, the CI mainly depends on the differences between plant diameters and the distance between plants. In this closed environment, the competitive relationships between plants can indicate the strength of the overall competition within the ecological environment.

If we consider journals that focus on one subject, we are investigating a relatively closed ecological environment. Then, all the individual journals can be viewed as separate plants. As shown in Figure 1, the respective "diameters" (D_i and D_j) of journals i and j, and the "distance" (L_{ij}) between them are the major factors of the competitive relationship.

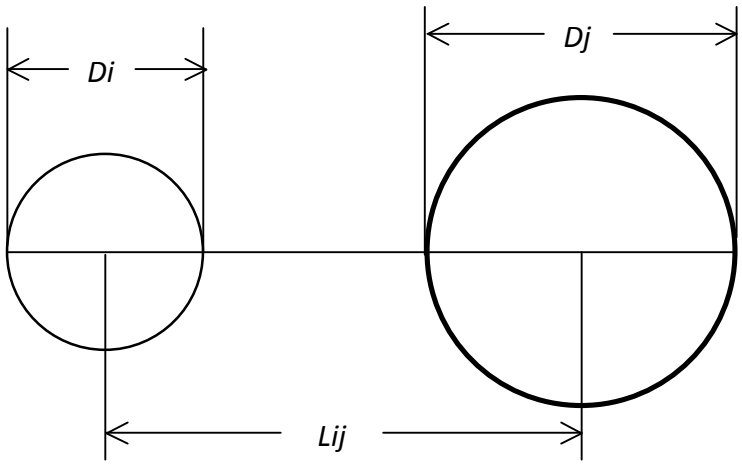


Figure 1. Main factors influencing the competitive relationship between S&T journals.

The number of total citations can be used as an alternative indicator to reflect the influence of the journal

The absolute influence of the journal can be seen as the plant thickness (diameter). Typically, a thicker plant is more capable of competing for resources and fighting rivals. Similarly, more influential journals are generally stronger in terms of their access to excellent manuscripts, funding, and attention. Journals with weaker influences are under more pressure from competitors.

The absolute influence of journals can be quantified using three main indicators: total citations (TC), IF, and the number of published papers.

Among these indicators, the IF is more likely to fluctuate. The number of papers is more vulnerable to subjective factors and can sometimes change dramatically. For example, a change to the journal's publishing cycle from bimonthly to monthly will lead to a sudden increase in the number of papers, and an accordingly sharp drop in the IF (because of a doubled denominator). Compared with the IF and paper number, the total citation indicator is relatively more stable and objective. It visually reflects the influence of journals, is less effected by other factors, and has a distinct advantage in terms of long term monitoring.

Additionally, the IF depends on the average number of citations of paper in a journal, so the total citation is equal to the IF multiplied by the number of papers. From this point of view, the total citation is monotonic in the mathematical sense.

Considering the above discussion, the total citation can be used as an alternative indicator of the influence of a journal. Therefore, in this study, we use the total citation as the diameter (D_i) of journal *i*. That is,

$$D_i = TC_i, \tag{1}$$

where TC_i is the total citation of journal *i.*

The similarity of two journals can be compared using the "distance" between them

It is widely accepted within the ecological community that competition is most intense when the same species live in the same environment (Clements, 1905). The similarity between two journals is also an important factor in their competitive relationship. In other words, a greater similarity between two journals leads to more intense competition. The similarity between two journals can be compared using the "distance" between them (L_{ij}).

Zheng, Na & Guozhen (2012) calculated a citation matrix for a sample of Chinese journals, which is classified into 61 subjects. They calculated the similarities for each journal in a specific subject area, and then constructed the similarity matrix for the journals. We used the same definition, and calculated the distance between periodicals using

$$\mathrm{Lij} = \frac{1}{Sij} - 1, \quad (2)$$

where S_{ij} is the cosine similarity indicator between i and j. S_{ij} is in the range of [0,1], and l_{ij} is in the range of [0,∞]. A S_{ij} value that is closer to 1 means that journals i and j are more similar. Accordingly, the distance L_{ij} is closer to zero. Conversely, if S_{ij} is closer to zero, i and j are less similar and the distance L_{ij} is closer to infinity.

Calculating the competition pressure between S&T journals

We used Hegyi's quantitative measurement for plant competition in ecology (Hegyi, 1974). Suppose that there are n journals for a subject, the target journal is called i and is set as the "basic journal", and the other is called j and considered a "rival journal". Then, $CRij$ is the competitive pressure on journal i from rival j. It is calculated using

$$\mathrm{CRij} = \frac{Dj}{Di \cdot Lij}. \quad (3)$$

We can assume that the competitive pressure on i from j is inversely proportional to the absolute influence of i, is directly proportional to the absolute influence of the rival, and is inversely proportional to the distance between the journals. This assumption is consistent with an intuitive understanding of the competitive relationship.

Combining Equations (1), (2), and (3), we get

$$\mathrm{CRij} = \frac{TCj}{TCi \cdot \left(\frac{1}{Sij} - 1\right)}, \quad (4)$$

where TC_i and TC_j represent the TC for i and j, and S_{ij} is the cosine similarity between periodicals.

CR_{ij} and CR_{ji} represent the competitive relationship between i and j. The cosine similarity S_{ij} measures the angular distance between a journal and its rival, so S_{ij} and S_{ji} are equal. However, CR_{ij} and CR_{ji} are not equal if TC_i is not equal to TC_j. Equation (4) implies that C_{ij} and C_{ji} have a mutually reciprocal relationship.

We can conclude from the definition that the basic journal is under less competitive pressure if it has a higher total citation value than its competitor, and vice versa. The more similar the journals are, the greater the competitive pressure. A journal does not compete with itself, so CR_{ii} is zero.

Calculating the competitive pressure on basic journal i

Suppose that, within its discipline, basic journal i has n-1 rival journals. Then, CI_i is the total competitive pressure on journal i from all of its rivals,

$$\mathrm{CIi} = \sum_{n}^{j=1} CRij. \quad (5)$$

Overall competitive strength for a specific subject

The number of competing journals depends on the subject classification. To compare disciplines, we define the overall competitive strength as CIS. It is the average competitive pressure for all journals, i.e.,

$$\mathrm{CIS} = \frac{1}{n}\sum_{n}^{i=1} CIi. \quad (6)$$

Analysis and Results

We calculated the mutual citation, similarity, competitive relationship, and competitive pressure matrices for the journals, and the CI for mycology using Journal Citation Reports (*JCR*) data from 1997, 2000, 2005, 2010, and 2003.

The inter-citation matrices for the target subject, and the similarity and competitive relationships

We used journals focussed on mycology to demonstrate how to calculate and analyse inter-citations within the target subject, and the similarities and competitive relationships between journals.

There are 23 journals indexed in the *JCR 2013* for mycology (*n=23*). The inter-citation matrix (*C*) was constructed by calculating the inter-citations of each pair of journals. We used the cosine similarity method to transform the inter-citation matrix to the similarity matrix, *R*. The cosine similarity is calculated using

$$\text{Cosine}(x,y) = \frac{\sum_{i=1}^{n} x_i\, y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}. \quad (7)$$

We transformed *R* into a net document and used Pajek to produce Figure 2, which shows the mutual citation network for mycology according to *JCR 2013*. Each node represents a journal, and a node's area represents the journal's TC. The location of the journal and the thickness of the link represent its similarity with its rivals.

From another perspective, we considered the whole subject area as an ecological space. Then, the 23 journals are independent plants. Figure 2 can be regarded as an ecological system with 23 plants, as viewed from above. The differences between the plant diameters and distances between plants determine the CI and the state of the journals.

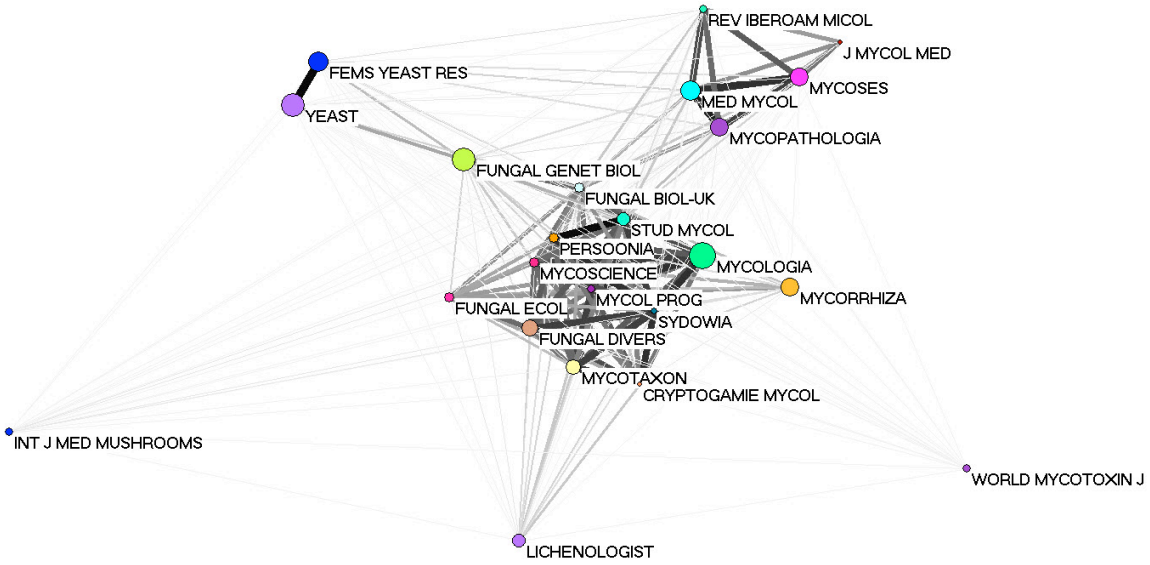


Figure 2. Mutual citation network of journal focussed on mycology, according to *JCR 2013.*

We applied Equation (4) to construct the competitive pressure matrix (*CR*) for the 23 journals, by considering each journal's TC and the cosine similarities between each journal pair.

Competitive pressure for a journal (CI)

Equation (5) shows that the CI of a journal is a combination of the competitive pressure from all of each rivals. We measured the competitive pressure of the all journals using competitive relationship matrices for mycology at five time points.

Table 1 shows that there were large differences in the competitive pressures of the rival journals. The maximum was 408.198 and the minimum was 0.022. In *JCR 2013,* two journals had competitive pressures over 100, 15 were between 10 and 100, and six were under 10.

Table 1. Competitive intensity (CI) for mycology journals.

Title	*1997*	*2000*	*2005*	*2010*	*2013*
CRYPTOGAMIE MYCOL	79.15	278.326	37.227	90.551	140.329
EXP MYCOL	13.81				
FEMS YEAST RES				17.673	32.585
FUNGAL BIOL-UK				81.125	48.575
FUNGAL DIVERS			28.170	8.875	14.402
FUNGAL ECOL				16.954	23.032
FUNGAL GENET BIOL	4.394	14.820	2.985	1.929	3.222
INT J MED MUSHROOMS				0.341	2.175
J MED VET MYCOL	13.572				
J MYCOL MED	42.521	18.324	31.853	17.819	41.412
LICHENOLOGIST			3.753	3.057	3.249
MED MYCOL		28.391	5.748	7.315	18.067
MIKOL FITOPATOL	3.280	1.854	2.389		
MYCOL PROG				189.149	98.921
MYCOL RES	3.751	6.649	11.217	11.919	
MYCOLOGIA	4.663	7.341	12.558	5.09	6.046
MYCOPATHOLOGIA	11.130	4.616	5.069	6.109	17.724
MYCORRHIZA	4.993	8.529	4.174	2.036	2.292
MYCOSCIENCE				30.886	53.764
MYCOSES	10.392	3.991	3.422	12.211	18.333
MYCOTAXON	16.890	20.216	18.220	15.182	16.865
PERSOONIA	94.223	84.520	408.198		92.237
REV IBEROAM MICOL				31.666	35.185
STUD MYCOL	139.528	69.935	51.901	31.591	36.342
SYDOWIA			116.148	298.986	230.812
WORLD MYCOTOXIN J					0.095
YEAST	0.031	0.022	0.318	5.028	15.638

Table 2 shows the competitive intensities compared with the IF and TC, for mycology journals in 2013. The rankings based on the IF and TC is different from the CI rankings. Some journals are ranked in the top 10 in terms of TC and IF but have low CIs, and some are ranked in the bottom five in terms of TC and IF but have higher CI*s*. Therefore, a more powerful journal has lower competitive pressure. We have only listed the results based on the 2013 data, but they were similar for 1997, 2000, 2005, and 2010. The difference between a journals' absolute and relative influence is related to its competitive pressure.

There are certainly some exceptions. Journals that are extremely similar have a significant influence on the competitive pressure. For example, some journals have TCs that are greater than one thousand and are very similar to other journals with the same mass influence, so they also have high competitive pressures. However, some journals are focused on narrow fields and have distinctive characteristics, and therefore do not have much competition because there are not many similar journals, although their TC may be high.

Table 2. Competitive intensity (CI) compared with impact factor (IF) and total citations (TC), for mycology journals in 2013.

Title	*CI 2013*	*rank*	*IF 2013*	*rank*	*TC 2013*	*rank*
CRYPTOGAMIE MYCOL	140.329	2	1.153	18	254	22
FEMS YEAST RES	32.585	10	2.436	7	2935	5
FUNGAL BIOL-UK	48.575	6	2.139	10	790	14
FUNGAL DIVERS	14.402	17	6.938	2	2120	9
FUNGAL ECOL	23.032	11	2.992	5	701	15
FUNGAL GENET BIOL	3.222	20	3.262	4	4298	2
INT J MED MUSHROOMS	2.175	22	1.123	19	554	19
J MYCOL MED	41.412	7	0.4	22	247	23
LICHENOLOGIST	3.249	19	1.613	14	1285	12
MED MYCOL	18.067	13	2.261	9	3132	4
MYCOL PROG	98.921	3	1.543	16	623	18
MYCOLOGIA	6.046	18	2.128	11	5754	1
MYCOPATHOLOGIA	17.724	14	1.545	15	2913	6
MYCORRHIZA	2.292	21	2.985	6	2650	7
MYCOSCIENCE	53.764	5	1.288	17	926	13
MYCOSES	18.333	12	1.805	12	2451	8
MYCOTAXON	16.865	15	0.643	21	1959	10
PERSOONIA	92.237	4	4.225	3	669	16
REV IBEROAM MICOL	35.185	9	0.971	20	649	17
STUD MYCOL	36.342	8	9.296	1	1461	11
SYDOWIA	230.812	1	0.213	23	355	21
WORLD MYCOTOXIN J	0.095	23	2.38	8	454	20
YEAST	15.638	16	1.742	13	4268	3

Figure 3 shows the difference between the CI rankings for a set of journals between 1997 and 2000, and a second set of journals between 2005 and 2013. For the first set, the CI rankings for most of the 14 journals decreased from 1997 to 2013, and only four were in the top ten. This typically means that the competitive pressures of traditional journals (with a longer publishing history) were declining. At the same time, most of the second set started in a high competitive pressure situation, and approximately half of them remained in the top ten of the CI ranking. This means these new journals had to face more challenges.

Competitive intensity for a subject

Equation (6) shows that the CI for a subject is the average competitive pressure of all the journals. We calculated the CIs for mycology in 1997, 2000, 2005, 2010, and 2013.

Table 3 shows that the competitive intensity for a subject (CIS) increased from 1997 to 2005, but the number of journals only increased from 15 to 17. We can see that the CIS decreased between 2005 and 2010 because the number of journals increased from 17 to 23 (by approximately 35%). By analysing the relationship between the subject's scale and CIS, we can see that more journals correspond to low CIs. From 2010 to 2013, the number of journals was stable at 23 so the CIS increased. In the absence of any other influences, the CIS will continue to increase.

By analysing the competitive pressure on each journal and the CIS, we can determine the state of the competitive environment using a quantitative method, and compare the competitive

relationships of different journals and subjects. Through a comparative analysis, we can research reasons for any differences and provide S&T publications with scientific data and tools. Additionally, the data can be used to monitor the S&T journals environment at a macro level, and help decision makers with regard to administration.

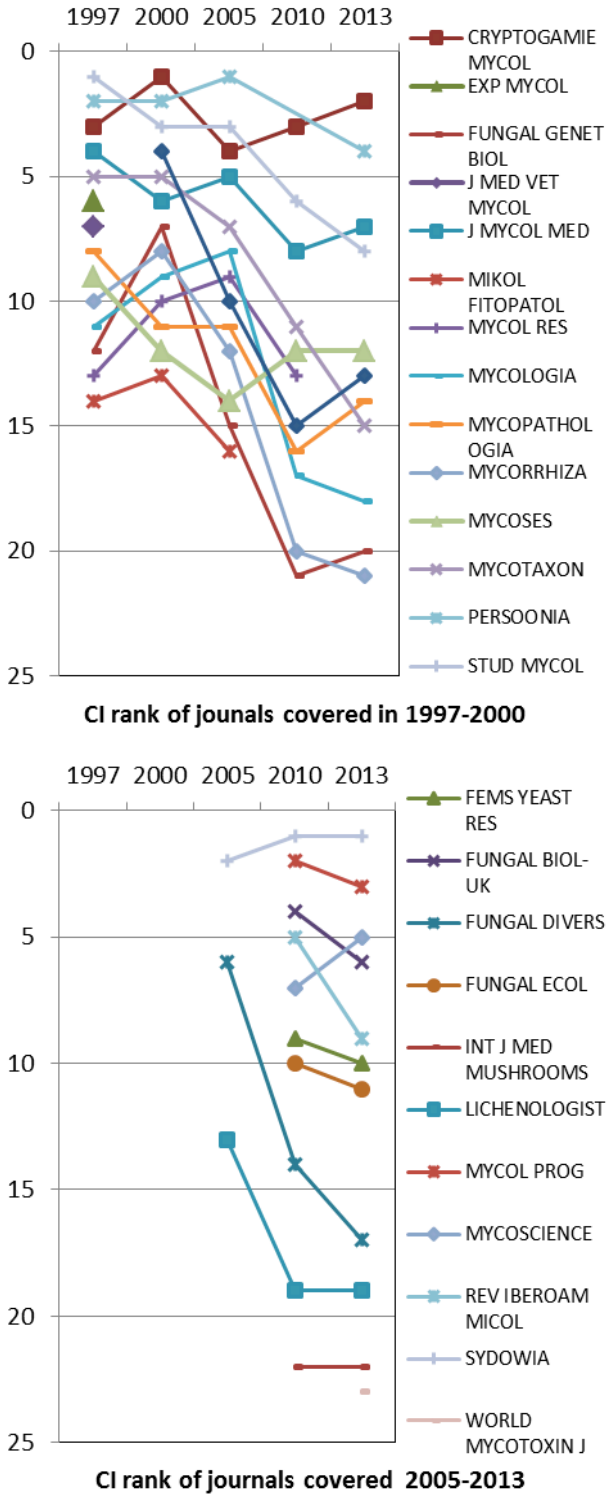


Figure 3. Relationship between competitive intensity (CI) and time.

Table 3. Competition intensity (CI) and number of journals for mycology

	1997	*2000*	*2005*	*2010*	*2013*
number of journal	15	14	17	23	23
CIS	29.489	39.110	43.726	38.500	41.361

Conclusions

There is vast difference in the CIs between subjects and competition pressures between journals.

We have measured journals' competition pressures and the CIS using quantitative methods. The differences between journals' competitive environments may be caused by many related factors. Different journal attributes are related to competitive pressure. For example, the competitive environment and resources vary among multidisciplinary, ordinary professional, and specialized professional journals. Fundamental research or academic journals and engineering or application journals have different competitive features. Chinese journals are obviously different to English language journals. So the factors that influence competitive pressure and intensity, measurements of these related factors, and mechanisms that influence journals' competitive environments must be studied further.

The competitive pressure from a powerful rival may be equal to the pressure from several weakly similar journals.

The ecological concept of CI is a combination of all kinds of competitive pressure. So the competitive pressure on a journal is a combination of the competitive pressure from all of its rivals. The competitive pressure from a powerful rival may be equal to the pressure from several weakly similar journals. The combination of competitive pressure for each journal may be different, which can lead to a high competitive pressure and number of rivals. It can be used as reference when analysing a target journal's competition.

A journal's homogeneity is important when developing S&T journals. Using our quantitative method, we found that homogeneity is obvious in some fields, especially journals that lack "personality". Such journals have higher competitive pressures. The homogeneity of a journal increases its competitive pressure, and the homogeneity of a subject hinders a favourable competitive environment. There is typically fierce competition between two journals that are very similar. Abnormal cooperative relationships exist between some journals, who adopt inter-citation journal group models. These very similar journals pursue high IFs and cited rates. The academic misconduct phenomenon is one problem that results from a journal's homogeneity.

More study is required for multidisciplinary or interdisciplinary journals.

In our method, each journal only belongs to one subject. However, developments in science and technology have led to fusions and evolutions in subject areas. Most articles belong to more than one subject area. At the same time, some journals are multidisciplinary, so it can be difficult to define their subject. We measured a journal's competitive pressure in terms of only one subject. Future research is required to determine how to measure and compare competitive pressure and similarities for multidisciplinary or interdisciplinary subjects.

A favourable competitive environment is only possible at the proper scale

The scale of the subject (number of journals) is related to its competitive pressure and intensity. A favourable competitive environment is only possible at the proper scale. If there

are too many or too few journals the CI decreases. In S&T journal administration, the distribution and trends of the CIs can be used as a reference to promote the development of favourable and sustainable environments.

The research findings in this study can be used as a reference for a new journal when choosing a subject and field.

In management science, there are "red ocean" and "blue ocean" strategies when facing competitive environments. The red ocean strategy directly reacts to competition, whereas the blue ocean strategy avoids direct competition and exploits new markets (Chan & Mauborgne, 2005). When facing competition from rivals, S&T journals must choose an optimal path based on the current environment and future positioning. Journals with relative advantages tend to use red ocean strategies, proactively consolidating and extending their advantages. Relatively weak journals use blue ocean strategies, seeking paths that reduce homogeneity problems and competitive pressures. The findings of this study can be used as a reference for a new journal when choosing a subject and field. In a fiercely competitive fields, it is difficult to successfully launch a new journal without obvious diversity.

Acknowledgments

This research was supported by National Key Technology Support Program of China. (Project Number: 2015BAH25F01)

References

Bonitz, M. & Scharnhorst, A. (2001). Competition in Science and the Matthew Core Journals. *Scientometrics*, 51, 37-54.

Cai, Y. (2003). On Market Competitiveness of Sci-tech Journals. *Chinese Journal of Scientific and Technical Periodicals*. 14, 345-348

Campanario, J.M. (1996). The Competition for Journal Space among Referees, Editors, and Other Authors and Its Influence on Journals' Impact Factors. *Journal of the American Society for Information Science*, 47,184-192

Clements, F.E. (1905). *Research Methods in Ecology*. Lincoln Nebraska: University Publishing Company.

Chen, X. (2005). On sets of core competitiveness of scientific and technical journal. *Media*. 2005, 9, 55

Gao, J. (2004). Market competitive game of sci-tech periodicals. *Acta Editologica*, 16, 319-320.

Hegyi, F. (1974). A simulation model for managing jack-pine stands. In *Growth models for tree and stand simulation*. J. Fries (Ed.). Stockholm: Royal College of Forestry.

Institute of Scientific and Technical Informaiton of China. (2011). *Chinese S&T Journal Citation Reports 2011*. Beijing: Scientific and Technical Documention Press.

Kim, W.C. & Mauborgne, R. (2005). *Blue Ocean Strategy*. Beijing: The Commercial Press.

Leydesdorff, L., Wagner, C.S., & Bornmann, L. (2014). The European Union, China, and the United States in the top-1% and top-10% layers of most-frequently cited publications: Competition and collaborations. *Journal of Informetrics*, 8, 606-617

Leydesdorff, L. (2006). Can Scientific Journals be Classified in terms of Aggregated Journal-Journal Citation Relations using the Journal Citation Reports? *Journal of the American Society for Information Science and Technology*, 57, 601-613.

Leydesdorff, L. (2004). Clusters and Maps of Science Journals Based on Bi-connected Graphs in the Journal Citation Reports. *Journal of Documentation*, 60, 317-427.

Tao, Y., Daoping, W. & Gaoming. Z. (2007). Think deeply in ecology of sci-tech periodicals' survival and development. *Acta Editologica.*, 19, 3-5.

Xinyan, L. (2008). Study on Competition Intensity in Scientific and Technical Journals Publishing Industry. (Unpublished Master's Dissertation) Institute of Scientific and Technical Information of China.

Zheng, M., Na, W. & Guozhen, Z. (2012). The Analysis of Mutual Citation Network on Patterns of Chinese S&T Core Journal Groups. *Studies in Science of Science*. 2012, 30, 983-991.

Zhu, J., & Mei, H. (1999). Relation between competitiveness and quality for S&T journal. *Science Technology and Publication*, 5, 27-29

SciELO Citation Index and *Web of Science*: Distinctions in the Visibility of Regional Science

Diana Lucio-Arias[1], Gabriel Velez-Cuartas[2] and Loet Leydesdorff [3]

[1] *dlucioarias@gmail.com*
Colombian Observatory of Science and Technology, Bogota (Colombia)

[2] *gabrielvelezcuartas@gmail.com*
Grupo de Investigación Redes y Actores Sociales, Departamento de Sociología, Facultad de Ciencias Sociales y Humanas, Universidad de Antioquia; Calle 70 No. 52-21, Medellín (Colombia)

[3] *loet@leydesdorff.net*
University of Amsterdam, Amsterdam School of Communication Research, Amsterdam (The Netherlands)

Abstract
In this study we compare the visibility and performance of Latin American and Caribbean (LAC) Science in terms of its presence in the core collection indexes included in the *Web of Science* (WoS) —*Science Citation Index Expanded*, *Social Sciences Citation Index*, and *Arts & Humanities Citation Index*—and the *Scielo Citation Index* (SciELO CI)—which was recently integrated into the WoS platform. The purpose of this comparison is to provide some inputs to reconstruct the role of SciELO as a communication platform for science produced in Latin America and the Caribbean, and to provide some reflections on the potential impacts—in terms of a better understanding of the global scientific scenery—of the articulation of SciELO CI into WoS: Are there significant differences in the region´s scientific results when studied from publications included in SciELO CI versus those included in the traditional core collection of the WoS? Are regional exercises, such as SciELO, successful in enhancing the visibility of regional scientific production?

Conference Topic

Journals, databases and electronic publications

Introduction

Although the participation of Latin American and Caribbean (LAC)-edited journals in WoS has increased over time, this growth is not comparable to the growth in the participation of scientific articles with at least one author affiliated to an institution in LAC. This increase in participation has been interpreted as a successful integration of LAC science into the world repertoires despite a persistent and notorious gap in the making of good scientific journals (Meneghini, Mugnaini & Packer, 2006). The difference in the nature and characteristics of the journals considered and included in each of the indices justifies our expectation of finding significant differences in the science produced in LAC and communicated through WoS or SciELO CI indexed journals: while the inclusion policy of WoS targets the top quality journals by discipline, the program SciELO has had an inclusive policy aimed at increasing visibility and circulation of LAC journals and their content.[1]

[1] SciELO (Scientific Library on Line) was a program that was initiated in Brazil in 1997 with the purpose of offering a core of Brazilian scientific journals in an open access mode through internet. The program had a successful expansion in the region and now includes, in addition to Brazilian, journals from Chile, Cuba, Spain, Venezuela, Colombia, Argentina, Costa Rica, Mexico, Portugal, Peru, and Uruguay. It is important to note that the SciELO program transcends the SciELO citation index which is the subject of this study. Not all the scientific journals that belong to the SciELO collection and whose content has been made available through SciELO´s program belong to ScieLO´s citation index.

Another difference in the origins of SciELO and WoS that might be helpful in explaining the differences in regional scientific communication is related to the disciplinary context of each of the indexes. A lot has been written about the "natural" or hard sciences origin of WoS, which derived from the Science Citation Index (Garfielfd, 1971), but was expanded to include a broader range of journals and then accompanied by the Social Science Citation Index and later on by the Arts & Humanities Citation Index. The three indexes have been operative since 1978. SciELO, on the other hand, resulted from cooperation of the Fundacao de Amparo a Pesquisa do Estado do Sao Paulo (FASPEP) and the Latin American and Caribbean Center for Health Sciences Information (Bireme) of Panamerican and World Health Organization (PHO/WHO).

We believe that SciELO´s contribution to global science relies on its impact in the circulation of LAC scientific production and therefore the visibility of this production. In the last 15 years, SciELO played an important role in the development of capabilities in LAC to produce world-class scientific results, particularly though the consolidation of a regional base of high-quality scientific journals. The financial requirements to maintain such an exercise updated, expanding and relevant (Aguillo, 2014), together with the potential of SciELO indexed journals to provide a representation of LAC science, might explain the interest behind the inclusion of the regional exercise in the Thomson Reuters owned databases.

The inclusion of SciELO into WoS has had a mixed reception in the LAC scientific community. In 2007, an alliance between Scopus and SciELO raised expectations of all SciELO information to be included in Scopus (Elsevier, 2007). The potential impacts of the inclusion of the journals, and the ambiguity of whether all SciELO journals would be included in Scopus raised some concerns in the LAC scientific community. The negotiations behind SciELO´s inclusion either in Scopus or WoS, was perceived by some editors of LAC journals as a "sell-out" of SciELO´s principles and allowed uncertainty in the future of the regional journal structure that SciELO had aimed to consolidate.

With this paper we expect to contribute on the relevance of both indexes and the complementarities between them as they represent different styles of scientific communication that transcend the center-periphery debate on scientific production. This section is followed by a section in which we introduce the data and methods employed for this study. The results section will focus on the differences between the indices; specifically in the geographical, collaborative aspects, and cognitive characteristics of the communications in each. We finish this contribution with some reflections on the challenges and opportunities of the integration of SciELO into WoS.

Data and Methods

We downloaded all the bibliographical information from the core collection of the WoS (SCI expanded, SSCI, A&HCI) for 79,924 documents that responded to the search query for affiliation country to any LAC countries AND publication year 2012. The same information was downloaded for 30,518 documents that responded to the same search query in the SciELO CI available through WoS. While participation of LAC authors explains 73% of the total publications in SciELO CI, in WoS, this participation is lower than 5%.[2] The organization of the information into relational databases was possible through dedicated routines available at http://www.leydesdorff.net/scielo and http://www.leydesdorff.net/software/isi/index.htm.

[2]In January 2015, a total of 1,899,805 documents were included in WoS with publication year 2012, and 41,621 in SciELO CI.

In order to assess some of the differences in the sets of data considered in this analysis, we provide some descriptive statistics in Table 1. We include the mean and the standard deviation to provide some order of magnitude and dispersion among attributes.

From Table 1, differences among the types of communications included in each set are evident. The mean (μ), represents the average number of authors, addresses, citations, cited references and subject categories per document and the standard deviation (σ) is included to illustrate dispersion in these data. The documents in journals indexed in WoS have more citations, and more frequently result from collaborations among larger number of authors in European or American institutions. These documents are more codified (in terms of the cited references used) as well, and, in general, have a significantly larger impact (in terms of citations received). The mean and standard deviation of the journals are included to represent the average number of LAC documents per journal. Although fewer journals concentrate LAC scientific production in SciELO CI than that in WoS, dispersion among different titles is greater; as can be expected, SciELO CI indexed journals have a larger participation of LAC authors compared with authors from other countries. A total of 163 journals are indexed in both WoS and SciELO CI.

Table 1. Differences in the sets of LAC publications from SciELO CI and WoS Core collection.

LAC publications	**SciELO CI**			**WoS Core Collection**		
Records	30,518			79,924		
Statistics	**N**	**μ**	**σ**	**N**	**μ**	**σ**
Authors	91,269	3.8	2.4	306,560	14	144,3
Addresses	11,858	2.3	1.5	168,390	3.9	14.3
Times cited	7,733	0.3	0.7	274,225	3.4	18.6
Cited references	681,151	26.2	19.1	1,969,653	37	29
Subject Categories	186	1.2	0.7	246	1.5	0.8
Journals	750	40.7	44.5	7,268	10.9	28.0

We use the Overlay maps Toolkit available at http://www.leydesdorff.net/overlaytoolkit (Rafols, Porter & Leydesdorff, 2012) to provide the different visualizations of the relations among disciplines in each of the document sets (SciELO CI and WoS core collection). We rely on these visualizations to suggest disciplinary differences in each of the sets of documents. We expect some of these differences to reflect on diverse goals and interests in the management of each of the indices and which were shortly introduced above.

To reflect upon the distinctions in the collaborative nature of the communications in each index, we build a collaboration network between countries using Pajek.

Results

In this section we provide some results on the differences between communications in the Core Collection of WoS and the recently integrated SciELO CI, focusing on the regional, collaborative and cognitive aspects underlying these communications. In Table 2, we provide the number of records in each of the sets by country of origin of the authors. To normalize for documents with a high number of co-authorships we include a fractional counting of documents considering the total number of signing authors.

The divergence in the countries' participation in the scientific production of LAC can result from (a) the degree in which the specific country has become articulated in the SciELO program and the efforts in increasing the SciELO journal list of each country. As can be expected, the most important SciELO journal collection is from Brazil and it includes 337 journal titles, Colombia follows with a total of 184 journal titles, Mexico has 149, Argentina

and Chile 107 and 106 journal titles each. Another explanation is (b) the specific country´s treatment and importance of national scientific journals.

The policy effort supporting national scientific journals varies in the region where some countries privilege international publication while others aim at balancing international visibility with support to local journals and local publishers (Vessuri, Guédon & Cetto, 2014). Different publication strategies are also evident from Table 2 where the effect of fractional counting seems to be more drastic for communications in journals indexed in WoS Core collection than in SciELO CI. Colombia, for example, has relied on collaborating with international peers to increase their participation in international journals and databases (Lucio-Arias, 2013).

Table 2. Regional distribution of papers in WoS Core collection and SciELO CI.

Country	SciELO CI		WoS	
	Records	Fractional	Records	Fractional
Brazil	19,537	11,929.5	44,812	21,844.1
Colombia	3,065	2,312.2	4,007	1,734.9
Chile	2,409	1,754.3	7,277	3,562.0
Mexico	2,336	1,529.2	13,041	5,879.3
Cuba	1,979	1,053.5	966	320.8
Argentina	1,625	1,223.8	9,975	4,953.8
Venezuela	526	340.8	1,240	543.9
Peru	480	344.0	975	336.1
Costa Rica	284	189.4	514	310.8
Uruguay	99	51.8	868	195.3
Ecuador	53	25.0	465	153.4
Bolivia	42	20.0	85	17.0
Guatemala	23	11.4	52	8.0
Panama	22	8.0	416	120.7
Puerto Rico	22	8.0	N/A	N/A
Paraguay	27	10.7	43	6.1
El Salvador	11	5.1	24	3.1
Jamaica	10	3.1	9	1.8
Nicaragua	20	8.4	31	4.3
Honduras	3	1.0	25	2.8
Dominica	1	0.2	2	0.4
Dominican Republic	1	0.2	33	4.4

The alliances and collaborations reflect important differences in the networks of collaboration that emerge from LAC scientific communications in each of the indices considered (See Figures 1 and 2).

Collaborations in WoS suggest the importance of North America and Europe as allies in the production of scientific knowledge in the region. Collaboration of LAC countries with peers "from the north" dominates scientific communications where LAC participate. Regional collaboration seems not very relevant and in fact not as important as collaboration with Asia, Africa and Oceania. South-South collaboration has received a lot of attention (Arunachalam & Doss, 2000; Chandiwana & Ornbjerg, 2003) and has become an important issue in the

development policy agenda.[3] We believe, nevertheless, that South-South collaboration depicted in Figure 1 is mostly mediated by developed countries and does not represent necessarily a transfer and exchange of resources and knowledge.

The resulting map of collaborations in LAC scientific communications in journals indexed in SciELO CI, suggest a more pronounced strategy based on the regional conjugation of research efforts. Collaboration with Europe is mainly oriented towards Spain and Portugal, suggesting language and cultural similarities as a strong motivation to collaborate. Collaboration with North America and particularly with the United States might rely on geographic proximity as this is stronger in the case of Mexico.

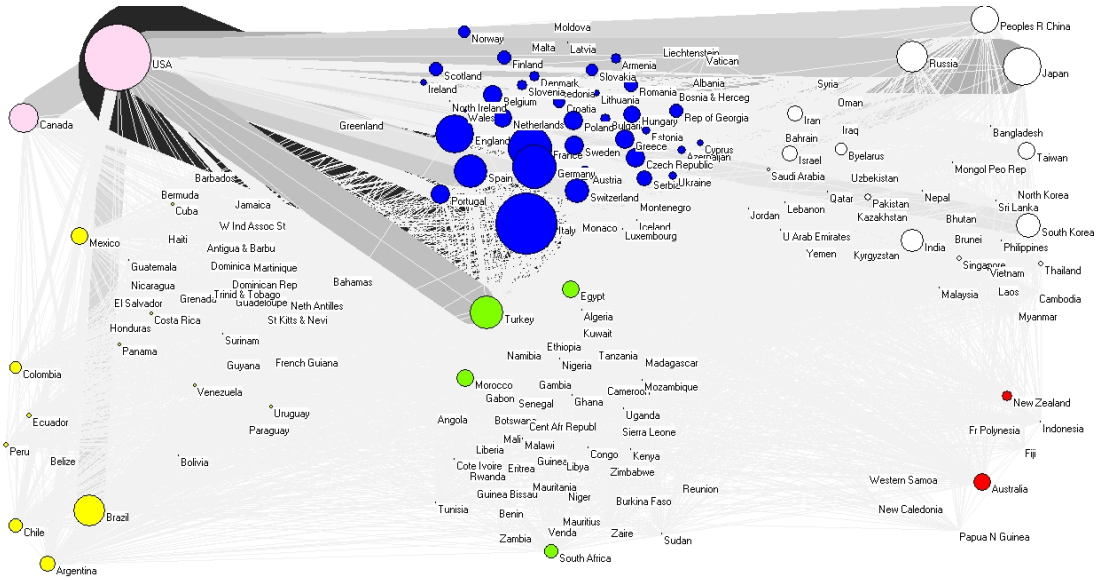


Figure 1. International Collaboration from LAC communications in WoS Core Collection.

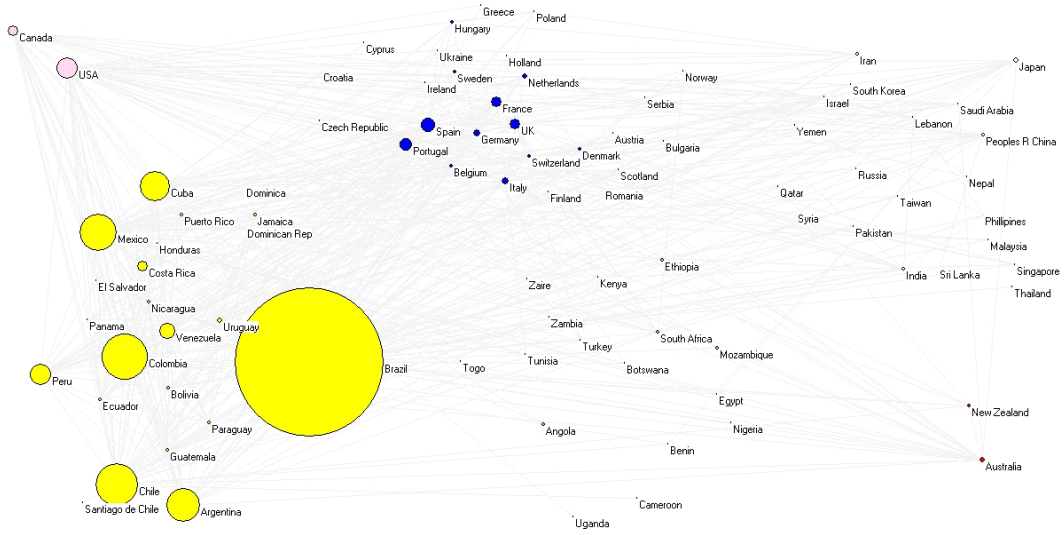


Figure 2. International Collaboration from LAC communications in SciELO CI.

Although it deserves further research, we expect collaborations in SciELO to be a better representation of South-South cooperation, which implies an exchange of resources and ideas within developing countries to solve similar development problems. Collaboration in Figure 2

[3] There is a United Nations Office for South-South cooperation with a website at http://ssc.undp.org/content/ssc.html.

within LAC, Africa and Asia might be a better representation of South-South cooperation. We expect less mediation of the North in the South-South collaboration for the case of SciELO CI indexed communications.

In summary, the differences between Figures 1 and 2 suggest distinct communication practices when (a) aiming at results with international visibility than when the main goal is (b) regional or local diffusion of scientific results through regional journals. While for WoS (Figure 1) strong ties can be indicated with North America and Europe, regional collaboration seems dominant in Figure 2. The participation of the USA in Figure 1 and Brazil in Figure 2 should be interpreted considering that these countries have the highest numbers of indexed journals in each of the respective databases.

This can also result from the different disciplines represented in each index. While WoS has some dominance of "hard" sciences, which are more prone to be published in English and in collaboration, for SciELO CI the disciplinary participation seems to favor the social sciences (see Figure 3 and 4).

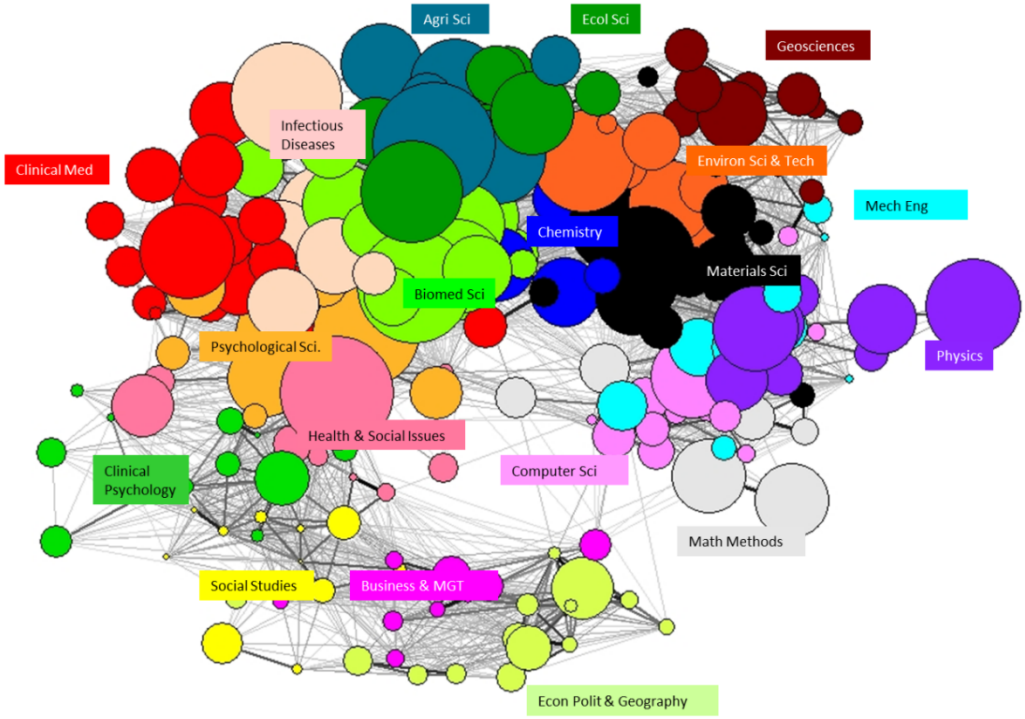


Figure 3. LAC map of Science, WoS Core Collection; 224 Web of Science Categories.

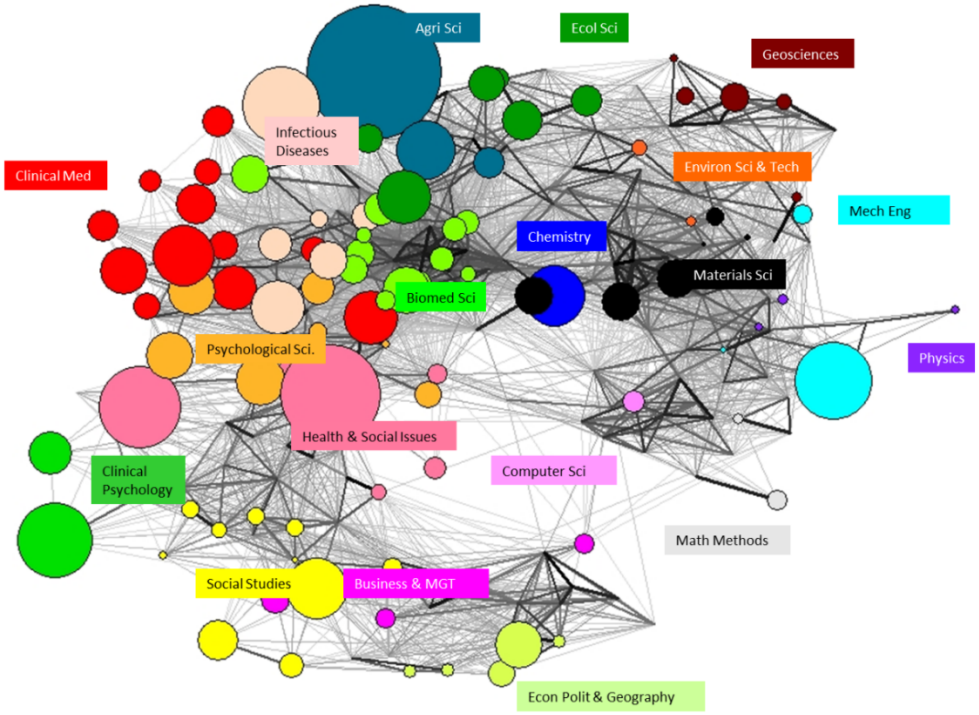


Figure 4. LAC map of Science, SciELO CI.; 224 Web of Science Categories.

Figures 3 and 4 suggest differences in the thematic orientation of the communications in each index. Contributions from the natural sciences are better represented in WoS Core Collection; nevertheless, SciELO CI provides a valuable insight into the regional scientific production in the social and health sciences (where social aspects of the health and medical sciences like research in public health has a better representation), and agriculture. Our expectation is that in-depth analysis of the subjects addressed by the communications would exhibit differences in the sets; communications in SciELO CI will address topics of regional relevance.

Reflections and Further Work

In the last twenty years, scientific development together with technological change and productive innovation have raised interest in the LAC countries, and as a consequence been targeted on the public-policy agenda. Important aspects in the institutionalization of scientific research, such as the consolidation of public institutions for the promotion of science technology and innovation, strengthening of public research institutes, the growth of PhD programs, and the formation and formalization of a journal structure, to socialize scientific results obtained in the region, have also characterized these last decades.

Although growth in the participation of LAC scientific production in traditional databases, such as Web of Science and Scopus, has also been the norm in this period, a common concern in the community has been the challenges to properly socialize scientific results when they are of little interest for mainstream scientific journals. The perseverance in LAC scientific communications of Spanish and Portuguese, as the main languages for communication, particularly in sciences with an important social component, demands alternative means of communication outside international journals as they might have their own structures. Leydesdorff and Bornmann (in press), for example, found a specific citation pattern of Spanish and Portuguese journals in library and information sciences (LIS).

This demand has been acknowledged and as a consequence, most LAC countries have an important structure of national journals. This poses other types of challenges in terms of

research assessment and evaluation. While rankings of international journals and measures based on citations allow researchers and librarians to make informed decisions on the expected quality of a scientific journal´s content, this distinction is more difficult and in occasions impossible when considering national publications. The proliferation of local journals edited by faculties or departments for the diffusion of mainly their own researchers' findings makes the distinction among journals harder.

The need to assess and monitor research results comes together with the demand for a transparent classification among scientific communications. How to assess scientific communications included in international journals versus regional or national journals? In part as a response to this need, different LAC countries have joined the SciELO program. SciELO, in our perspective, has had a positive impact on the consolidation of regional research capabilities and in providing a proper infrastructure for regional exchange and communication.

As was suggested in the collaboration networks analyzed, the SciELO program seems to have transcended the LAC region and includes authorships from Africa and Asia suggesting a platform for South-South collaboration. Other causes for the dominance of the international collaborations in scientific communications in WoS are the cognitive dominance of the biomedical and natural sciences, where collaboration among geographical dispersed groups of individuals is very common. The type of research that results in publications indexed in WoS Core Collection might also cause the dominance of international collaboration in WoS when compared to SciELO CI. Researchers from LAC countries might have a marginal participation in these collaboration networks. This position results of a collaboration among many authors and contributions in the form of data processing instead of cognitive contributions and argumentations. Successful collaborations in the region should hold the researchers in leadership positions (Moya Anegón et al., 2013).

From a cognitive perspective, the inclusion of SciELO CI into WoS offers new opportunities of coverage of disciplines and specialties where the particularities of the territory and the social context are important. Public health, social sciences and agriculture are relevant in SciELO CI; the participation of the LAC scientific communications in these disciplines in the core collection of the WoS has traditionally been low. In this sense, the 15% overlap of Scielo CI journals in both indexes suggests that the inclusion of SciELO CI in the WoS benefits WoS in terms of coverage of regional scientific advances, particularly of communications that have a local object of study and where communication is more original and responds to regional capabilities, but also regional issues and problems.

The inclusion of SciELO CI has raised some concerns among the editors of Spanish[4] and Portuguese journals that have benefitted from a special treatment and inclusion in WoS but that do not have an important position in SciELO CI. Editors of these journals fear that the policy of articulation of SciELO CI into the WoS might result in exclusion of their journals from WoS.

Inclusion of SciELO CI into WoS, responds to the need for a more inclusive representation of scientific results despite regional constrains and conditions. This has resulted from the competition of services offered by Thomson Reuters and Elsevier. The strategies aimed at improving regional visibility are different in Scopus and in the Web of Science. While Scopus has aimed at increasing coverage by increasing their base of regional journals, the globalization of the Web of Science (Testa, 2011) has meant the articulation of regional exercises. The Chinese Journal Database has been hosted in the WoS since 2008, the

[4] FECyT (Spain´s foundation for science and technology) has had an important role in certifying quality of its quality journals in order to support their inclusion in the WoS after an alliance with Thomson Reuters around 2007 (FECyT, 2011)

inclusion of SciELO CI and the Korean Journal Database has been operative since 2014. We believe that the strategy followed by Thomson Reuters provides the cumulative expertise of circulation and visibility promoted regionally, by programs similar to SciELO. We would like to explore this issue further in the future to understand how the inclusion of SciELO CI might put the WoS back in the competition for visibility of regional results.

References

Aguillo, I. (2014). Políticas de información y publicación científica. *El Profesional de la Información*, 23 (2), 113-118.

Arunachalam, S. & Doss, M.J. (2000). Mapping international collaboration in science in Asia through co-authorship analysis. *Current Science*, 79 (5), 621-628

Chandiwana, S. & Ornbjerg, N. (2003). Review of North South and South South cooperation and conditions necessary to sustain research capability in developing countries. *Journal of Health Population and Nutrition*, 21(3), 288-97.

Elsevier. (2007). Elsevier News America Latina. Retrieved on January 10, 2015 from: http://www.elsevier.com.br/bibliotecadigital/news_dez07/pdf/edicao_03_esp_ok.pdf

FECYT. (2011). Análisis de la presencia de las revistas científicas españolas en el JCR de 2010. Retrieved on January 10, 2015from: http://icono.fecyt.es/informesypublicaciones/Documents/2011_07_27RevEspanolasJCR2010.pdf

Garfield, E. (1971). The mystery of the transposed journal lists—wherein Bradford's Law of Scattering is generalized according to Garfield's Law of Concentration. *Current Contents*, 3(33), 5–6.

Lucio-Arias, D. (2013). Colaboraciones en Colombia, un análisis de las coautorías en el Web of Science 2001-2010. In, J. Lucio (Ed.). Observando el sistema nacional de ciencia y tecnología, sus actores y sus productos. Bogotá: OCyT.

Leydesdorff, L., & Bornmann, L. (in press). The Operationalization of "Fields" as WoS Subject Categories (WCs) in Evaluative Bibliometrics: The cases of "Library and Information Science" and "Science & Technology Studies". *Journal of the Association for Information Science and Technology*. http://arxiv.org/abs/1407.7849.

Meneghini, R., Mugnaini, R. & Packer, A.L. (2006). International versus national oriented Brazilian scientific journals. A scientometric analysis based on SciELO and JCR-ISI databases. *Scientometrics*, 69(3), 529-538.

Rafols, I., Porter, A. L. & Leydesdorff, L. (2010). Overlay science maps: a new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, 61(9), 871–1887.

Vessuri, H., Guédon, J.C., & Cetto, A.M. (2014). Excellence or quality? Impact of the current competition regime on science and scientific publishing in Latin America and its implications for development. *Current Sociology*, 62(5), 647-665.

Testa, J. (2011). The globalization of the Web of Science. http://wokinfo.com/media/pdf/globalwos-essay.pdf.

Book Bibliometrics – A New Perspective and Challenge in Indicator Building Based on the Book Citation Index

Pei-Shan Chi[1], Wouter Jeuris[1], Bart Thijs[1] and Wolfgang Glänzel[1,2]

peishan.chi@kuleuven.be, wouter.jeuris@kuleuven.be, bart.thijs@kuleuven.be, wolfgang.glanzel@kuleuven.be
[1]KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
[2]Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

Abstract
This study aims to gain a better understanding of communication patterns in different publication types and the applicability of the Book Citation Index (BKCI) for building indicators for use in both informetrics studies and research evaluation. The authors investigate the differences not only in citation impact between journal and book literature, but also in citation patterns between edited books and their monographic authored counterparts. The complete 2005 volume of the Web of Science Core collection database including the three journal databases and the BKCI has been processed as source documents. Annual cumulative citation rates in a three-year (x3) and a nine-year (x9) citation window are applied to compute the citation impact of different types of publications. The ratio x3/x9 is utilized as a kind of prospective Price index to examine the extent of ageing. The results of this study show that books are more heterogeneous information sources and addressed to more heterogeneous target groups than journals. Comparatively, the differences between edited and authored books in terme:s of the citation impact are not so impressive as books vs. journals. Humanities have the most different citation impact between books and journals, whereas life sciences have the most similar impact between two groups.

Conference Topic
Journals, databases and electronic publications; Citation and co-citation analysis

Introduction

Some consequences of the absence of books in bibliometric analyses

In contrast to the natural and life sciences, social scientists and humanists publish in different formats, specifically, they rather produce books and contributions to edited volumes and monographs than journal articles (Bourke & Butler, 1996; Pestaña, Gómez, Fernández, Zulueta & Méndez, 1995; Nederhof, 2006; Sivertsen & Larsen, 2012). Books should not be ignored by bibliometrics, not only because they are a major output type but also due to their high impact. Hicks (1999) states that the best social science is often found in books, which is reflected in their citation rates. The danger of ignoring books is illustrated by research, which explores the differences between the worlds of book and journal publishing (e.g., Nederhof, van Leeuwen & van Raan, 2010; Butler & Visser, 2006; Amez, 2013; Clemens, Powell, Mcllwaine & Okamoto, 1995; Hicks & Potter, 1991; Bourke & Butler, 1996; Chi, 2014a). Furthermore, citations to and from books are distributed differently from those to and from journal articles, and often originate from outside the cited work's specialty (Broadus, 1971). Some studies show that books reference more books than articles, and journal articles refer to more articles than books (Larivière, Archambault, Gingras & Vignola-Gagné, 2006; Line, 1979), indicating that citations from journal articles are not the largest source of citations obtained by book publications.

Even though the importance of books in scholarly communication, notably in the social sciences and humanities, was proved by previous studies, only few and small-scale case studies investigating the characteristics of books were conducted by bibliometricians due to the lack of a reliable and comprehensive data source providing citation links. These studies either investigate the citations of so-called non-source items in the references of Web of

Science (WoS) journal papers (Butler & Visser, 2006; Hammarfelt, 2011; Amez, 2013; Chi, 2014a) or analyse citations in other alternative databases such as Google Books or Google Scholar (Kousha & Thelwall, 2009; Kousha, Thelwall & Rezaie, 2011; Samuels, 2011, 2013). All in all, large-scale bibliometric studies analysing the citation patterns of book literature have not been conducted in the past decade.

A new approach to explore citation patterns of books and its limitations

In 2011, Thomson Reuters released a new collection in the WoS, Book Citation Index (BKCI), to allow users to discover book literature and trace its comprehensive citation links alongside journal literature (Adams & Testa, 2011). BKCI covers over 60,000 editorially selected books starting from 2005 with an additional 10,000 new titles each year (Book Citation Index, 2015).

Even though the BKCI broadens the coverage of WoS and allows researchers to tackle studies based on numerous and qualified bibliographic data of books and book chapters in different aspects, the new database is not fully developed yet (Leydesdorff & Felt, 2012; Torres-Salinas, Robinson-García, Jiménez-Contreras & Delgado López-Cózar, 2012; Gorraiz, Purnell & Glänzel, 2013; Torres-Salinas, Robinson-García, Campanario & Delgado López-Cózar, 2013a; Torres-Salinas, Rodríguez-Sánchez, Robinson-García, Fdez-Valdivia & García, 2013b; Torres-Salinas, Robinson-García, Cabezas-Clavijo & Jiménez-Contreras, 2014). Some limitations mentioned in previous studies include:

- Coverage
  BKCI indexes 61% of 60,000 books in the social sciences and humanities (in November 2014, see Book Citation Index, 2015), which is not too arguable due to the nature of the publication behavior of scholars in different fields. However, its indexing bias in terms of language, country, and publisher is large. For example, 96% of the indexed books are written in English (Torres-Salinas et al., 2014) and the United States and England account for 35% of all publications and 75% of publishers in BKCI (Gorraiz et al., 2013; Torres-Salinas et al., 2014). Furthermore, Springer, Palgrave and Routledge alone account for 50% of the total database (Torres-Salinas et al., 2014) evincing a rather high concentration of publishers.
- Completeness of records
  Gorraiz et al. (2013) report the absence of affiliation data in BKCI but it has been confirmed by Torres-Salinas et al. (2014) that their later downloaded data does include affiliation information which could be used to analyse research units such as countries or institutions. Moreover, the low share of BKCI indexed items with references data (<30%, see Chi, 2014b) would also limit the validity of relevant studies.
- Document type classification
  A further limitation of the BKCI comes from the lack of a clear distinction of document types due to the different forms of book literature.
  - *Books*
    Gorraiz et al. (2013) argue that 'book' might be considered to be at a higher hierarchical level as 'journal' instead of being treated as a document type, and consequently point out the lack of cumulative citation counts from different hierarchies in BKCI. It is in line with the warning raised by Leydesdorff and Felt (2012) that monographs may be underrated in terms of citation impact or overrated using publication performance indicators. Furthermore, Gorraiz et al. (2013) question the fuzzy boundaries of subtypes of book and how to treat new editions.

  - *Monographs and edited volumes*

It was discovered that edited books usually have a greater impact than non-edited books (Leydesdorff & Felt 2012, Torres-Salinas et al., 2014, Chi, 2014a; Amez, 2013). This may be because of the effects of working collectively with a more diverse content and the higher average number of book chapters per book (Torres-Salinas et al., 2014). However, a global consensus on how to cite the book editor(s), the book author(s) or the author(s) of the book chapter is lacking (Gorraiz et al., 2013). Even though it is possible to distinguish bibliometrically between monographs and edited volumes among the type 'book', a normalization for the credit of a monograph is required (Leydesdorff & Felt, 2012).

- *Book series and annual series*
  BKCI covers annual series, which are part of the journal and series literature and indexed by other collections of WoS as well. They are assigned to the pubtype 'Journal' in BKCI (the other two pubtypes are 'Books' and 'Books in series'), and all are published by the publisher Annual Reviews. Leydesdorff and Felt (2012) indicate the problems from ignoring differences between book series and annual series. As noticed by Torres-Salinas et al. (2012, 2013b), this publisher presents an outlier pattern showing a behavior more closely linked to journals rather than monographs.

The research purposes of this study

In this study, we analyse and compare BKCI items jointly with journals literature to answer the following open questions based on the revealed limitations of using the database. Some of these questions have already been addressed but not yet answered by, e.g., Adams & Testa (2011) and Gorraiz et al. (2013). These issues apply to differences in citation impact between journal and book literature but also to the question whether edited books with different contributors for each chapter essentially deviate in their citation patterns from their monographic authored counterparts.

1. What is the feature of books in the sciences (including life sciences, natural sciences, technical sciences), social sciences and humanities through the lens of the BKCI?
2. Is there any difference between the ageing of periodical and monographic literature?
3. Is there a difference in citation patterns of edited and authored books?

The findings are expected to allow a better understanding of communication patterns in different publication types and the applicability of the BKCI for building indicators for use in both informetrics studies and research evaluation.

Methodology

Data sources

The complete 2005 volume of the Web of Science Core collection database including the three journal databases Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI) and Arts & Humanities Citation Index (A&HCI) as well as the Book Citation Index (BKCI) has been processed as source documents. The two proceedings editions of the core collection have been excluded because of the large overlap among the book, proceedings and journal databases (cf. Gorraiz et al., 2013). The choice of volume 2005 was made for two reasons, particularly, because 2005 was the first BKCI volume and this allowed us to trace citations till end of 2013, i.e., for a full period of nine years.

In addition, we have split up the BKCI database into two parts, namely those books that could be identified as edited books and the rest, which was considered to refer to authored books. Overlap with proceedings and journals were removed to obtain a correct dataset for the

analysis. Only so-called citable document types have been taken into account, that is, articles, letters and reviews for journals, books and citable book chapters for the BKCI. All documents extracted from the BKCI have been analysed both individually and aggregated to the book level.

Subject classification

All items extracted from the database have been assigned to the 74 individual subfields according to the *modified* Leuven-Budapest classification system. Multiple assignments are quite frequent at this level of granularity. The original scheme was introduced by Glänzel and Schubert (2003) and has been recently modified to provide a better categorisation for the social sciences and humanities. The modified version has been developed for the use with the BKCI but is also fully compatible with the journal and proceedings editions of the WoS Core Collection as it is based on the WoS and Journal Citation Reports (JCR) subject categories. Major fields and subfields in the sciences of the previous version have not been changed. The modified classification scheme is presented in Figure 1.

THE LEUVEN – BUDAPEST CLASSIFICATION SCHEME FOR THE SCIENCES, SOCIAL SCIENCES AND HUMANITIES

0. **MULTIDISCIPLINARY SCIENCES**
   - X0 multidisciplinary sciences

1. **AGRICULTURE & ENVIRONMENT**
   - A1 agricultural science & technology
   - A2 plant & soil science & technology
   - A3 environmental science & technology
   - A4 food & animal science & technology

2. **BIOLOGY (ORGANISMIC & SUPRAORGANISMIC LEVEL)**
   - Z1 animal sciences
   - Z2 aquatic sciences
   - Z3 microbiology
   - Z4 plant sciences
   - Z5 pure & applied ecology
   - Z6 veterinary sciences

3. **BIOSCIENCES (GENERAL, CELLULAR & SUBCELLULAR BIOLOGY; GENETICS)**
   - B0 multidisciplinary biology
   - B1 biochemistry/biophysics/molecular biology
   - B2 cell biology
   - B3 genetics & developmental biology

4. **BIOMEDICAL RESEARCH**
   - R1 anatomy & pathology
   - R2 biomaterials & bioengineering
   - R3 experimental/laboratory medicine
   - R4 pharmacology & toxicology
   - R5 physiology

5. **CLINICAL AND EXPERIMENTAL MEDICINE I (GENERAL & INTERNAL MEDICINE)**
   - I1 cardiovascular & respiratory medicine
   - I2 endocrinology & metabolism
   - I3 general & internal medicine
   - I4 hematology & oncology
   - I5 immunology

6. **CLINICAL AND EXPERIMENTAL MEDICINE II (NON-INTERNAL MEDICINE SPECIALTIES)**
   - M1 age & gender related medicine
   - M2 dentistry
   - M3 dermatology/urogenital system
   - M4 ophthalmology/otolaryngology
   - M5 paramedicine
   - M6 psychiatry & neurology
   - M7 radiology & nuclear medicine
   - M8 rheumatology/orthopedics
   - M9 surgery

7. **NEUROSCIENCE & BEHAVIOR**
   - N1 neurosciences & psychopharmacology
   - N2 psychology & behavioral sciences

8. **CHEMISTRY**
   - C0 multidisciplinary chemistry
   - C1 analytical, inorganic & nuclear chemistry
   - C2 applied chemistry & chemical engineering
   - C3 organic & medicinal chemistry
   - C4 physical chemistry
   - C5 polymer science
   - C6 materials science

9. **PHYSICS**
   - P0 multidisciplinary physics
   - P1 applied physics
   - P2 atomic, molecular & chemical physics
   - P3 classical physics
   - P4 mathematical & theoretical physics
   - P5 particle & nuclear physics
   - P6 physics of solids, fluids and plasmas

10. **GEOSCIENCES & SPACE SCIENCES**
    - G1 astronomy & astrophysics
    - G2 geosciences & technology
    - G3 hydrology/oceanography
    - G4 meteorology/atmospheric & aerospace science & technology
    - G5 mineralogy & petrology

11. **ENGINEERING**
    - E1 computer science/information technology
    - E2 electrical & electronic engineering
    - E3 energy & fuels
    - E4 general & traditional engineering

12. **MATHEMATICS**
    - H1 applied mathematics
    - H2 pure mathematics

13. **SOCIAL SCIENCES I (GENERAL, REGIONAL & COMMUNITY ISSUES)**
    - Y1 education, media & information science
    - Y2 sociology & anthropology
    - Y3 community & social issues

14. **SOCIAL SCIENCES II (ECONOMIC, POLITICAL & LEGAL SCIENCES)**
    - L1 business, economics, planning
    - L2 political science & administration
    - L3 law

15. **ARTS & HUMANITIES**
    - K0 multidisciplinary
    - K1 arts & design
    - K2 architecture
    - K3 history & archaeology
    - K4 philosophy & religion
    - K5 linguistics
    - K6 literature

Figure 1. The modified version of the Leuven-Budapest classification scheme for the WoS.

Data processing

In order to analyse citation impact and ageing patterns over subfields, we have calculated the following statistics:

- Annual citation rates (both increments and cumulated) for the year of publication 2005 (1) till 2013 (9). In this study, however, we only use cumulative citation impact in a three-year (x_3) and a nine-year (x_9) citation window.
- The ratio x_3/x_9 as a kind of prospective Price index and an indicator of ageing.

We have calculated all statistics on the basis of both individual book chapters, where available, and for the complete books. Chapters were considered the equivalent of journal articles in terms of the aggregation level. Unfortunately, chapter-based citation statistics proved not to be reliable since citations to individual chapters could not be identified in many cases as they were assigned to the book in the database. This is not necessarily due to the database producer: often the authors of the citing documents are responsible for this uncertainty. In order to avoid biased indicators or otherwise incomplete or distorted results we decided to use only citation indicators for complete books, which, of course, results in a serious loss of information and a more intricate interpretation. This applies above all to edited books, where chapters are authored by different contributors, and a distinction between different chapters would be of paramount importance.

A further issue is the small size of the publication set resulting from this restriction. We have found many subfields with fewer than 30 books each: This threshold might be critical for the interpretation and reliability of statistics like mean values and shares (e.g., Glänzel & Moed, 2013). Furthermore, we have not assigned books to corporate addresses of authors/editors because the availability of author affiliation in books is rather low (see, e.g., Gorraiz et al., 2013).

Results

It is not the aim of the present paper to study the subject coverage of the BKCI database since, on one hand, we can refer to the study by Adams and Testa (2011) in the context of broader subject areas and, on the other hand, a subject analysis at the level of subject categories can easily be conducted using the analyse tool of the web version of Thomson Reuters WoS Core Collection. Nevertheless we would just like to mention in passing that we can confirm that subfields in the social sciences and humanities have a better representation in the BKCI than in the other databases of the WoS.

Ten subfields had a share larger than 5% in the 2005 volume of the BKCI: Among those 10 subfields applied mathematics was the only representative of the sciences. Slightly more than 12% of all books could be assigned each to business, economics, planning and political science & administration, respectively. All books in the humanities (except for multidisciplinary and arts & design) as well as education, media & information science and sociology & anthropology in the social sciences were among the top ten in terms of subject representation.

In the first step we looked at citation patterns of book and journals literature by disciplines in a nine-year citation window. What we intended to do was not to compare citation impact over across fields but to compare subject-specific citation patterns between journals and books. It is a well-known fact that the subject is one of the factors influencing citation impact; the document type is another one (cf. Glänzel, 2013). Thus the publication type such as journal, proceeding, or monograph is expected to play a role in this context as well. Figure 2 plots the mean citation rates of subfields based on the nine-year citation window of books against the corresponding journal indicators. The volume year of the source items was 2005. Only subfields have been chosen in which at least 30 books have been published in that year. Subfields are ranked according the subfield impact in the BKCI. The results are somewhat unexpected here: Not the life sciences – as expected from journal literature – exhibit the highest citation impact for books but disciplines in chemistry and the geosciences. Consequently, the correlation between the corresponding x_9 values is medium ($r = 0.420$). In this respect, there are no dramatic differences between edited and authored books. The correlation between these two book types with $r = 0.762$ is relatively strong.

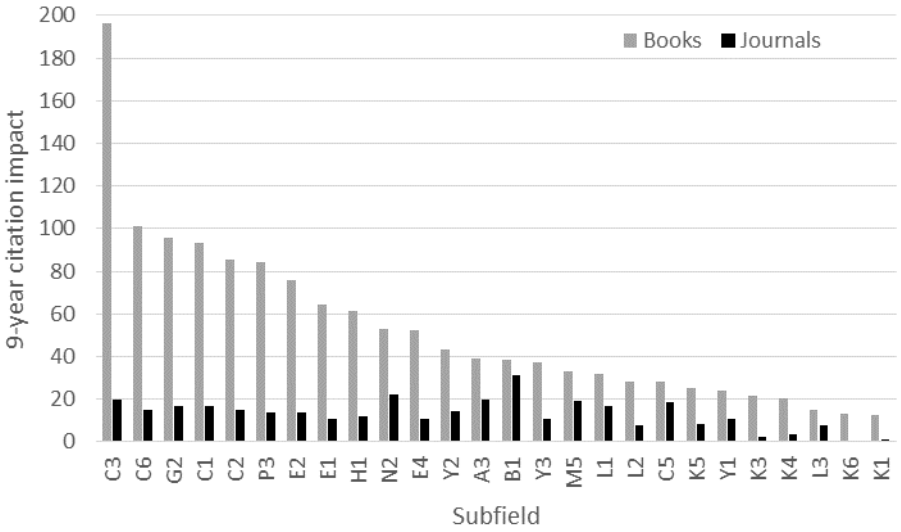


Figure 2. Most cited subfields in the mirror of the BKCI vs. SCIE/SSCI/AHCI. [Data sourced from Thomson Reuters Web of Science Core Collection].

It is known from journal literature that ageing is the fastest in the life and the natural sciences, followed by applied sciences, mathematics, social sciences and humanities (see Glänzel & Schoepflin, 1999). Ageing patterns can be characterised as a combination of phases of maturing and decline in citation processes (Glänzel & Schoepflin, 1995; Moed, van Leeuwen & Reedijk, 1998). The transition from the first to the second phase is marked by a peak in the annual increments of citation impact. This peak ranges according to the ageing of the discipline under study typically between the second and the fifth year beginning with the date of publication. The ratio (x_3/x_9) can thus serve as a proxy for literature ageing in the mirror of citation processes.

The plot of the prospective 'Price Index' (x_3/x_9) of books indexed in the 2005 volume of the BKCI against the corresponding journal indicators for the same volume is shown in Figure 3. The x_3/x_9 ratios are ranked in descending order according to the journal database editions of the WoS. At the left-hand side the disciplines with the fastest aging (highest ratios) can be found, while the low end is formed by slow-ageing subfields (cf. black bars in Figure 3). The grey bars representing the subfields in the BKCI show a rather subject-balanced situation. High (between 20% and 25%) as well as low (between 10% and 15%) shares can be found in both science and SSH subfields. The correlation between the x_3/x_9 ratios for books and journals is practically zero. This is illustrated in Figure 4. We just mention in passing that also the correlation between the corresponding ratios of edited and authored books is low ($r = 0.110$) as well. This substantiates that citation processes of books are more complex as these apparently depend on more factors than in the case of journal literature. Notably ageing seems not to be principally characterised by subject-specific peculiarities. Books are thus more heterogeneous information sources and addressed to more heterogeneous target groups than journals (and possibly proceedings).

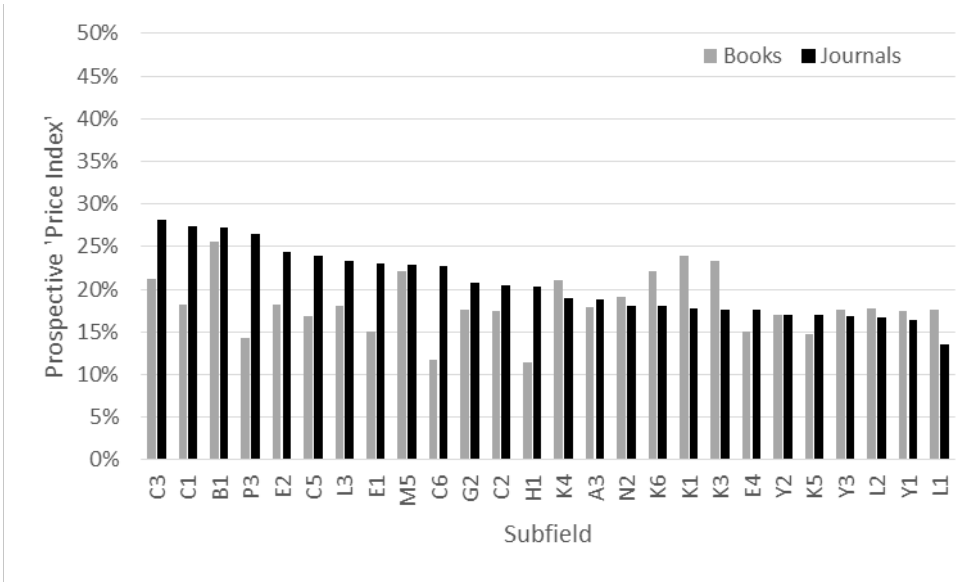


Figure 3. Prospective 'Price Index' of subfields in the BKCI vs. SCIE/SSCI/AHCI. [Data sourced from Thomson Reuters Web of Science Core Collection].

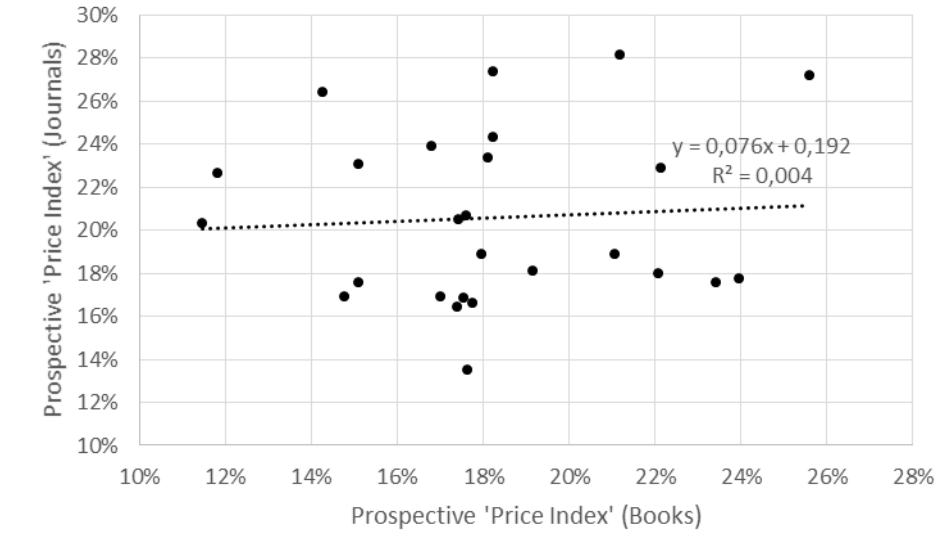


Figure 4. Scatter plot of prospective 'Price Index' of subfields in the BKCI vs. SCIE/SSCI/AHCI. [Data sourced from Thomson Reuters Web of Science Core Collection].

Conclusion

It is confirmed in this study that subfields in the social sciences and humanities have a higher representation in the BKCI (59%) than they have in the other databases of the WoS (12%). Disciplines in chemistry and the geosciences, instead of life sciences, have the highest citation impact for books. Humanities is the field having the highest difference between citation impact of books and journals. In contrast, life sciences have the most similar impact in books and journals. Compared to other sciences, technical sciences have relatively moderate characteristics in different perspectives.

It is not surprising to see that the social sciences and humanities have the largest increase of both the coverage and citation impact in the BKCI compared to journal literature in the other databases of the WoS. The BKCI could be an initial approach to explore wider targets of bibliometric analyses in the social sciences and humanities. The books in the basic sciences have unexpectedly high citation impact, whereas books in the life sciences do not reflect the

dominant position in journal literature but have been found to be on a relatively similar scale of citation counts as journals. This may imply that using BKCI data for bibliometric analyses in basic sciences would be a powerful approach to drag in more citation information.

For the ageing of periodical and monographic literature, the results of this study indicate a clear boundary between the two groups. The differences between books and journals are obvious, but the ageing of books is balanced between subjects. The differences between edited and authored books in terms of the 9-year citation impact are not so impressive as the other group books and journals. However, their disparities in ageing ratios are more evident than those of citation impact. The more complex citation processes of books, compared to journal literature, are shown in this study, the more heterogeneous characteristics of books should therefore be addressed.

The different ageing patterns of book and journal literature, i.e., books do not have as strong discipline specific patterns as journals, may lead to a universal condition for applying or building indicators in the collections of BKCI. It especially needs to be taken into account while designing indicators that are sensitive to the observed citation period. Moreover, the heterogeneous characteristics of books from their different formats such as edited or authored volumes result in more complex citation patterns than journals. These findings on the differences between periodical and monographic literature are worth further studies of indicator design to take into account.

References

Adams, J. & Testa, J. (2011). Thomson Reuters Book Citation Index. In E. Noyons, P. Ngulube, J. Leta (Eds.), *The 13th Conference of the International Society for Scientometrics and Informetrics* (Volume I, 13–18). Durban, South Africa: ISSI, Leiden University and the University of Zululand.

Amez, L. (2013). Citation patterns for social sciences and humanities publications. In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger & H. Moed (Eds.), *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference* (Volume II, 1891–1893). Vienna: AIT GmbH.

Book Citation Index (2015). Retrieved January 16, 2015 from: http://wokinfo.com/products_tools/multidisciplinary/bookcitationindex/

Bourke, P. & Butler, L. (1996). Publication types, citation rates and evaluation. *Scientometrics*, 37(3), 473–494.

Broadus, R. N. (1971). The literature of the social sciences: A survey of citation studies. *International Social Sciences Journal*, 23(2), 236–243.

Butler, L. & Visser, M. S. (2006). Extending citation analysis to non–source items. *Scientometrics*, 66(2), 327–343.

Chi, P. S. (2014a). Which role do non-source items play in the social sciences? A case study in political science in Germany. *Scientometrics*, 101(2), 1195–1213.

Chi, P. S. (2014b). T*he Characteristics and Impact of Non-Source Items in the Social Sciences – A Pilot Study of Two Political Science Departments in Germany*. Berlin: Humboldt-Universität zu Berlin.

Clemens, E. S., Powell, W. W., Mcllwaine, K. & Okamoto, D. (1995). Careers in print: Books, journals, and scholarly reputations. *American Journal of Sociology*, 101(2), 433–494.

Glänzel, W. (2013), *Bibliometrics as a research field.* Course script, KU Leuven, 3rd Edition.

Glänzel, W. & Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature. *Journal of Information Science*, 21(1), 37–53.

Glänzel, W. & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing and Management*, 35(3), 31–44.

Glänzel, W. & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, 56(3), 357–367.

Glänzel, W. & Moed, H. F. (2013). Opinion paper: thoughts and facts on bibliometric indicators. *Scientometrics*, 96(1), 381–394.

Gorraiz, J., Purnell, P. & Glänzel, W. (2013). Opportunities and limitations of the book citation index. *Journal of the American Society for Information Science and Technology*, 64(7), 1388–1398.

Hammarfelt, B. (2011). Interdisciplinarity and the intellectual base of literature studies: Citation analysis of highly cited monographs. *Scientometrics*, 86(3), 705–725.

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, 44(2), 193–215.

Hicks, D. & Potter, J. (1991). Sociology of scientific knowledge: A reflexive citation analysis or science disciplines and disciplining science. *Social Studies of Science*, 21(3), 459–501.

Kousha, K. & Thelwall, M. (2009). Google book search: Citation analysis for social science and the humanities. *Journal of the American Society of Information Science and Technology*, 60(8), 1537–1549.

Kousha, K., Thelwall, M. & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, 62(11), 2147–2164.

Larivière, V., Archambault, E., Gingras, Y. & Vignola-Gagné, E. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, 57(8), 997–1004.

Leydesdorff, L. & Felt, U. (2012), Edited volumes, monographs and book chapters in the Book Citation Index. *Journal of Scientometric Research*, 1(1), 28–34.

Line, M. B. (1979). The influence of the type of sources used on the results of citation analyses. *Journal of Documentation*, 35(4), 265–284.

Moed, H. F., van Leeuwen, T. N. & Reedijk, J. (1998). A new classification system to describe the ageing of scientific journals and their impact factors. *Journal of Documentation*, 54 (4), 387–419.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, 66(1), 81–100.

Nederhof, A. J, van Leeuwen, T. N. & van Raan, A. F. J. (2010). Highly cited non-journal publications in political science, economics and psychology: A first exploration. *Scientometrics*, 83(2), 363–374.

Pestaña, A., Gómez, I., Fernández, M. T., Zulueta, M. A. & Méndez, A. (1995). Scientometric evaluation of R&D activities in medium-size institutions: A case study based on the Spanish Scientific Research Council. In M. Koenig & A. Bookstein (Eds.), *Proceedings of the Fifth Biennial International Conference of the International Society for Scientometrics and Infometrics* (425–434). Medford: Learned Information, Inc.

Samuels, D. J. (2011). The modal number of citations to a political science article is greater than zero: Accounting for citations in articles and books. *PS: Political Science and Politics*, 44(4), 783–792.

Samuels, D. J. (2013). Book citations count. *PS: Political Science and Politics*, 46(4), 785–790.

Sivertsen, G. & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: an empirical analysis of the potential. *Scientometrics*, 91(2), 567–575.

Torres-Salinas, D., Robinson-García, N., Jiménez-Contreras, E. & Delgado López-Cózar, E. (2012). Towards a 'Book Publishers Citation Reports'. First approach using the 'Book Citation Index'. *Revista Española de Documentación Científica*, 35(4), 615–620.

Torres-Salinas, D., Robinson-García, N., Campanario, J. M. & Delgado López-Cózar, E. (2013a). Coverage, specialization and impact of scientific publishers in the Book Citation Index. *Online Information Review*, 38(1), 24–42.

Torres-Salinas, D., Rodríguez-Sánchez, R., Robinson-García, N., Fdez-Valdivia, J. & García, J. A. (2013b). Mapping citation patterns of book chapters in the book citation index. *Journal of Informetrics*, 7(2), 412–424.

Torres-Salinas, D., Robinson-García, N., Cabezas-Clavijo, Á. & Jiménez-Contreras, E. (2014). Analyzing the citation characteristics of books: edited books, book series and publisher types in the book citation index. *Scientometrics*, 98(3), 2113–2127.

When is an Article Actually Published? An Analysis of Online Availability, Publication, and Indexation Dates

Stefanie Haustein[1], Timothy D. Bowman[1] and Rodrigo Costas[2]

[1] *stefanie.haustein@umontreal.ca, timothy.bowman@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal (Canada)

[2] *rcostas@cwts.leidenuniv.nl*
Center for Science and Technology Studies, Leiden University, Wassenaarseweg 62A, 2333 AL Leiden (The Netherlands)

Abstract
With the acceleration of scholarly communication in the digital era, the publication year is no longer a sufficient level of time aggregation for bibliometric and social media indicators. Papers are increasingly cited before they have been officially published in a journal issue and mentioned on Twitter within days of online availability. In order to find a suitable proxy for the day of online publication allowing for the computation of more accurate benchmarks and fine-grained citation and social media event windows, various dates are compared for a set of 58,896 papers published by Nature Publishing Group, PLOS, Springer and Wiley-Blackwell in 2012. Dates include the online date provided by the publishers, the month of the journal issue, the Web of Science indexing date, the date of the first tweet mentioning the paper as well as the Altmetric.com publication and first seen dates. Comparing these dates, the analysis reveals that large differences exist between publishers, leading to the conclusion that more transparency and standardization is needed in the reporting of publication dates. The date on which the fixed journal article (Version of Record) is first made available on the publisher's website is proposed as a consistent definition of the online date.

Conference Topic
Journals, databases and electronic publications

Introduction
The process of scholarly communication, which usually begins with the formulation of a research idea and hypothesis and ends with publishing results to share them with the scientific community (Garvey & Griffith, 1964), has been sped up by means of electronic publishing (Dong, Loh, & Mondry, 2006; Wills & Wills, 1996). The publication delay, which Amat (2008, p. 382) defined as the "chronological distance between the stated date of reception of a manuscript by a given journal and its appearance on any print issue of that journal", has been accelerated by email and online manuscript handling systems as well as online publication (Wills & Wills, 1996). The delay period consists of the review process, which constitutes the main delay and ends with the acceptance of the manuscript, followed by technical delays of journal production and paper backlog.

Various studies have analyzed publication delays and found differences between scientific fields, journals, and publishers (e.g., Abt, 1992; Amat, 2008; Björk & Solomon, 2013; Das & Das, 2006; Diospatonyi, Horvai, & Braun, 2001; Dong et al., 2006). Since long delays interfere with priority claims and slow down scientific discourse, publication speed plays an important role for authors and scholarly communication (Rowlands & Nicholas, 2006; Schauder, 1994; Tenopir & King, 2000). Short publication delays can therefore be considered as a quality indicator reflecting the up-to-dateness of scientific journals (Haustein, 2012). Publishers have begun to reduce delays by making so-called *early view*, *in press*, *ahead of print* or *online first* versions of accepted papers available before they appear in an (print) issue. It has been shown for food research journals that online ahead of print publication has reduced publication delay by 29% (Amat, 2008), while Das and Das (2006) reported for 127 journals in 2005 average lags of three months between online and print issues publications

with particular differences between publishers. Tort, Targino, and Amaral (2012) showed that this lag increased significantly over time for six neuroscience journals. Online dates are now being recorded in bibliometric databases like Scopus, which impacts bibliometric analyses (Gorraiz, Gumpenberger, & Schlögl, 2014; Heneberg, 2013). Together with the increasing popularity of preprint servers (such as arXiv and SSRN) and institutional repositories, such *in press* versions have helped to speed up the read-cite-read cycle. As a result manuscripts increasingly cite papers that have not been officially published in a journal issue. Although scholarly communication has always involved sharing different versions of a manuscript with colleagues before, during, and after formal publication—such as exchanging drafts for feedback before submission or diffusing preprints after acceptance—, the electronic era makes these versions 'public', searchable, and (often) permanently retrievable on the web. To define and distinguish between various versions, the National Information Standards Organization (NISO) agreed upon the following versions of a journal article (NISO/ALPSP Working Group, 2008):

- Author's Original (AO) – manuscript ready to submit.
- Submitted Version Under Review (SMUR) – manuscript under formal peer review.
- Accepted Manuscript (AM) – version of journal article accepted for publication.
- Proof (P) – copy-edited version of accepted article.
- Version of Record (VoR) – fixed version of journal article formally published.
- Corrected Version of Record (CVoR) – VoR in which errors have been corrected.
- Enhanced Version of Record (EVoR) – VoR updated or enhanced with supplementary material.

It is important to note that by the NISO definition, the VoR is defined as a "fixed version of a journal article that has been made available by any organization that acts as a publisher by formally and exclusively declaring the article 'published'" (NISO/ALPSP Working Group, 2008, p. 3). This definition includes early views and in press articles without information on volume and issue or other identifiers as long as the content and layout of the article are fixed.

When it comes to bibliometric indicators, the acceleration of the publication process has been reflected in obsolescence patterns (Egghe & Rousseau, 2000) as well as citing half-lives (Luwel & Moed, 1998). These increasing online-to-print lags were shown to artificially increase citation rates including the immediacy index and impact factor (Heneberg, 2013; Seglen, 1997; Tort et al., 2012). The speed of scholarly communication becomes particularly visible in the context of social media metrics (the so-called altmetrics); for example, mentions of scientific documents on Twitter happen within hours (and sometimes within minutes) of online availability (Shuai, Pepe, & Bollen, 2012).

We argue that in the fast-moving digital era, the use of the publication *year* of the journal issue as the smallest level of time aggregation for bibliometric indicators is becoming insufficient, particularly in research evaluation contexts, due to the following factors:

a. acceleration of the read-cite-read cycle due to electronic publishing;
b. commonplace of online publication before publication of the journal issue; and
c. increasing online-to-print lags.

Following NISO's terminology, we suggest that the date of the first public online appearance of the VoR is the most relevant and should be used as the basic time unit to determine the official publication date of a paper. This would allow for the construction of more accurate citation and social media event windows, for example, citation windows of equal length (in days or months) for papers published in January or December, as well as the construction of more exact benchmarks by aggregating citations and social media events per week (e.g., tweets and Facebook shares) or month (citation rates) depending on the evaluation context.

Although many publishers now report online publication dates, many different dates are presented and the information provided varies between publishers, as no official standards

exist on publication dates. This paper explores and aims to verify various 'publication' dates in order to find a good proxy for the actual date of online availability. Thus, the paper aims to answer the following research questions:

1. Which publishers specify online dates and how do they provide them?
2. How reliable are dates provided by the publishers and how do they compare to each other?
3. What other existing dates can be used as a proxy of the online publication date of the VoR?

Methods and Materials

The dataset of this study was retrieved from the Web of Science (WoS) (as the major citation database) and is restricted to the publication year 2012 to limit effects of changes over time. To validate the publication dates provided by the publishers, the dates of the first tweet mentioning the particular paper were obtained from Altmetric.com. We argue that a tweet cannot link to a paper before it exists, thus the first tweet cannot have appeared before the online publication date. Tweets captured by Altmetric.com are linked to the documents via the DOI resulting in 313,301 WoS 2012 papers with at least one event captured by Altmetric.com (Haustein, Costas, & Larivière, 2015). Altmetric records that contained an arXiv ID or Astrophysics Data System (ADS) ID were removed to exclude tweets to preprints, which could have been made public before the online publication of the VoR. Twitter mentions are thus restricted to the mentions or links to the publisher's website, DOI, or PubMed ID.

Table 6. Top 10 publishers according to number of papers with types of dates available according to data provided by the publisher via API (a), in the metadata (m) of the webpage, on the webpage only (w), or as dynamic content only (d). Publishers selected for this study are highlighted in grey.

Publisher	Papers	Received	Revised	Accepted	Version of Record	Online	Publication	Date	Journal Issue	Journal Issue Online
Elsevier	51,292	d	d	d		d	a		w	
Wiley-Blackwell	47,958	w		w		m,w[i]	m		w,m	w
Lippincott	21,944							m	w,m	
Springer	19,225					m	m,a	m	w,m,a	
PLOS	16,208	w		w			a,m		a,m	
BMC	11,930	w		w			w,m		w,m	
NPG	11,181	w,m		w,m		m,a	w,m,a		w,m,a	
ACS	11,024							m,w	w	
Oxford	10,368	w		w		w		m	w,m	
Sage	8,776				w	w		m	w,m	

[i] Wiley provides two online dates "article published online" as well as "online date". See explanations below.

The top 10 publishers[1] of papers in the WoS-Altmetric dataset can be found in Table 1 together with the date information provided via API, in the metadata, in the webpage only, or as dynamic content of the webpage. It can be seen (in the headings of the table) that multiple terms exist to describe the online publication date and that multiple types of dates are made available on the website, in the metadata, or via the API; these include received, revised, accepted, version of record, online, publication, and date. Based on checking samples of articles for each of the publishers, we assume that the dates provided as *Version of Record*, *Online*, *Publication* and *Date* (Table 1) refer to (first) online appearances of the VoR required

[1] Publisher names from WoS were cleaned searching for name variants, but mergers and acquisitions were not accounted for. For example, BMC is considered an independent publisher, although it was acquired by Springer in 2008.

for this study. Wiley-Blackwell, Springer, PLOS, and Nature Publishing Group (NPG) were chosen due to their coverage and the technical feasibility of retrieving online date information. While Elsevier was the most represented publisher in this sample, it was difficult to obtain the required date information for their articles using PHP because this information is inserted dynamically into the webpage using JavaScript; Elsevier offers an API, but when queried[2] it was found to provide access to only the issue date and not to the online publication dates required for this study.

Using the DOI, the respective publishers' web platforms were queried to retrieve online dates. PLOS, Springer, and NPG each offer an API, but it was found that in some instances additional date information was only made available by searching the web page. In order to obtain the dates for Wiley, Springer and NPG, a PHP script was written that retrieved the HTML of the page. The HTML was then searched for metadata containing date information (e.g. <meta name="prism.publicationDate" content="2012-01-05"/>). When date information was found, it was saved to a relational database for evaluation. In instances where the article website had no (or missing) metadata available, the HTML was parsed and the contents of specific HTML tags found to contain date information was extracted and saved to a relational database; for the Wiley articles, a second script was written to retrieve dates not found in the metadata.

To compare different dates available and test in how far they can be used as proxies for online publication dates, other date information was obtained from WoS and Altmetric, so that together with the information from publishers the following dates were available:

- *online date*: retrieved from the publishers websites as part of the article metadata. For NPG ("Advance Online Publication"[3]), Springer ("Online First"[4]), and Wiley-Blackwell ("Early View"[5]) this date marks when the VoR was made publicly available on the publisher's website. For PLOS the online date equals the publication date because there is no difference between online and issue dates.
- *journal issue date*: the date from the journal issue as recorded by WoS. Since only a minority of papers provided the day of the month, the journal issue date was converted to the first of each month. Based on all 1.3 million papers in WoS published in 2012, 3.2% were published in issues spanning several months (such as JAN-FEB for a double issue). These were converted to the first day of the first month. A small percentage (0.5%) of papers appeared in seasonal issues (SPR, SUM, FAL, WIN). Since the data indicates that these are published at the beginning, middle, as well as the end of the particular season, these dates were disregarded. An additional 11.3% of all 2012 papers did not provide any issue date. Figure 1 provides an overview of the distribution of the 1.3 million WoS 2012 papers per journal issue date information.
- *Altmetric publication date*: the publication date as recorded by Altmetric.com, which is a mix of the journal issue date and online date (personal communication with Euan Adie and Jean Liu) as retrieved from the publisher. This is also the date Altmetric.com uses to compute the Altmetric score and provide benchmarks for papers of the same age. As shown in Figure 2, particular peaks can be observed for January 1 of each year as well as the first or last of each month. This might reflect common publishing practices, but could also be caused by aggregating data without actual day (and month) information. It was found that 15.1% of Altmetric.com records[6] did not have any publication date or they had incorrect dates (e.g. dates up to 2037).

[2] Using the http://api.elsevier.com/content/abstract/doi/{doi} API call
[3] http://www.nature.com/authors/author_resources/about_aop.html
[4] http://www.springer.com/authors/journal+authors/helpdesk?SGWID=0-1723213-12-817311-0
[5] http://olabout.wiley.com/WileyCDA/Section/id-404512.html#ev
[6] Based on 2.1 million Altmetric.com records collected in August 2014.

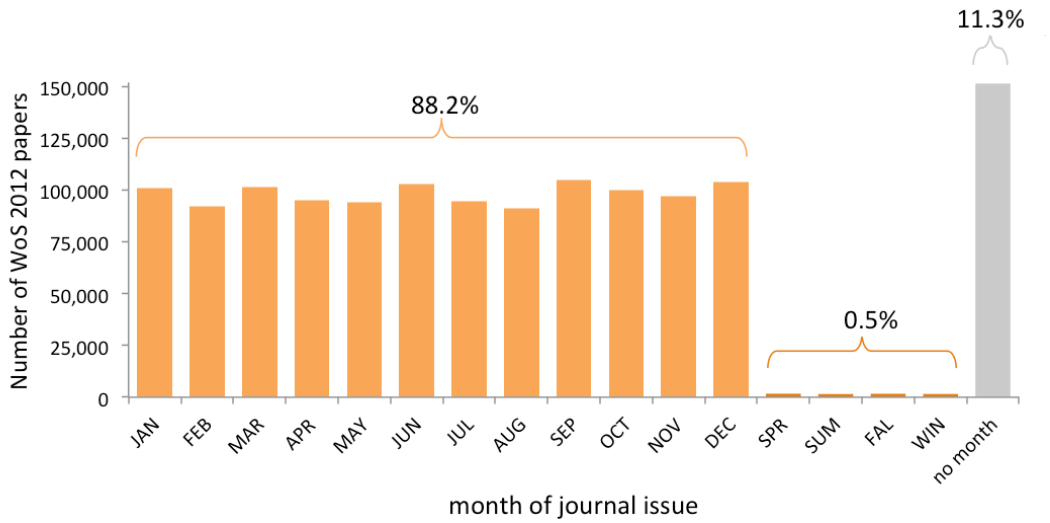


Figure 1. Number of WoS 2012 papers per months of journal issue.

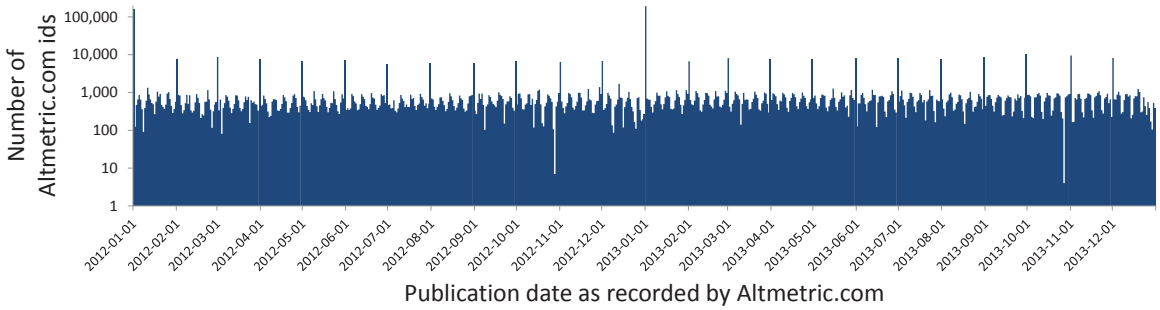


Figure 2. Number of Altmetric.com ids per Altmetric.com publication date from January 2013 to December 2014.

- *Altmetric first seen date*: the datestamp when Altmetric.com captured the first event for a particular document, which is missing for 4% of all records.[7]
- *First tweet date*: the datestamp of the first tweet[8] captured by Altmetric.com (excluding all papers with links to arXiv IDs or ADS IDs to ensure that the tweet did not refer to a preprint).
- *WoS indexing date*: the day when the document was indexed by WoS, which for 2012 papers was mostly during (37.7%) or in the month before (11.5%) or after (29.4%) the journal issue month.

In addition to the dates above we were also able to retrieve the following information for the papers published by Wiley-Blackwell:

- *Manuscript received*: the date the AO was submitted.
- *Manuscript accepted*: the date the AM was accepted.
- *Article first published online*: we could not determine the exact meaning of this date; for 95.6% of the total 34,507 Wiley-Blackwell documents it was identical with the *online date* and for 1.6% it was missing. For 2.3% of papers the *article first published online date* occurred before the *online date* by, on average, 35 days, which suggest

[7] Based on 2.1 million Altmetric.com records collected in August 2014.
[8] Twitter is the most common source covered by Altmetric.com (Robinson-García, Torres-Salinas, Zahedi, & Costas, 2014), so it makes sense to work with this date and not from other less common sources (e.g. Facebook or blogs).

that it marks the publication of the AM. However, in 137 cases (0.4%), it followed the *online date* by, on average, 52 days.

The final dataset—that is, the match of WoS, Altmetric.com, and papers with online dates retrieved from the four publishers—included 71,175 papers. For better comparison, it was restricted to papers for which all five dates tested as proxies for online publication (i.e., journal issue, Altmetric publication and first seen date, first tweet and WoS indexing date) were available. This amounted to a total of 58,896 papers, 12.5% NPG, 16.3% PLOS, 24.6% Springer and 46.6% Wiley-Blackwell.

Results and Discussion

Descriptive statistics comparing the online date to the five potential proxies are presented in Table 2, highlighting particular differences for the four publishers. Based on the assumption that the online date provided by the publishers were correct, the Altmetric publication date, first seen date, as well as the first tweet date seem to be the best proxies for online publication, while the journal issue and WoS indexing date show the largest deviations from the online publication dates. These differences reflect the nature of these dates. For example, Altmetric collects its publication dates from the publishers websites and while first tweets are known to happen shortly after publication (Shuai et al., 2012), WoS processing takes more time, namely, on average between 39 days for PLOS or 163 days for Springer papers. The 61 (NPG), 84 (Wiley-Blackwell), and 146 (Springer) days between online and journal issue date mostly reflect the backlog between online availability and publication of the journal issue. Although the (print) issue is generally assumed to follow online publication chronologically, results in Table 2 show that for 3.47% of Springer, 9.09% of Wiley-Blackwell, and 20.04% of NPG papers analyzed the online date came after the journal issue date, which is considered negative delay (Das & Das, 2006).

Although Altmetric and Twitter dates work better than journal issue and WoS indexing, none of the dates seem to reflect the online date well and large differences can be observed between publishers, in particular for Wiley-Blackwell, which questions the validity of any of the five dates as a reliable proxy of the publication of the VoR across publishers. The Altmetric publication date, which overall shows the smallest difference compared to the online date provided by the publishers—on average, 9 days for Springer, 12 days for NPG, 27 days for PLOS, and 121 for Wiley-Blackwell—is also problematic, because it is set to a date prior to online publication in 43.37% of Springer, 55.38% of NPG, 63.83% of Wiley-Blackwell, and 66.49% of PLOS papers. The variance between publishers affects Altmetric scores (but arguably also citation scores) when benchmarking a paper's scores against that of papers of the same reported age.

Based on the assumption that a tweet cannot mention a paper before it exists in the online space it links to, the online dates provided by Wiley-Blackwell seem to be the most problematic (Figure 3), as 14.52%[9] of the 27,432 analyzed papers had tweets linking to them before the date that the publisher identifies as the online publication date. On the other hand, none of the PLOS papers and few of the Springer (0.08%) articles were mentioned on Twitter before the online publication date. Although all of the papers analyzed have been tweeted, the mean number of days between online date and first tweet was higher than expected, ranging from 15 days for PLOS to 92 days for Springer. Moreover, the first mention on Twitter happened on the day of online publication for 1.06% (Springer) and 34.47% (NPG) sampled papers, which—particularly considering that about 80% of recent papers are never tweeted

[9] Results change only slightly when using the *article first published online* date, i.e. 14.61% of Wiley-Blackwell papers had a tweet appear before this date.

(Haustein, Costas, & Larivière, 2015)—limits the usefulness of the first tweet date as a proxy for online publication.

Table 2. Statistics for chronological distance (in number of days) of the journal issue month, Altmetric publication and first seen date, first tweet date and WoS indexing date with the online date for NPG, PLOS, Springer and Wiley-Blackwell.

Chronological distance to online date in number of days		NPG *n=7,391*	PLOS *n=9,600*	Springer *n=14,473*	Wiley-Blackwell *n=27,432*
Journal issue month[i]	% before	20.04%	n/a[ii]	3.47%	9.09%
	% identical	5.47%		0.11%	0.29%
	% after	74.50%		96.42%	90.62%
	mean	61		146	84
	standard deviation	78		111	93
	min	-330		-269	-423
	max	548		1,850	1,032
Altmetric publication date	% before	55.38%	66.49%	43.37%	63.83%
	% identical	39.35%	31.41%	34.11%	2.81%
	% after	5.28%	4.44%	22.52%	33.36%
	mean	12	27	9	121
	standard deviation	68	79	48	322
	min	-3,013	-697	-519	-16,761
	max	411	526	1,850	5,016
Altmetric first seen date	% before	3.48%	0.00%	0.08%	14.59%
	% identical	32.88%	36.64%	1.04%	14.26%
	% after	63.64%	63.36%	98.89%	71.15%
	mean	35	12	90	63
	standard deviation	87	49	164	122
	min	-459	0	-257	-533
	max	890	602	1,843	1,228
First tweet date	% before	3.52%	0.00%	0.08%	14.52%
	% identical	34.37%	37.23%	1.06%	15.21%
	% after	62.21%	62.77%	98.85%	70.27%
	mean	37	15	92	65
	standard deviation	92	59	169	127
	min	-459	0	-257	-533
	max	890	811	1,843	1,393
WoS indexing date	% before	2.72%	0.00%	0.10%	0.05%
	% identical	0.01%	0.00%	0.00%	0.00%
	% after	97.27%	100.00%	99.90%	99.95%
	mean	83	39	163	97
	standard deviation	81	20	113	94
	min	-302	9	-252	-359
	max	576	262	1,866	1,049

[i] First of the journal issue month as recorded by WoS.
[ii] PLOS does not distinguish between online and issue date, so that the two dates are actually identical.

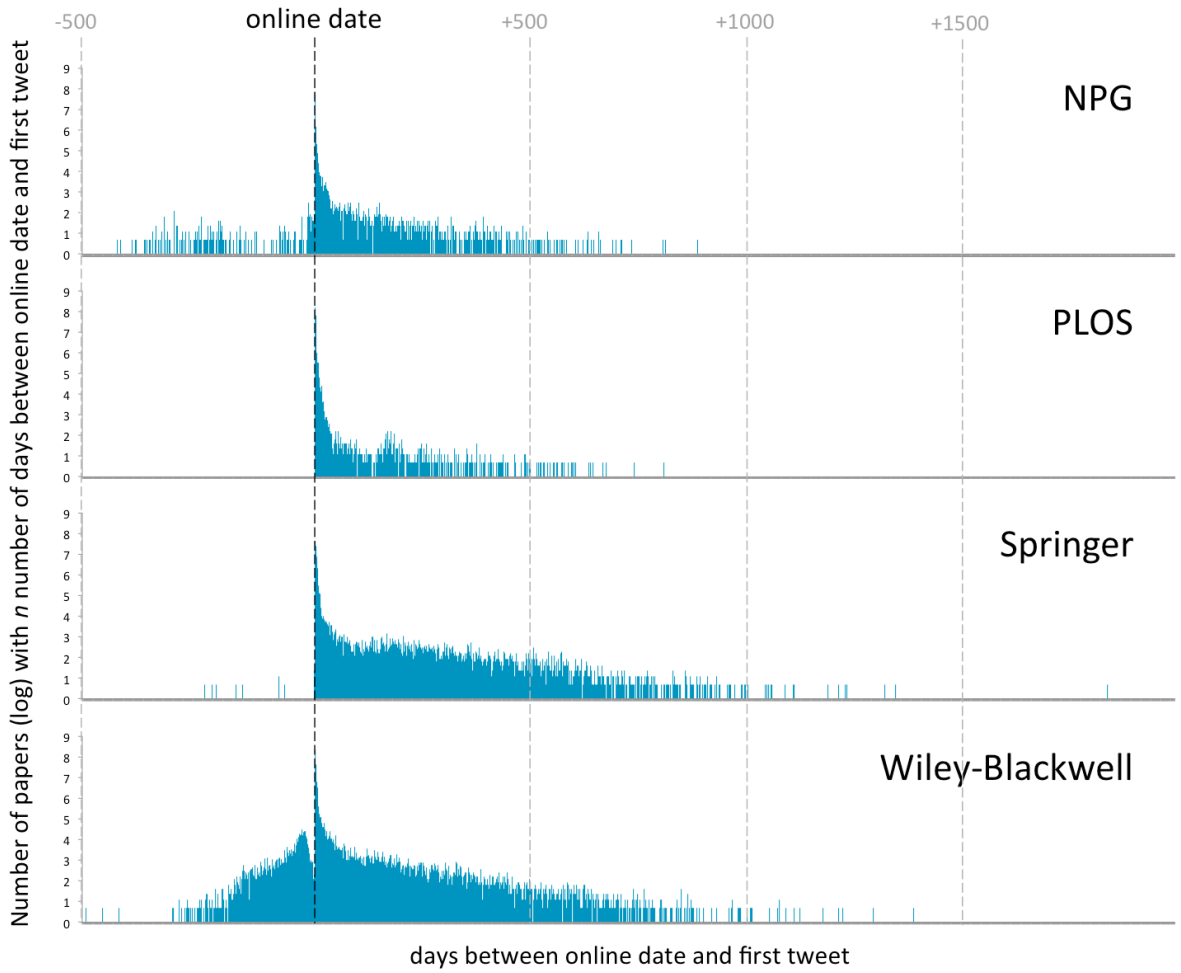


Figure 3. Number of papers (log) with *n* days between online date and first tweet per publisher.

Conclusions and Outlook

Currently none of the investigated dates represent a good proxy for the date a journal article was actually available online. In particular, the finding that a considerable amount of Wiley-Blackwell papers had been mentioned on Twitter before the online date, suggests that inconsistencies exist in terms of how publishers report online dates. This applies to the technical aspects as well as to actual content and vocabulary used. Thus, even when online dates can be retrieved from the publishers' websites or via API, they do not seem to always (and in a similar way for every publisher) mark the actual point in time when something was made accessible online. There is, thus, an urgent need for transparency and standardization of various dates reported by publishers in order to assure comparability of online dates across publishers. Adopting the vocabulary developed by NISO, specific dates could be reported for each version of the journal article, and the first appearance of the VoR would thus mark the date the fixed version of the document appeared online. A standardized vocabulary and a common definition of what various publication dates mean would not only improve benchmarking in the context of research evaluation but would also help to accurately determine the start of open access embargo periods required by certain funders, such as the NIH in the United States or the European Research Council. Currently these embargo periods, delaying green open access by a couple of months to years to protect publishers' revenue, are supposed to begin with publication of the article, which can refer to either journal issue or

online date.[10] Setting the start date of the embargo to the online publication date of the VoR would remove a potential loophole that allows the publishers to increase the embargo period during which they have the exclusivity of access.

Until such a standard is implemented, research on metrics should focus on obtaining more publisher-independent date information. One potential proxy for online publication could be the date when a DOI resolved successfully for the first time. Recently CrossRef has implemented the DOI Chronograph, a tool which tracks various deposits of metadata by the publisher as well as the first day of successful DOI resolution (Wass, 2015). Future work will investigate in how far these dates can be used to create fine-grained benchmarks needed in the context of social media metrics. Regarding citations, where monthly proxies are sufficient, the WoS Indexing date should be further investigated.

Acknowledgments

The authors would like to thank Euan Adie and Altmetric.com for access to their data and acknowledge funding from the Alfred P. Sloan Foundation, grant no. 2014-3-25.

References

Abt, H. A. (1992). Publication practices in various sciences. *Scientometrics*, *24*(3), 441–447. doi:10.1007/BF02051040

Amat, C. B. (2008). Editorial and publication delay of papers submitted to 14 selected Food Research journals. Influence of online posting. *Scientometrics*, *74*(3), 379–389. doi:10.1007/s11192-007-1823-8

Björk, B.-C., & Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics*, *7*(4), 914–923. doi:10.1016/j.joi.2013.09.001

Das, A., & Das, P. (2006). Delay between online and offline issue of journals: A critical analysis. *Library & Information Science Research*, *28*(3), 453–459. doi:10.1016/j.lisr.2006.03.019

Diospatonyi, I., Horvai, G., & Braun, T. (2001). Publication speed in analytical chemistry journals. *Journal of Chemical Information and Modeling*, *41*(6), 1452–1456. doi:10.1021/ci010033d

Dong, P., Loh, M., & Mondry, A. (2006). Publication lag in biomedical journals varies due to the periodical's publishing model. *Scientometrics*, *69*(2), 271–286. doi:10.1007/s11192-006-0148-3

Egghe, L., & Rousseau, R. (2000). The influence of publication delays on the observed aging distribution of scientific literature. *Journal of the American Society for Information Science*, *51*(2), 158–165. doi:10.1002/(SICI)1097-4571(2000)51:2<158::AID-ASI7>3.0.CO;2-X

Garvey, W. D., & Griffith, B. C. (1964). Scientific information exchange in psychology: The immediate dissemination of research findings is described for one science. *Science*, *146*(3652), 1655–1659. doi:10.1126/science.146.3652.1655

Gorraiz, J., Gumpenberger, C., & Schlögl, C. (2014). Usage versus citation behaviours in four subject areas. *Scientometrics*, *101*(2), 1077–1095. doi:10.1007/s11192-014-1271-1

Haustein, S. (2012). *Multidimensional Journal Evaluation. Analyzing Scientific Periodicals beyond the Impact Factor*. Berlin / Boston: De Gruyter Saur. doi:10.1515/9783110255553

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: the effect of document properties and collaboration patterns. *PLoS ONE*, *10*(3), e0120495. doi:10.1371/journal.pone.0120495

Heneberg, P. (2013). Effects of print publication lag in dual format journals on scientometric indicators. *PLOS ONE*, *8*(4), e59877. doi:10.1371/journal.pone.0059877

Luwel, M., & Moed, H. F. (1998). Publication delays in the scientific field and the their relationship to the ageing of scientific literature. *Scientometrics*, *41*(1-2), 29–40.

NISO/ALPSP Journal Article Versions (JAV) Technical Working Group. (2008). *Journal Article Versions (JAV): Recommendations of the NISO/ALPSP JAV Technical Working Group*. Baltimore. Retrieved from http://www.niso.org/publications/rp/RP-8-2008.pdf

Robinson-García, N., Torres-Salinas, D., Zahedi, Z., & Costas, R. (2014). New data, new possibilities: exploring the insides of Altmetric.com. *El Profesional de La Informacion*, *23*(4), 359–366. doi:10.3145/epi.2014.jul.03

Rowlands, I., & Nicholas, D. (2006). The changing scholarly communication landscape: An international survey of senior researchers. *Learned Publishing*, *19*, 31–55. doi:10.1087/095315106775122493

[10] http://authorservices.wiley.com/bauthor/faqs_fundingbodyrequirements.asp

Schauder, D. (1994). Electronic publishing of professional articles: Attitudes of academics and implications for the scholarly communication industry. *Journal of the American Society for Information Science*, *45*(2), 73–100. doi:10.1002/(SICI)1097-4571(199403)45:2<73::AID-ASI2>3.0.CO;2-5

Seglen, P. O. (1997). Citations and journal impact factors: questionable indicators of research quality. *Allergy*, *52*(11), 1050.

Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: article downloads, Twitter mentions, and citations. *PLoS ONE*, *7*(11), e47523. doi:10.1371/journal.pone.0047523

Tenopir, C., & King, D. W. (2000). *Towards Electronic Journals: Realities for Scientists, Librarians, and Publishers*. Washington, DC: Special Libraries Association.

Tort, A. B. L., Targino, Z. H., & Amaral, O. B. (2012). Rising publication delays inflate journal impact factors. *PLoS ONE*, *7*(12), e53374. doi:10.1371/journal.pone.0053374

Wass, J. (2015). Introducing the CrossRef Labs DOI Chronograph. *Crosstech blog post 12 January 2015*. Retrieved January 21, 2015, from http://crosstech.crossref.org/2015/01/introducing-chronograph.html

Wills, M., & Wills, G. (1996). The ins and the outs of electronic publishing. *Internet Research*, *6*(1), 10–21. doi:10.1108/10662249610123647

Analysis of the Obsolescence of Citations and Access in Electronic Journals at University Libraries

Chizuko Takei[1], Fuyuki Yoshikane[2] and Hiroshi Itsumura[3]

[1] *naoe.chizuko@ynu.ac.jp*
University of Tsukuba, Graduate School of Library, Information and Media Studies, 1-2 Kasuga, Tsukuba, Ibaraki (Japan)

[2] *fuyuki@slis.tsukuba.ac.jp,* [3] *hits@slis.tsukuba.ac.jp*
University of Tsukuba, Faculty of Library, Information and Media Science, 1-2 Kasuga, Tsukuba, Ibaraki (Japan)

Abstract

This study analyzes the correlation between the obsolescence of citations and access concerning a broad range of subjects, including fields that have not been dealt with in previous studies, shedding light on the differences between these two types of obsolescence and the characteristics for each field. The analysis investigates approximately 1,200 journals that were randomly sampled from 11 subject fields in SpringerLink and 20 subject fields in ScienceDirect. Metrics such as cited half-life and download half-life are employed to examine the relationship between the rate of obsolescence of citations and access. As a result, no strong correlation between citations and access is observed in most fields with regard to the short-term obsolescence. As for the long-term obsolescence, on the other hand, comparatively strong and significant correlations are seen in natural sciences other than medicine-related fields ($p < 0.05$).

Conference Topic

Journals, databases and electronic publications

Introduction

This study analyzes the relationship between the obsolescence of citations and access for usage of electronic journals in Japanese university libraries. The Big Deal, which is a package contract for electronic journals, has been rapidly adopted among Japanese university libraries. Irrespective of the university's size, the Big Deal drastically increased the number of accessible titles of journals at contract universities. However, with ongoing budget cuts and increasing journal prices, price hikes for the Big Deal are putting pressure on library budgets. This situation makes it difficult for libraries not only to subscribe to new journals but also to maintain existing subscriptions. As withdrawal from the Big Deal results in a drastic decrease in the number of accessible titles of journals, and thereby a collapse of the library's academic information framework, collection building of journal backfiles is necessary to alleviate the impact of these losses.

The collection development of journal backfiles differs from that of current files, which have a strong tendency to become fixed owing to budgetary considerations. This is because library staffs at many universities select and propose journal backfiles to be introduced under their own direction, for example, by utilizing special proposals received from publishers shortly before the accounting period. However, few Japanese universities have sought to implement a planned introduction of journal backfiles by scrutinizing the level of on-campus demand and the effectiveness of such an introduction.

As Takei, Yoshikane, and Itsumura (2013) pointed out, effective methods of collecting journal backfiles have rarely been studied in the literature. Investigating the development of backfiles requires perspectives focusing on the articles that fall into disuse, that is, obsolescence. Slower obsolescence represents stronger demand of researchers for older articles in the concerned field. Obsolescence analysis has been performed on library

collections to evaluate a decrease in the use of documents over time. The obsolescence of books is assessed on the basis of the number of times a book is used by lending year and accession year. In contrast, obsolescence of journals is based on citations and access to documents. Understanding the relationship between the obsolescence of citations and access will make it possible to estimate the obsolescence of access on the basis of information regarding the obsolescence of citations. This relationship has already been examined in certain fields, such as chemistry, and for specific journals, as will be described in the next section. However, the nature of documental use (citations and access) varies by field, and trends in the differences between the obsolescence of citations and access may also differ by field. Thus, this study employs several indices of obsolescence, some of which had not been adopted before our previous study (Takei, Yoshikane & Itsumura 2013), and analyzes obsolescence of access and citations for a wide range of subjects, including fields that have not previously been examined. We shed light on the differences between both types of obsolescence and their characteristics in each field.

Related Research

There are some indices for analyzing the relationship between citations and downloads (access). Impact Factor (IF), Immediacy Index (II), and Cited Half-life (CHL) are major indices of citations, while Download Impact Factor (DIF), Download Immediacy Index (DII), Download Half-life (DHL), and Usage Half-life (UHL), which is used as a synonym of DHL, are indices of downloads. According to the definition of Journal Citation Reports (JCR), IF is "the average number of times articles from the journal published in the past two years have been cited in the JCR year," II is "the average number of times an article is cited in the year it is published," and CHL is "the median age of the articles that were cited in the JCR year." IF and II indicate how frequently articles in the journal are cited within several years after publication and immediately after publication, respectively. CHL shows the degree of demand for older articles in the journal. In contrast, DIF and DII analogically apply the definitions of IF and II to downloads, respectively, and both DHL and UHL replicate the definition of CHL to access. Using these indices, many studies have been conducted on the relationship between citations and downloads to evaluate journal collections. For instance, Duy and Vaughan (2006) analyzed local citation data and IF with journal usage in the fields of chemistry and biochemistry. Good correlations were seen between local citation data and journal usage, whereas no significant correlation was observed between IF and journal usage. Other examples can be found in Chu and Krichel (2007), McDonald (2007), Bollen and van de Sompel (2008), and Watson (2009). In particular, there are some studies on obsolescence of access and citations related to electronic journals. For instance, Nicholas et al. (2005) surveyed synchronous obsolescence of access, revealing that over half of all usage was accounted for by items published within the last 15 months. Moreover, several studies have analyzed the relationship between obsolescence of citations and access by calculating and comparing the densities of citations and access (e.g., Kurtz et al., 2005; Moed, 2005; Brody et al., 2006).

In recent years, Schloegl and Gorraiz (2010; 2011) conducted more multifaceted studies related to oncology and pharmacology, using indices such as IF, II, and CHL. In the case of oncology journals in 2006, the results indicated that the means of UHL and CHL were 1.7 years and 5.6 years, respectively. Similar results were found in the case of pharmacology journals in the same year. Furthermore, they calculated CHL and found a medium-sized correlation between CHL and UHL in pharmacology ($r = 0.42$). Wan et al. (2010) examined the relationship between DII and citation indicators using the Chinese full-text database, the Chinese National Knowledge Infrastructure (CNKI).They found that DII had the potential to be a predictor for other indices such as h-index. While a moderate correlation between DII

and II was observed in the field of agriculture and forestry (r = 0.57), a strong correlation was found in psychology (r = 0.8). In addition, Gorraiz, Gumpenberger and Schloegl (2013) investigated the differences in obsolescence between citations and downloads in five fields in ScienceDirect, and Guerrero-Bote and Moya-Anegon (2013) observed the influence of language on the relationship between citations and downloads.

However, these analyses have only been performed for limited fields, including organic chemistry, astronomy, and astrophysics, and for selected journals in those fields. Although our previous work analyzed the obsolescence of citations and access with regard to all fields in Springer's SpringerLink and suggested the predictability of the long-term obsolescence of access on the basis of that of citations (Takei et al., 2013), its sample size for each field was small and insufficient for generalizing the results for the whole field.

Therefore, this study examines Elsevier's ScienceDirect in addition to SpringerLink to increase the sample size. SpringerLink is a collection comprising 11 fields focusing on Science, Technology, and Medicine (STM), whereas ScienceDirect is a collection comprising 23 fields including social sciences as well as STM. Analyzing both collections will enable a survey for a wider range of fields; besides, as for the fields included in both, it will facilitate an analysis based on more samples. It is assumed that indices of obsolescence that are effective for predicting the effects of backfiles will differ by field. Utilizing data of the two collections, we clarify the relationship in obsolescence between citations and downloads for each field.

Methodology

This study targeted Yokohama National University (YNU) in Japan, a medium-sized national university without a medical school. YNU consists of four undergraduate colleges (Education and Human Sciences, Economics, Business Administration, and Engineering Science) and five graduate schools (Education, International Social Sciences, Engineering, Environment and Information Sciences, and Urban Innovation). The university comprises around 600 full-time teaching staff and 10,000 students (around 2,600 graduate and 7,500 undergraduate students).

The survey employed the 2009–2012 editions of JCR as citation data, and statistics on the use of full text by publication year in the style of COUNTER Journal Report 5 for SpringerLink (2010–2012) and ScienceDirect (2001–2012) as access data. COUNTER Journal Report 5 defines the number of downloads, the number of times accessed, and the number of times used as the number of times the "full text" of an article is used. As with many studies, we employed this definition and referred to it as access count. COUNTER report has some limitations, for example, it does not reflect all of researchers' activities or could not distinguish the number of access by unique users. However, it reflects a certain amount of user's needs and it is useful to evaluate journal collections. We examined all the 11 fields in SpringerLink and 20 of the 23 fields in ScienceDirect (excluding Decision Science, Nursing and Health Professions, and Veterinary Science and Veterinary Medicine, for which the number of journals suitable for our analysis was less than 10). Because, for both collections, statistics contained sections in which the access count for multiple publication years had been summed up, the access count was divided by the number of years in the section to calculate the access count for each year.

The main concern of this study is to examine the practical predictability of local usage (i.e., access count in a given university) for each field based on global citation data, which is easily available from JCR, for collection management. Although local data does not always correspond with global data as shown in earlier studies (e.g., Duy & Vaughan, 2006; Bollen & van de Sompel, 2008), there may be a certain relationship between them because the

former is a part of the latter and the former partly reflects the latter. Thus, we compared local access data to global citation data in order to reveal the predictability of local access.

The sampling procedure was as follows. First, from all 2,782 journals in SpringerLink and all 1,792 journals in ScienceDirect, we extracted the journals whose fields could be identified on the basis of the title lists of publishers, excluding journals whose full text had never been accessed at YNU. As for ScienceDirect, where journals are classified into multiple fields, this study employed the fields first listed in Web of Science to ensure the same analysis conditions as for SpringerLink. Consequently, 1,567 and 1,657 journals were selected from SpringerLink and ScienceDirect, respectively.

Next, journals with index values listed in the relevant edition of JCR were sampled and rearranged in descending order of cumulative ratio of access counts for each field. These journals were separated into three layers according to the cumulative ratio of access counts as illustrated in Figure 1, i.e., less than 70%, 70% up to (not including) 90%, and 90% and above.

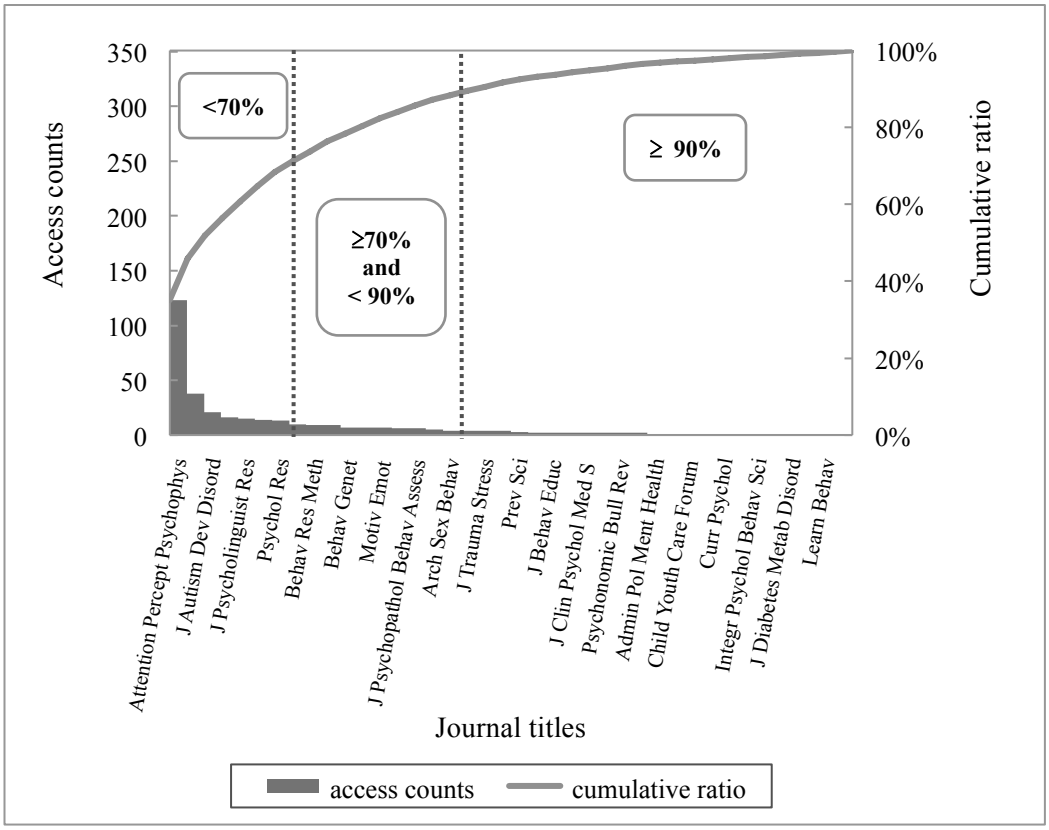


Figure 1. An example of 3 layers according to the cumulative ratio of access counts (Behavioral Science in SpringerLink).

To examine overall trends in each field, 15 journals were then randomly sampled from each of the layers in each field other than the three fields of ScienceDirect described above; for layers with less than 15 journals, all journals were considered. On this occasion, we sampled the journals that fulfilled the following conditions to obtain data for calculating the indices regarding obsolescence as of 2011 and 2012:

(a) Journals whose access count in 2011 and 2012 is not zero to analyze long-term obsolescence.

(b) Journals included in collections from 2011 to 2012 to analyze short-term obsolescence.

(c) Journals that fulfill the conditions of both (a) and (b) to examine the relationship between the two types of obsolescence.

As a result, the number of titles that became the targets of research was as follows:

SpringerLink: (a) 417, (b) 469, (c) 135

ScienceDirect: (a) 773, (b) 752, (c) 571

Tables 1 and 2 show the number of titles by field in the collections of SpringerLink and ScienceDirect, respectively. With regard to the sampling condition (c), we excluded 6 fields of SpringerLink (Behavioral Science; Business and Economics; Computer Science; Humanities, Social Sciences and Law; Mathematics and Statistics; and Medicine) and one field of ScienceDirect (Psychology) for which we obtained only 10 samples or less.

Table 1. Number of titles by field in SpringerLink

Subject	Sampling condition (a)	Sampling condition (b)	Sampling condition (c)
Behavioral Science (BS)	17	30	N/A
Biomedical and Life Sciences (BL)	45	45	32
Business and Economics (BE)	29	40	N/A
Chemistry and Materials Science (CM)	45	45	35
Computer Science (CS)	40	45	N/A
Earth and Environmental Science (EE)	45	45	30
Engineering (EG)	42	42	16
Humanities, Social Sciences and Law (HS)	30	42	N/A
Mathematics and Statistics (MS)	45	45	N/A
Medicine (MD)	34	45	N/A
Physics and Astronomy (PA)	45	45	22
Whole	417	469	135

Table 2. Number of titles by field in ScienceDirect

Subject	Sampling condition (a)	Sampling condition (b)	Sampling condition (c)
Agricultural and Biological Sciences (AB)	41	41	41
Biochemistry, Genetics and Molecular Biology (BG)	45	45	45
Business, Management and Accounting (BM)	36	34	20
Chemical Engineering (CE)	40	40	40
Chemistry (CH)	36	35	35
Computer Science (CS)	45	45	35
Earth and Planetary Sciences (EP)	45	45	43
Economics, Econometrics and Finance (EF)	45	45	30
Energy (EN)	22	21	16
Engineering (EG)	45	45	45
Environmental Science (ES)	36	36	35
Health Sciences (HE)	45	43	20
Immunology and Microbiology (IM)	37	37	17
Materials Science (MT)	43	42	43
Mathematics (MA)	36	36	21
Neuroscience (NS)	38	34	12
Pharmacology, Toxicology and Pharmaceutical Science (PT)	30	29	18
Physics and Astronomy (PA)	33	33	32
Psychology (PC)	36	29	N/A
Social Sciences (SS)	39	37	23
Whole	773	752	579

Sampling conditions: (a) Journals whose access count in 2011 and 2012 is not zero to analyze long-term obsolescence; (b) Journals included in collections from 2011 to 2012 to analyze short-term obsolescence; (c) Journals that fulfill the conditions of both (a) and (b) to examine the relationship between the two types of obsolescence.

This study employs the following indices as measures of obsolescence:

(1) Obsolescence of citations:
  (1A) Cited Half-life (CHL)
  (1B) Immediacy Index/Impact Factor (II/IF), i.e., ratio between II and IF
(2) Obsolescence of access:
  (2A) Download Half-life (DHL)
  (2B) Download Immediacy Index/Download Impact Factor (DII/DIF), i.e., ratio between DII and DIF

CHL and DHL express slower obsolescence, while II/IF and DII/DIF express faster obsolescence, as values become higher. In addition, whereas CHL and DHL are indices of obsolescence of use that take into consideration long periods of time, II/IF and DII/DIF particularly focus on the change in usage during several years after publication. DII/DIF, the ratio between DII and DIF, had not been used in obsolescence analysis before our previous study (Takei et al., 2013). However, given that the use of journals is generally concentrated at the time immediately after publication, it seems that DII/DIF would also prove useful as an index representing the nature of documental use in each field. For example, as for 2012, DII/DIF of Medicine is 5368.33 whereas DII/DIF of Earth and Environmental Science is 41.17 in SpringerLink. This means that the former field tends to progress quickly and the "latest" findings attract a lot of attention in the field whereas the latter field is inclined to emphasize not only the "latest" results but also previous ones. Therefore, DII/DIF was used in combination with II/IF in this study. The survey examined the degree of accordance—that is, correlation—of obsolescence between citations and access for each field with respect to the long-term (CHL and DHL) and the short-term (II/IF and DII/DIF). First, the values of these indices were calculated as of 2012. Data for CHL, II, and IF was obtained from the JCR of 2012. DHL, DII, and DIF analogically apply the definitions of CHL, II, and IF in JCR, respectively, to access count. To compute these indices, we set the sampling conditions (a) and (b) described above. In the analysis of short-term obsolescence based on the sampling condition (b), DII and DIF were used with the addition of one to avoid division by zero. Furthermore, to compare the tendencies in 2012 with those in the preceding year (i.e., to observe changes in documental use), the values as of 2011 were also obtained in the same manner.

If good correlations are found between the indices of citations and access in some fields, the information of CHL or II/IF obtained from JCR greatly helps us to determine the strategy to collect journal backfiles for these fields. That is, the correlations suggest the predictability of the use of journal backfiles by the information that can be obtained before introducing them.

Results

First, to determine the degree of accordance of obsolescence of citations and access, correlations between each pair of indices were observed: (A) between CHL and DHL; and (B) between II/IF and DII/DIF. The samples for analyzing (A) and (B) were extracted on the sampling conditions (a) and (b), respectively. The distributions of II/IF and DII/DIF had high values of skewness (2.71–12.97). Moreover, we cannot obtain exact values for CHL from JCR, in which the maximum value of CHL is 10, that is, even if its true value is greater than 10, CHL is described as 10. Thus, Spearman's rank correlation coefficient ρ was employed instead of Pearson's product-moment correlation coefficient r, which should be applied to interval or ratio scale data following a normal distribution.

Table 3 shows the correlation coefficients for (A) CHL and CHL and those for (B) II/IF and DII/DIF by field. There are differences between SpringerLink and ScienceDirect, both in the number and scope of fields. Therefore, to make it easier to compare the results of both collections, we reclassified all fields into the following 6 fields: Humanities and Social Sciences, Medicine, Chemistry and Engineering, Mathematics and Computer Science, Agricultural and Environmental Science, and Physics, as shown in Table 3.

As for 2012, the correlation coefficients for all fields were (A): $\rho = 0.50$ ($p < 0.05$) and (B): $\rho = 0.04$ ($p < 0.05$) in SpringerLink; (A): $\rho = 0.30$ ($p < 0.05$) and (B): $\rho = 0.03$ in ScienceDirect. While a moderate correlation was observed for (A), almost no correlation was found for (B). With regard to individual fields, in the case of (A), the strongest and statistically significant correlation was seen for Physics and Astronomy ($\rho = 0.59$, $p < 0.05$) in SpringerLink and for Energy ($\rho = 0.62$, $p < 0.05$) in ScienceDirect.

Table 3. Rank correlation ρ of obsolescence between citations and access.

Subject		2012 (A)		2012 (B)		2011 (A)		2011 (B)	
Humanities and Social Sciences	BS (S)	0.25		0.04		0.11		−0.10	
	BE (S)	0.46	*	0.07	*	0.32		−0.10	
	HS (S)	0.33		0.13		0.04		0.14	
	BM (E)	0.09		−0.27		−0.31		0.28	
	EF (E)	0.26		0.01		0.13		0.08	
	PC (E)	0.16		0.22		−0.04		0.00	
	SS (E)	0.05		−0.07		0.36	*	−0.04	
Medicine	BL (S)	0.51	*	0.28		0.29		0.40	*
	MD (S)	0.32		0.19		0.40	*	0.39	*
	HE (E)	0.09		−0.06		0.22		0.17	
	IM (E)	0.05		0.06		0.18		0.24	
	NS (E)	0.30		−0.31		0.18		0.08	*
	PT (E)	0.08		0.05		0.27		0.04	
Chemistry and Engineering	CM (S)	0.57	*	0.09		0.62	*	0.00	
	EG (S)	0.50	*	0.04	*	0.72	*	0.26	
	BG (E)	0.26		0.15		0.50	*	0.22	
	CE (E)	0.60	*	0.32	*	0.57	*	0.28	
	CH (E)	0.30	*	0.05		0.66	*	0.10	*
	EG (E)	0.34	*	0.04		0.42	*	0.26	
	MT (E)	0.56	*	0.07		0.56	*	0.03	
Mathematics and Computer Science	CS (S)	0.43	*	−0.06		0.45	*	0.09	
	MS (S)	0.43	*	0.07		0.52	*	−0.11	
	CS (E)	0.25		0.13		0.23		0.17	
	MA (E)	0.36	*	0.05		0.41	*	−0.20	
Agricultural and Environmental Science	EE (S)	0.47	*	0.02		0.53	*	0.03	
	AB (E)	0.15		0.04		0.36	*	0.18	
	ES (E)	0.46	*	−0.24		0.39	*	0.18	
Physics	PA (S)	0.59	*	0.08		0.39	*	−0.12	
	EP (E)	0.32	*	0.27		0.32	*	−0.21	
	EN (E)	0.62	*	0.11		0.73	*	0.23	
	PA (E)	0.35	*	0.10		0.33		−0.30	
Whole	(S)	0.50	*	0.04	*	0.45	*	0.01	
	(E)	0.30	*	0.03		0.37	*	0.08	*

(A): correlations between the indices of long-term obsolescence (CHL and DHL) on the sampling condition (a).
(B): correlations between the indices of short-term obsolescence (II/IF and DII/DIF) on the sampling condition (b)
(S): fields in SpringerLink. (E): fields in ScienceDirect. *Significant ($p < 0.05$)

In the case of (B), the correlation was significant and stronger in Chemical Engineering (ρ = 0.32, $p < 0.05$) in ScienceDirect than in other fields, and negative correlations were witnessed in some fields unlike in the case of (A). Meanwhile, as for 2011, the correlation coefficients for all fields were (A): $\rho = 0.45$ ($p < 0.05$) and (B): $\rho = 0.01$ in SpringerLink; (A): $\rho = 0.37$ ($p < 0.05$) and (B): $\rho = 0.08$ ($p < 0.05$) in ScienceDirect. With regard to individual fields, the correlation between indices changed according to the base years of observation. In the case of (A), for example, while Energy showed the strongest significant correlation both in 2012: ρ = 0.62 ($p < 0.05$) and in 2011: $\rho = 0.73$ ($p < 0.05$), the correlation for Chemistry varied from ρ = 0.66 ($p < 0.05$) in 2011 to 0.30 ($p < 0.05$) in 2012 in ScienceDirect. In the case of (B), for example, the correlation for Medicine varied from $\rho = 0.39$ ($p < 0.05$) in 2011 to 0.19 in 2012 in SpringerLink.

Concerning the 6 fields after reclassification, somewhat strong and significant correlations were seen between the indices of long-term obsolescence (CHL and DHL) in natural sciences other than Medicine, particularly in Physics and in Chemistry and Engineering.

Engineering (EG), Computer Science (CS), and Physics and Astronomy (PA) are included in both SpringerLink and ScienceDirect. Comparing SpringerLink and ScienceDirect, we find differences in the degree of correlation for these fields. The access count of the latter fluctuated considerably by year compared to that of the former in YNU. The gap between global data and unrepresentative local data might result in these differences.

Furthermore, we examined the correlations of pairs of indices for journal usage, including pairs other than (A) and (B), based on the sampling condition (c). To enable comparison with the results of previous studies and to take into account the strength of raw values, Pearson's product-moment correlation r was also studied along with Spearman's rank correlation ρ. When calculating the product-moment correlations, the data was logarithmically transformed to reduce skewness of distribution. As examples, Tables 4 and 5 show the correlation coefficients for SpringerLink (in 2012). Similar results were also obtained for SpringerLink (in 2011) and ScienceDirect (in 2011 and 2012). An example of these was shown in Table 6. The gray-colored cells in the tables indicate the correlations between the indices for citations and access, and moreover, the cells enclosed in boxes indicate the correlations between the indices relating to the obsolescence of citations and access. Little difference exists between the results of the three types of correlations, i.e., the rank correlation and the product-moment correlations before and after logarithmic transformation.

Table 4. Rank correlation ρ between indices for all 6 fields in 2012 in SpringerLink on the sampling condition (c).

	II	IF	DII	DIF	CHL	DHL	II/IF	DII/DIF
II	1	0.81 *	0.17 *	0.24 *	−0.04	−0.01	0.53 *	0.00
IF		1	0.05	0.20 *	−0.01	0.07	0.01	−0.15
DII			1	0.55 *	0.07	−0.19 *	0.10	0.57 *
DIF				1	0.21 *	0.01	0.05	−0.30 *
CHL					1	0.53 *	−0.03	−0.11
DHL						1	−0.10	−0.20 *
II/IF							1	0.12
DII/DIF								1

*Significant ($p < 0.05$)

Among pairs of the indices relating to obsolescence, while the strongest significant correlation (around 0.5, $p < 0.05$) was observed between CHL and DHL, which are the indices corresponding to (A), only weak correlations were found in the remaining pairs. However, an exception was found for Energy (ScienceDirect in 2011): a strong and positive

correlation was also seen between II/IF and DII/DIF, the indices corresponding to (B), as shown in Table 7.

Table 5. Product-moment correlation *r* after logarithmic transformation between indices for all 6 fields in 2012 in SpringerLink on the sampling condition (c).

	II	IF	DII	DIF	CHL	DHL	II/IF	DII/DIF
II	1	0.82 *	0.09	0.18 *	−0.03	0.05	0.57 *	−0.08
IF		1	0.04	0.19 *	−0.01	0.08	0.00	−0.15
DII			1	0.63 *	0.07	−0.21 *	0.10	0.57 *
DIF				1	0.19 *	0.01	0.03	−0.28 *
CHL					1	0.56 *	−0.04	−0.11
DHL						1	−0.03	−0.27 *
II/IF							1	0.08
DII/DIF								1

*Significant ($p < 0.05$)

Table 6. Rank correlation *ρ* between indices for all 6 fields in 2011 in SpringerLink on the sampling condition (c).

	II	IF	DII	DIF	CHL	DHL	II/IF	DII/DIF
II	1	0.81 *	0.11	0.02	0.00	0.20 *	0.59 *	0.07
IF		1	0.16	0.13	0.08	0.19 *	0.08	0.04
DII			1	0.58 *	−0.04	−0.22 *	−0.09	0.58 *
DIF				1	0.07	−0.14	−0.22 *	−0.27 *
CHL					1	0.54 *	−0.05	−0.08
DHL						1	0.15	−0.12
II/IF							1	0.10
DII/DIF								1

*Significant ($p < 0.05$)

Table 7. Rank correlation *ρ* between indices for Energy in 2011 in ScienceDirect on the sampling condition (c).

	II	IF	DII	DIF	CHL	DHL	II/IF	DII/DIF
II	1	0.86 *	0.73 *	0.62 *	−0.12	−0.30	0.71 *	0.33
IF		1	0.49	0.69 *	−0.30	−0.37	0.36	0.05
DII			1	0.55 *	−0.01	−0.19	0.74 *	0.71 *
DIF				1	−0.06	−0.07	0.29	−0.08
CHL					1	0.77 *	0.23	0.15
DHL						1	0.01	0.02
II/IF							1	0.64 *
DII/DIF								1

*Significant ($p < 0.05$)

Discussion and Conclusions

Results of the analysis indicated that, for 8 fields of SpringerLink and 7 fields of ScienceDirect, statistically significant positive correlations of over 0.4 were observed between CHL and DHL, which are the indices of long-term obsolescence, in both or either year. Furthermore, having reclassified all fields of both collections into 6 fields, comparatively strong and significant correlations were seen between CHL and DHL in natural sciences other

than Medicine, particularly in Physics and in Chemistry and Engineering. This result suggests that, to a certain degree, it is possible to predict the long-term obsolescence of access on the basis of the value of CHL obtained from JCR with regard to natural sciences.

In addition to Spearman's rank correlation coefficients ρ, we also examined the correlations between indices for all fields using Pearson's product-moment correlation coefficients r, and no major differences were observed between both types of correlations. Comparing with previous studies such as Schloegl and Gorraiz (2010; 2011) and Wan et al. (2010), our results indicated the same tendency regarding the indices of long-term obsolescence (CHL and DHL). However, in the case of other indices, a different tendency was observed. Wan et al. (2010), for example, investigated many indices and reported the following correlations between indices: DII and II showing $\rho = 0.24$ ($p = 0.0964$), DII and IF showing $\rho = 0.41$ ($p = 0.0034$), II and IF showing $\rho = 0.59$ ($p < 0.0001$) in agriculture and forestry; DII and II showing $r = 0.8$ in psychology. Meanwhile, in this study, almost no correlations were witnessed between DII and II and between DII and IF in most fields, whereas strong and significant correlations were observed between II and IF ($\rho = 0.81$, $r = 0.82$) as indicated in Tables 4 and 5. This is thought to be partly due to the characteristics of local use along with differences in the fields and databases. For example, citation speed in YNU may be slower than that of global trends, or research areas of researchers in YNU may be specific and narrow, i.e., a large proportion of the journals that they read may not be core journals for their research and thus their research activities (citations) may not correspond to global trends. If one focuses on this issue, the relationship between local access and local citation should be investigated. In addition to this, citation age may also influence the results. Citation age is larger than publication time lag of the citing article, which is mostly around one year. In contrast, downloads (access) tend to be concentrated in the publication year, that is to say, there is little time lag. This might cause different tendencies of downloads and citations in the short-term (e.g., weak correlation between DII and II in Tables 4–6).

Furthermore, the results of 2011 and 2012 for both collections indicate that the degree of correlation in several fields such as Chemistry may vary considerably by year, and the indices with a strong correlation differ depending on the field. Regarding the variation in the indices of short-term obsolescence (II/IF and DII/DIF), we can guess that it would be easily influenced by such factors as the change in the number of papers, the frequency of publication, and special issues of journals. In contrast, regarding the variation in the indices of long-term obsolescence (CHL and DHL), factors such as the transfer to another publisher, title change, and discontinuation of publication may exert influence.

This study focused on the relationship between the obsolescence in local access and global citation for the purpose of grasping the predictability of the former based on the latter. Although one should take into consideration various ways such as cost-effectiveness (e.g., Bergstrom et al., 2014) when introducing journal backfiles efficiently, our approach would also be useful for making a decision.

In future research, aiming to clarify the characteristics themselves of document use by researchers in Japan, we will investigate the citation data in Japanese universities, including YNU, and compare it with the corresponding access data. Moreover, we would like to observe the obsolescence of access and citation for a longer period for further examination of the tendency concerning the variation in the relationship between them.

Acknowledgments

This work was partially supported by Grant-in-Aid for Scientific Research (C) 23500294 (2013) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and we would like to show our gratitude to the support.

References

Bergstrom, T. C., Courant, P. N., McAfee, R. P. & Williams, M. A. (2014). Evaluating big deal journal bundles. *Proceedings of the National Academy of Sciences*, 111(26), 9425–9430.

Bollen, J. & van de Sompel, H. (2008). Usage impact factor: The effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology*, 59(1), 136-149.

Brody, T., Harnad, S. & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8), 1060-1072.

Chu, H. & Krichel, T. (2007). Downloads vs. citations in economics: Relationships, contributing factors & beyond. In: *Proceedings of the 11th International Society for Scientometrics and Informetrics Conference* (pp. 207-215). Madrid, Spain.

Duy, J., & Vaughan, L. (2006). Can electronic journal usage data replace citation data as a measure of journal use? An empirical examination. *The Journal of Academic Librarianship*, 32(5), 512-517.

Gorraiz, J., Gumpenberger, C., & Schloegl, C. (2013). Differences and similarities in usage versus citation behaviours observed for five subject areas. In: *Proceedings of the 14th International conference of the international Society for Scientometrics and Informetrics (ISSI2013)* (pp. 519-535). Vienna: University of Wien.

Guerrero-Bote, V. P., & Moya-Anegon, F. (2013). Relationship between downloads and citation and the influence of language. In: *Proceedings of the 14th International conference of the international Society for Scientometrics and Informetrics (ISSI 2013)* (pp. 1469-1484). Vienna: University of Wien.

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Murray, S. S., Martimbeau, N., & Elwell, B. (2005). The bibliometric properties of article readership information. *Journal of the American Society for Information Science and Technology*, 56(2), 111-128.

McDonald, J. D. (2007). Understanding journal usage: A statistical analysis of citation and use. *Journal of the American Society for Information Science and Technology*, 58(1), 39-50.

Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology*, 56(10), 1088-1097.

Nicholas, D., Huntington, P., Dobrowolski, T., Rowlands, I., Jamali, M. H. R., & Polydoratou, P. (2005). Revisiting 'obsolescence' and journal article 'decay' through usage data: An analysis of digital journal use by year of publication. *Information Processing and Management,* 41(6), 1441-1461.

Schloegl, C., & Gorraiz, J. (2010). Comparison of citation and usage indicators: The case of Oncology journals. *Scientometrics*, 82(3), 567-580.

Schloegl, C., & Gorraiz, J. (2011). Global usage versus global citation metrics: The case of Pharmacology journals. *Journal of the American Society for Information Science and Technology,* 62(1), 161-170.

Takei, C., Yoshikane, F., & Itsumura, H. (2013). Use of electronic journals in university libraries: an analysis of obsolescence regarding citations and access. In: *Proceedings of the 14th International Conference of the International Society for Scientometrics and Informetrics (ISSI 2013)* (pp. 1772-1783). Vienna: University of Wien.

Wan, J.-K., Hua, P.-H., Rousseau, R., & Sun, X.-K. (2010). The journal download immediacy index (DII): Experiences using a Chinese full-text database. *Scientometrics*, 82(3), 555-566.

Watson, A. B. (2009). Comparing citations and downloads for individual articles. *Journal of Vision*, 9(4), 1-4.

Dynamics between National Assessment Policy and Domestic Academic Journals

Eleonora Dagienė[1] and Ulf Sandström[2]

[2] *eleonora.dagiene@vgtu.lt*
Vilnius Gediminas Technical University, Sauletekio al. 11, LT-10223 Vilnius (Lithuania)

[2] *ulf.sandstrom@indek.kth.se*
KTH, Indek — Department of Industrial Economics and Management,
Lindstedtsvägen 30, 10044 Stockholm (Sweden)

Introduction

Normally research assessment methodologies assume that the highest scores should be given to articles published in recognised high impact journals. While these high impact journals are mostly published in the US and UK, lower citation rates are particular to journals published in other countries. Subsequent to expansion of the Web of Science in 2007–2009, the research platform was generously augmented with scientific journals issued by local publishers of non-English speaking countries (Leeuwen et al., 2001; van Raan, van Leeuwen, & Visser, 2011). Analysts agree that papers in national journals are usually less frequently cited in comparison to articles published in English (Haiqi & Yamazaki, 1998; Meneghini & Packer, 2007; Moed, 2002; Ponomariov & Toivanen, 2014; Russell, 1998; Tijssen et al., 2006). Research evaluations in several Eastern European countries largely build on data from Thomson Reuters and Elsevier databases. An overview provided by Dejan Pajić (Pajić, 2014) demonstrates that methodologies of most countries award papers in leading international journals rather than national ones. In some countries, articles published in national journals either receive a lower score or are given no score. The Lithuanian methodology is but an illustration of this.

The way a journal reflects the internationalized nature of science may be determined by many methods, one of which is based on the distribution of authoring and citing countries (Zitt & Bassecoulard, 1998).

The aim of the paper is to analyse the impact of the national assessment policy on the development of research journals published in the same country.

Lithuanian Assessment Methodologies and Journal Publishing in Lithuania 2005–2013

Five Lithuanian research assessment methodologies were designed in the period 2005–2010. It should be underlined that there is a great difference between assessment of papers in Sciences and papers in Social Sciences & Humanities. While in Social Sciences and Humanities, researchers have to be published in peer-reviewed journals only, papers in the Sciences have especially high requirements: to gain a score, they have to be published in journals indexed by Web of Science and have an impact factor. The methodology of 2010 was grossly disadvantageous to most Lithuanian journals as it was centred on papers published in high ranking journals (Maskeliūnas, 2011). Lithuanian research journal publishing and other quantitative indicators as well as technical publishing issues have already been analysed in several papers (Dagiene, 2011, 2013). In 2006, Thomson Reuters Web of Science database had only 5 indexed Lithuanian journals; while in 2007, it had 21; and since 2008, there were 29 journals in WoS with Lithuania as the publishing country. One supplementary journal—BALT J OF MANAGEMENT—has been added to this list although its country of origin is England and it is published by Emerald, the Editor-in-Chief and the Managing Editor are from Lithuania.

Data and Methodology

All data analysed in this research has been retrieved from the Web of Science databases: SCIE, SSCI and A&HCI. All indicators employed in this research and listed below have been analysed for two periods: 2008–2010 and 2011–2013. This is done because Lithuanian methodology was changed in 2010, using not only journal impact factors but also JCR data with thresholds measuring the "citation quality" of journals. The main quantitative and qualitative indicators of the Lithuanian journals are presented in the appendix. NJCS – Normalized journal citation score is the impact of the journal set normalized in relation to its sub-fields (average=1.00) (Sandström, 2009).

Citation indicators showed an improvement over the recent years: in 2011–2013, the number of cites by foreign researchers increased by 10% compared to 2008–2010; besides, citation from core journals increased by 19%, which confirms the growing internationalization of Lithuanian journals.

Figure 1 presents dynamics of internationalization indicators of Lithuanian journals.

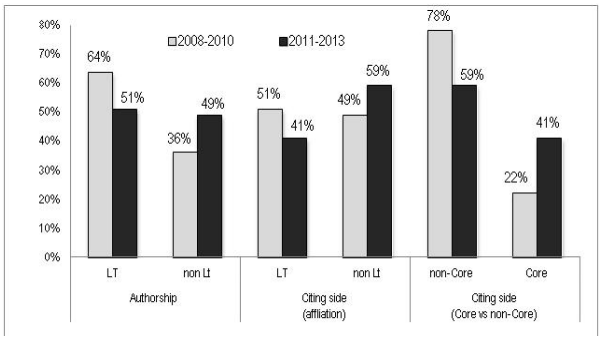


Figure 1. Dynamics of internationalization indicators of Lithuanian journals.

Authorship: from period I to period II, there's an overall drop in LT share and growth of foreign researchers from 36% to 49% if we count averages of all LT journals.

Conclusions

National policy has an influence on scholarly communication and puts the pressure on the national journals. There is some tension but also a response from the journals; thus, over a short period of time we see rather substantial changes.

Firstly, from 2008–2010 to 2011–2013, the relative share of the Lithuanian authors in authorship became smaller; secondly, papers published in Lithuanian journals are more often cited by researchers affiliated to non-Lithuanian institutions; thirdly, papers published in Lithuanian journals are more often cited by papers published in core journals defined as such by Leiden (CWTS 2014).

References

CWTS Leiden Ranking (2014) Retrieved, March 3, 2014, from: http://www.leidenranking.com/methodology/indicators

Dagiene, E. (2011). Changes in Lithuanian research journal publishing in 2009–2010. *Sciecominfo, 7*(1). Retrieved from http://nile.lub.lu.se/ojs/index.php/sciecominfo/article/view/4906

Dagiene, E. (2013). Progressive Opportunities for Research Journal Publishing. *Proceedings of the 5th Belgrade International Open Access Conference 2012: Journal Publishing in Developing, Transition and Emerging Countries* (pp. 11–23). Centre for Evaluation in Education and Science. doi:10.5937/BIOAC-94

Haiqi, Z., & Yamazaki, S. (1998). Citation indicators of Japanese journals. *Journal of the American Society for Information Science, 49*(4), 375–379. doi:10.1002/(SICI)1097-4571(19980401)49:4<375::AID-ASI7>3.0.CO;2-X

Leeuwen, T. N. Van, Moed, H. F., Tijssen, R. J. W., Visser, M. S., & Raan, A. F. J. Van. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics, 51*(1), 335–346. doi:10.1023/A:1010549719484

Meneghini, R., & Packer, A. L. (2007). Is there science beyond English? Initiatives to increase the quality and visibility of non-English publications might help to break down language barriers in scientific communication. EMBO Reports, 8(2), 112–6. doi:10.1038/sj.embor.7400906

Moed, H. F. (2002). Measuring China's research performance using the Science Citation Index. *Scientometrics, 53*(3), 281–296. doi:10.1023/A:1014812810602

Pajić, D. (2014). Globalization of the social sciences in Eastern Europe: genuine breakthrough or a slippery slope of the research evaluation practice? *Scientometrics*. doi:10.1007/s11192-014-1510-5

Ponomariov, B., & Toivanen, H. (2014). Knowledge flows and bases in emerging economy innovation systems: Brazilian research 2005–2009. *Research Policy, 43*(3), 588–596. doi:10.1016/j.respol.2013.09.002

Russell, J. M. (1998). Publishing patterns of Mexican scientists: Differences between national and international papers. *Scientometrics, 41*(1-2), 113–124. doi:10.1007/BF02457972

Sandström, U. (2009). Bibliometric evaluation of research programs. The Swedish Environmental Protection Agency, 81 p.

Tijssen, R. J. W., Mouton, J., van Leeuwen, T. N., & Boshoff, N. (2006). How relevant are local scholarly journals in global science? A case study of South Africa. *Research Evaluation, 15*(3), 163–174. doi:10.3152/147154406781775904

Van Raan, A. F. J., van Leeuwen, T. N., & Visser, M. S. (2011). Severe language effect in university rankings: particularly Germany and France are wronged in citation-based rankings. *Scientometrics, 88*(2), 495–498. doi:10.1007/s11192-011-0382-1

Zitt, M., & Bassecoulard, E. (1998). Internationalization of scientific journals: A measurement based on publication and citation scope. *Scientometrics, 41*(1-2), 255–271. doi:10.1007/BF02457982

Appendix. The main quantitative and qualitative indicators of the Lithuanian journals.

Journal title	Period I – 2008-10 II – 2011-13	THREE MOST FREQUENT COUNTRIES (TOP3) in the authors' affiliations	LT Authorship	TOP3 Authorship	Shift Towards International	NJCS 1=Global avg.
Included in Science Citation Index Expanded (SCI-EXPANDED) – Web of Science Core Collection						
BALT ASTRON	I	LITHUANIA | CZECH REPUBLIC | USA	22.17%	46.95%		0.11
	II	LITHUANIA | ESTONIA | USA	6.95%	34.89%	25.7%	0.07
BALT FOR	I	LITHUANIA | ESTONIA | FINLAND	35.96%	77.34%		0.21
	II	LITHUANIA | ESTONIA | FINLAND	30.54%	62.29%	19.5%	0.19
BALT J ROAD BRIDGE E	I	LITHUANIA | SOUTH KOREA | ITALY	62.95%	77.07%		0.65
	II	LITHUANIA | POLAND | ITALY	45.74%	66.60%	13.6%	0.68
BALTICA	I	LITHUANIA | ESTONIA | LATVIA	36.47%	70.20%		0.29
	II	LITHUANIA | ESTONIA | RUSSIA	74.93%	85.57%	-21.9%	0.12
CHEMIJA	I	LITHUANIA | IRAN | INDIA	94.01%	98.33%		0.14
	II	LITHUANIA | IRAN | BULGARIA	85.94%	91.06%	7.4%	0.08
ELEKTRON ELEKTROTECH	I	LITHUANIA | LATVIA | ROMANIA	61.67%	77.21%		0.25
	II	LITHUANIA | LATVIA | PEOPLES R CHINA	40.10%	58.08%	24.8%	0.21
INFORMATICA-LITHUAN	I	LITHUANIA | SLOVENIA | PEOPLES R CHINA	57.78%	74.81%		1.08
	II	LITHUANIA | PEOPLES R CHINA | TAIWAN	46.00%	62.77%	16.1%	1.04
INF TECHNOL CONTROL	I	LITHUANIA | POLAND | ALGERIA	81.15%	86.89%		0.34
	II	LITHUANIA | TAIWAN | PEOPLES R CHINA	61.51%	88.17%	-1.5%	0.56
J CIV ENG MANAG	I	LITHUANIA | POLAND | TURKEY	43.73%	69.33%		1.28
	II	LITHUANIA | POLAND | TAIWAN	30.03%	54.69%	21.1%	0.71
J ENVIRON ENG LANDSC	I	LITHUANIA | TURKEY | ESTONIA	70.28%	80.47%		0.47
	II	LITHUANIA | TURKEY | INDIA	71.68%	82.57%	-2.6%	0.26
J VIBROENG	I	LITHUANIA | LATVIA | POLAND	66.10%	82.03%		0.11
	II	LITHUANIA | PEOPLES R CHINA | POLAND	28.57%	84.18%	-2.6%	0.41
LITH J PHYS	I	LITHUANIA | UKRAINE | INDIA	88.91%	91.61%		0.12
	II	LITHUANIA | LATVIA | RUSSIA	69.43%	83.55%	8.8%	0.09
LITH MATH J	I	LITHUANIA | GERMANY | HUNGARY	72.27%	83.33%		0.42
	II	LITHUANIA | PEOPLES R CHINA | GERMANY	51.10%	75.64%	9.2%	0.31
MATER SCI-MEDZ	I	LITHUANIA | ESTONIA | CZECH REPUBLIC	83.44%	90.16%		0.18
	II	LITHUANIA | ESTONIA | LATVIA	64.50%	79.20%	12.2%	0.22
MATH MODEL ANAL	I	LATVIA | ESTONIA | LITHUANIA	20.61%	59.02%		0.51
	II	LATVIA | LITHUANIA | PEOPLES R CHINA	18.28%	55.28%	6.3%	0.51
MECHANIKA	I	LITHUANIA | ROMANIA | ALGERIA	71.28%	83.67%		0.51
	II	LITHUANIA | PEOPLES R CHINA | IRAN	48.57%	76.89%	8.1%	0.41
MED LITH	I	LITHUANIA | ESTONIA | USA	92.33%	94.77%		0.11
	II	LITHUANIA | LATVIA | ESTONIA	67.40%	84.24%	11.1%	0.17
NONLINEAR ANAL-MODEL	I	LITHUANIA | INDIA | BANGLADESH	64.86%	82.97%		0.50
	II	LITHUANIA | INDIA | PEOPLES R CHINA	47.62%	75.62%	8.9%	0.61
TRANSPORT-VILNIUS	I	LITHUANIA | PEOPLES R CHINA | TURKEY	56.83%	67.51%		1.19
	II	LITHUANIA | PEOPLES R CHINA | SERBIA	43.10%	65.38%	3.2%	0.56
VET ZOOTECH-LITH	I	LITHUANIA | POLAND | ESTONIA	82.13%	91.88%		0.13
	II	LITHUANIA | POLAND | ESTONIA	69.36%	83.67%	8.9%	0.11
ZEMDIRBYSTE	I	LITHUANIA | ITALY | POLAND	73.74%	86.59%		0.19
	II	LITHUANIA | TURKEY | POLAND	59.79%	80.30%	7.3%	0.35
Included in Social Sciences Citation Index (SSCI) and Arts & Humanities Citation Index (A&HCI) – Web of Science Core Collection						
BALT J OF MANAGEMENT	I	ESTONIA | LITHUANIA | USA	17.30%	62.89%		0.29
	II	ESTONIA | LITHUANIA | FINLAND	16.34%	67.91%	-8.0%	0.35
FILOS-SOCIOL	I	LITHUANIA | POLAND | NETHERLANDS	88.31%	96.10%		0.41
	II	LITHUANIA | POLAND | LATVIA	90.57%	96.60%	-0.5%	0.41
INT J STRATEG PROP M	I	LITHUANIA | FINLAND | ENGLAND	25.71%	58.57%		0.80
	II	LITHUANIA | PEOPLES R CHINA | ENGLAND	24.27%	59.75%	-2.0%	0.86
INZ EKON	I	LITHUANIA | ESTONIA | POLAND	93.03%	97.23%		0.92
	II	LITHUANIA | CZECH REPUBLIC | SPAIN	65.78%	77.47%	20.3%	0.77
J BALT SCI EDUC	I	TURKEY | USA | SLOVAKIA	3.92%	60.10%		0.09
	II	TURKEY | SLOVENIA | FINLAND	2.25%	74.36%	-23.7%	0.43
J BUS ECON MANAG	I	LITHUANIA | TURKEY | ESTONIA	52.07%	65.70%		1.52
	II	LITHUANIA | TURKEY | SPAIN	20.11%	49.84%	24.1%	0.99
LOGOS-VILNIUS	I	LITHUANIA | FRANCE	99.32%	100%		0.14
	II	LITHUANIA | POLAND | FRANCE	99.44%	100%	0.0%	0.35
PROBLEMOS	I	LITHUANIA | BYELARUS | POLAND	92.64%	96.93%		0.52
	II	LITHUANIA | ESTONIA | USA	82.81%	93.75%	3.3%	n.a.
TECHNOL ECON DEV ECO	I	LITHUANIA | POLAND | LATVIA	64.55%	80.43%		1.81
	II	LITHUANIA | PEOPLES R CHINA | POLAND	37.85%	62.22%	22.6%	2.46
TRANSFORM BUS ECON	I	LITHUANIA | POLAND | ROMANIA	42.41%	76.70%		0.51
	II	LITHUANIA | POLAND | ROMANIA	39.89%	79.45%	-3.6%	0.14

Correlation between Impact Factor and Public Availability of Published Research Data in Information Science & Library Science Journals

Rafael Aleixandre-Benavent[1], Luz Moreno-Solano[2], Antonia Ferrer Sapena[3], Enrique Alfonso Sánchez Pérez[4]

[1] *rafael.aleixandre@uv.es*
INGENIO (CSIC-Universitatd Politècnica de València) & UISYS-Universitat de València, Palacio Cerveró - Pza. Cisneros 4, 46003 Valencia (Spain)

[2] *luz.moreno@cchs.csic.es*
IFS, Centro de Ciencias Humanas y Sociales (CCHS). CSIC. Albasanz 26-28. 28037 Madrid (Spain)

[3] *anfersa@upv.es*
DCADHA. Universitat Politècnica de València. Spain. Camino de Vera s/n, 46022 Valencia (Spain)

[4] *easancpe@mat.upv.es*
Instituto Universitario de Matemática Pura y Aplicada. Universitat Politècnica de València. Camino de Vera s/n, 46022 Valencia (Spain)

Introduction

Scientists continuously generate research data but only a few part of them are published. If these data were accessible and reusable, researchers could examine them and generate new knowledge. Currently, the barriers to data sharing are phased out and public research organizations are demanding ever more insistently that publications resulting from publicly funded projects and data that support them should be published in open (Savage & Vickers, 2009). The purpose of this work is: a) to analyse policies concerning open availability of raw research data in journals in the Information Science & Library Science (ISLS); and b) to determine whether there is a correlation between the impact factor and policies of these journals concerning storage and reuse of scientific data.

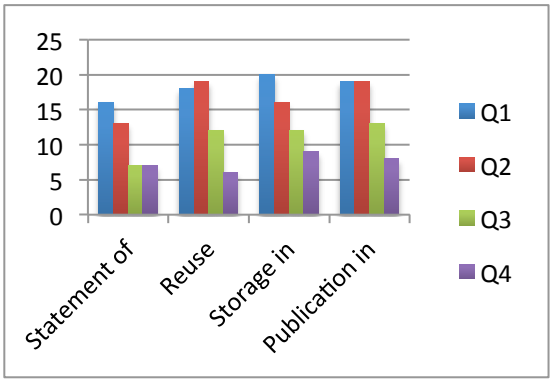
Method

We reviewed the policies related to public availability of papers and data sharing in the 85 journals included in the ISLS category of Journal Citation Reports, 2012 edition. We reported information about the statement of policy regarding: a) complementary material; b) reuse; c) storage in repositories; d) publication on a website; e) journal impact factor; and f) quartile (Q). We have performed a statistical analysis using Chi-square test of the difference regarding each point considered.

Results

The results obtained after analysing the four main variables are presented in Table 1. The variable "Statement of complementary material" was accepted in 50% of the journals. The results were quite similar between the first and second Q and between the third and fourth Q. Regarding the reuse of data, 65% of the journals support this possibility. The highest percentage of response was in the journals of the first Q that accept the reuse of data (86%). The variable "Storage in thematic or institutional repositories", 67% of the journals specified that it was possible. The percentage of journals that accepts storage in institutional repositories decreases by the quartile of journals (e.g., journals in lower quartiles are less supportive). For publication of the manuscript in a website, 69% of the journals accepted it (Figure 1).

Figure 1. Journals supporting each variable by quartile (Q).



Statistical analysis:
Chi-square tests suggest that there is a strong correlation between being a top quartile journal and allowing (a) complementary material (χ^2=11.318, p <.001); (b) reuse of research data (χ^2=19.888, p <.001); (c) storage in thematic and institutional repositories (χ^2=13.080, p <.001); and (d) in personal websites (χ^2=17.350, p <.001).

Conclusions

Our results show that, of the four variables analysed, three have an acceptance rate close to 70% (reuse, publication of the manuscript in a website and storage in thematic or institutional repositories), while the percentage of journals that include the ability to deposit data as supplementary material is lower (50%). These percentages are somewhat higher than those found in a previous study that analysed public availability of published research data in Substance abuse journals (Aleixandre-Benavent et al., 2014). In another study that analysed the same variable in high-impact journals (Alsheikh-Ali et al., 2011), 88% had a statement in their instructions to authors related to public availability and sharing of data, which is 38 percentage points above the average found in the LSIS journals (50%). We found a positive correlation between being a top journal in JCR and having an open policy. A previous paper pointed out that, despite the willingness of some journals to accept supplementary materials, policies, when present, were weak (Borrego & Garcia, 2013). As future research, it would be interesting to raise the question whether journals having high impact factor and open research data is related to the fact that these journals are often owned by rich publishers that are more open for new developments and also have the financial capacities to support such developments.

Acknowledgments

This work has benefited from assistance by the National R+D+I of the Ministry of Economy and Competitiveness of the Spanish Government (CSO2012-39632-C02-01) and Prometeo Program for excellent research groups of Generalitat Valenciana (GVPROMETEO2013-041).

References

Aleixandre-Benavent, R., Vidal-Infer, A., Alonso-Arroyo, A., Valderrama-Zurián, J.C., Bueno-Cañigral, F., & Ferrer-Sapena A. (2014). Public availability of published research data in substance abuse journals. *International Journal of Drug Policy*, *25*, 1143–1146.

Alsheikh-Ali, A.A., Qureshi, W., Al-Mallah, M.H., & Ioannidis, J.P.A. (2011). Public Availability of Published Research Data in High-Impact Journals. *PLoS ONE*, *6*(9): e24357.

Borrego, A., & Garcia, F. (2013). Provision of supplementary materials in library and information science scholarly journals. *Aslib Proceedings*, *65*(5): 503-514.

Savage, C.J., & Vickers, A.J. (2009). Empirical study of data sharing by authors publishing in PLOS journals. *PLoS ONE*, *4*(9), e7078.

Table 1. Results from main variables analysed in the 85 ISLS journals.

Quartile *	Statement of complementary material			Reuse			Storage in thematic or institutional repositories			Publication in website		
	A	NA	NS	A	NA	NS	A	NA	NS	A	NA	NS
	n (%)	n (%)	N (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
1	16 (76%)	-	5 (24%)	18 (86%)	-	3 (14%)	20 (95%)	-	1 (5%)	19 (90%)	-	2 (%)
2	13 (62%)	-	8 (38%)	19 (90%)	1 (5%)	1 (5%)	16 (76%)	-	5 (24%)	19 (90%)	1 (5%)	1 (5%)
3	7 (33%)	2 (10%)	12 (57%)	12 (57%)	3 (14%)	6 (29%)	12 (57%)	-	9 (43%)	13 (61%)	2 (10)	6 (29%)
4	7 (32%)	2 (9%)	13 (59%)	6 (27%)	1 (5%)	15 (68%)	9 (40 %)	1 (5%)	12 (55%)	8 (36%)	1 (5%)	13 (59%)
Total	43 (50%)	4 (5%)	38 (45%)	55(65%)	5 (6%)	25 (29%)	57 (67%)	1 (1%)	27(32%)	59 (69%)	4 (5%)	22 (26%)
	85			85			85			85		

Quartile on ISLS journals in JCR-2012. A: Accepted. NA: Not Accepted. NS: Not Specified

Use of CrossRef and OAI-PMH to Enrich Bibliographical Databases

Mehmet Ali Abdulhayoglu[1] and Bart Thijs[2]

[1]*Mehmetali.abdulhayoglu@kuleuven.be*
Centre For R&D Monitoring (ECOOM), K.U. Leuven, Waaistraat 6, B-3000 Leuven (Belgium)

[2]*Bart.thijs@kuleuven.be*
Centre For R&D Monitoring (ECOOM), K.U. Leuven, Waaistraat 6, B-3000 Leuven (Belgium)

Introduction

Today prominent and comprehensive databases such as Thomson Reuters' Web of Science (WoS) or Elsevier's Scopus are highly in use for bibliometric research. However, these databases do not index full texts hindering researchers to carry out more detailed analyses. Besides, it is possible that some indexed publications do not have DOI numbers playing an important role to access full texts. This paper focuses on how these above-mentioned deficiencies might be overcome by harnessing the Web sources CrossRef and OAI-PMH. Glenisson, Glänzel, Janssens, & De Moor (2005) and Alexandrov, Gelbukh, & Rosso (2005) stated and showed that full text can have an added value in comparison to abstract and title combination when mapping or clustering disciplines and subfields are in question. Therefore, automatic, rapid and free access to full texts of scientific publications might yield a significant contribution to bibliometric research.

Sources

CrossRef

CrossRef provides, besides its other valuable services, a Text and Data Mining (TDM) service enabling researchers to access full-texts of scientific papers for free (Lammey, 2014). This initiative might be a good alternative when considering the policies of the publishers over TDM hindering or retarding the scientific initiatives (Van Noorden, 2012). In this context, by means of a CrossRef REST API, which is free to be used by the public, the developer can access the metadata that CrossRef assembles from more than 4,400 publishers. Besides the metadata such as title, source (e.g. journal, book chapter etc.) name, co-author names, volume year, volume, issue, subject category, two additional important items might be given. These records are *license* and *links* where *link* gives the related full text link and *license* presents an URL link to the license which must be accepted when a GET request is triggered to access the full text. Figure 1 depicts how to access a full text through CrossRef for a given sample digital object identifier (DOI) and a java GET request. In CrossRef's web site, other methods are given to access full text. Since it is not mentioned in the site, we opt to give a java sample through a snippet.

Figure 1. Process of accessing a full text presented by CrossRef by applying *license* and *link* information.

As of 22/12/14, CrossRef has thousands of publications metadata having both full text and license info from the publishers using creative commons license (CC-BY) which encourages the reuse and distribution of content. These publishers are given in Figure 2.

Publisher	CrossRef Number	CrossRef & WoS Number
HINDAWI PUBLISHING CORPORATION	123552	30737
PENSOFT PUBLISHERS	2233	1712
AIP PUBLISHING	273	5
AMERICAN ASSOCIATION OF PHYSICISTS IN MEDICINE (AAPM)	39	11
AMERICAN VACUUM SOCIETY	4	1
ACOUSTICAL SOCIETY OF AMERICA (ASA)	1	0

Figure 2. Number of publications according to publishers using creative commons license (CC-BY) with full text info within CrossRef and within CrossRef-WoS DOI combination.

On the figure's last column, the number of publications, which appear in both CrossRef and WoS, is given for those WoS records only having a DOI. Even though only a few publishers are willing to allow their contents to be mined, we believe that this number will increase over time as also stated by Van Noorden (2014).

Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH)

OAI-PMH emerged aiming at enabling e-print archives to be interoperated (Van de Sompel & Lagoze, 2000). The content of the metadata depends on data provider, for example, while BMC

is providing full texts as well as other metadata, most of data providers such as arXiv do not provide full text or they just mention the URL link not guaranteeing that the full text can be freely downloaded. Below, some example links are given from arXiv and BMC which can be applied to harvest data.

http://www.pubmedcentral.nih.gov/oai/oai.cgi?verb=ListRecords&from=2014-01-01&metadataPrefix=pmc&set=bmcbiology (1)

http://export.arxiv.org/oai2?verb=ListRecords&metadataPrefix=arXiv&set=cs (2)

While former link gives the results only for the journal BMC Biology and those recorded in the repository later than 2014/01/01, later link invokes all the data from computer science discipline in arXiv repository without any date limitation. Note that both results will be invoked in accordance with their own XML schema.

Application

Combining WoS - arXiv - CrossRef

Leveraging arXiv repository, we harvested their OAI-PMH compatible data (See (2)) to combine with our WoS database by matching titles through a character N-Gram text matching process (Abdulhayoglu, Thijs, & Jeuris, 2014). In particular, from arXiv we retrieved title and DOI information for only the *computer science(cs)* discipline to deal with a relatively small data set. There were about 60,000 arXiv records while we have, in WoS, more than 35 million records indexed between 1991 and 2014. We searched for arXiv records within WoS and we found around 18,000 matches having a Salton similarity score higher than 0.90.

Besides 10,000 matches having identical titles, there were more than 7,000 matches having both Salton and Kondrak scores higher than 0.90. Finally, there were only about 200 matches having lower similarity Kondrak scores which can be re-checked manually or simply removed.

We examined the matches having very high similarity scores around 0.90-0.99 and saw that the small character corruptions might appear both on the database or repository side. Additionally, some terms might be given as a text string while it might appear as a symbol in the other source for exp. alpha and α. As a result a similarity score higher than 0.90, especially for Kondrak, can be applied for string matches. So, considering the observations just mentioned, we retained about 6,000 matches having both Salton and Kondrak scores higher than 0.90 and DOI information from the arXiv side.

The retrieved DOI numbers were supposed to be used for accessing full texts through CrossRef. However, a few accessed records have a CC-BY license and we could only grab 286 publications and download their full texts in pdf format. We controlled each full text whether they are correct by checking titles. During this optional process we applied a java pdf parser (*itextpdf*) and correctly extract the title information of those 286 publications. Besides *itextpdf*, CrossRef has its own tool named *pdfextract,* however, it is only applied on Linux environment. Lipinski, Yao, Breitinger, Beel, & Gipp (2013) compare some other extractors.

Conclusions and Discussions

Employing CrossRef and OAI-PMH, a process of accessing full texts of scientific publications indexed in WoS database is explained. Computer science articles from arXiv repository are matched with whole WoS database. Despite a high number of matches, the number of publications appearing within CrossRef repository having creative commons license is quite low. Though a small number of publications has creative commons license, CrossRef seems to ease the issue of accessing full texts freely in time (Van Noorden, 2014).

Acknowledgments

Authors would like to thank Rachael Lammey and Karl Ward from CrossRef, Meshna Koren from Elsevier, Mikail Shaikh from Springer and IT admins from arXiv for their valuable guidance and helps for their TDM systems.

References

Abdulhayoglu, M. A., Thijs, B., & Jeuris, W. (2014). Matching bibliographic data from publication lists with large databases using N-Grams. *Available at SSRN 2464065*.

Alexandrov, M., Gelbukh, A., & Rosso, P. (2005). An approach to clustering abstracts. In *Natural Language Processing and Information Systems* (pp. 275-285). Berlin: Springer.

Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, *41*(6), 1548-1572.

Lammey, R. (2014). CrossRef's Text and Data Mining Services. *Learned Publishing*, 27(4), 245-250.

Lipinski, M., Yao, K., Breitinger, C., Beel, J., & Gipp, B. (2013, July). Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In *Proc. 13th ACM/IEEE-CS Joint Conf. on Digital Libraries* (pp. 385-386). ACM.

Van de Sompel, H., & Lagoze, C. (2000). The Santa Fe convention of the open archives initiative. *D-Lib Magazine*, (2), 2011-10.

Van Noorden, R. (2014). Elsevier opens its papers to text-mining. *Nature*, *506*(7486), 17-17.

Van Noorden, R. (2012). Trouble at the text mine. *Nature*, *483*(7388), 134-135.

Does Scopus Put its Own Journal Selection Criteria into Practice?

Zehra Taşkın[1], Güleda Doğan[1], Sümeyye Akça[1], İpek Şencan[1], and Müge Akbulut[2]

[1] *ztaskin@hacettepe.edu.tr, gduzyol@hacettepe.edu.tr, sumeyyeakca@hacettepe.edu.tr, ipeksencan@hacettepe.edu.tr*
Hacettepe University, Department of Information Management, Ankara (Turkey)

[2] *mugeakbulut@gmail.com*
Yıldırım Beyazıt University, Department of Information Management, Ankara (Turkey)

Introduction

Scopus has been one of the main abstract and citation databases introduced by Elsevier in 2004 to the scientific area. With the multidisciplinarity and international coverage aspects, it is one of the largest databases of peer-reviewed literature in the fields of science, technology, medicine, social sciences, arts, and humanities. There have been several literature studies assessing different aspects of Scopus since the very beginning. The following consists mainly of a description of Scopus, comparing it with the other databases, from the point of usability and accessibility, evaluations regarding the number of citations, and so on. Although there have been many studies about content evaluation and comparisons with other databases, to our knowledge no study has been published focusing on the journal selection criteria of Scopus. The main goal of this study is to evaluate Scopus journals and draw a picture regarding the quality of the journals indexed in Scopus. The two research questions of this study are:

- Do the journals indexed in Scopus match with the Scopus indexing criteria?
- Is there any contribution of the journals that does not fulfil the criteria of Scopus with respect to diversity of authors, institutions and countries as well as internationality of referees, editors and authors?

Methodology

The universe of the study consists of the 2013 Scopus journal list downloaded from SCImago Journal Rank (SJR) on September 18th, 2014. Two groups of countries that have more than 1,000 journals and less than 100 journals in Scopus were left out of the content of this study because of their projected effects on the sample. As a result, 6,151 journals from 23 countries constituting the sample frame were sampled with the systematic sampling method with a rate of 1:30 and 203 journals were chosen for the sample in proportion to 23 countries' journal counts in Scopus.

These 203 journals were evaluated according to the criteria outlined in Table 1, which is mainly based on Scopus journal selection criteria.[1] The contextual criteria were removed because of the requirement to have a comprehensive knowledge of related field. Furthermore, revised Scopus criteria and some new added criteria are marked with grey in Table 1.

Table 1. Criteria selected and used to evaluate Scopus journal content.

Criteria categories	Criteria
Minimum technical criteria (Pre-selection conditions)	Peer-review content and have a publicly available description of the peer review process
	Have an International electronic Standard Serial Number (eISSN) as registered with the ISSN International Centre
	English abstracts and titles
	Regular publication
	References in Roman script
	Publicly available publication ethics and malpractice statement
Journal policy	Editorial policy available
	Type of peer review
	Reviewer list available online
	Diversity in geographical distribution of editors
	Volume of editorial board
	Diversity in geographical distribution of authors
Journal standing	Citedness of journal articles in Scopus
Publishing regularity	No delays or interruption in the publication schedule
Online availability	Full journal content available online
	Journal website available
	English language journal website available
General information about journal	Country of the journal
	Number of issues per year
	First publishing year of the journal
	Journal back issues available on the journal website

Findings and Results

There are only 13 journals providing all of the minimum technical criteria of Scopus. The majority of the journals (190) did not meet at least one criterion. Six journals fulfilled only one criterion of Scopus. Journals and their fulfilment of evaluation criteria are shown in Figure 1. The baseline of the radar graphic (Fig. 1) was created by using "yes"

[1] http://www.elsevier.com/online-tools/scopus/content-overview#content-policy-and-selection

answers to the criteria. We found that 32% of journals did not have an International Electronic Standard Serial Number available (eISSN). Most of the journals (82% and 69% respectively) did not match the criteria of reviewers list being available online and having publicly available publication ethics and malpractice statement. Journals were successful about applying the criteria of available references in Roman script, regular publication and English abstracts and titles.

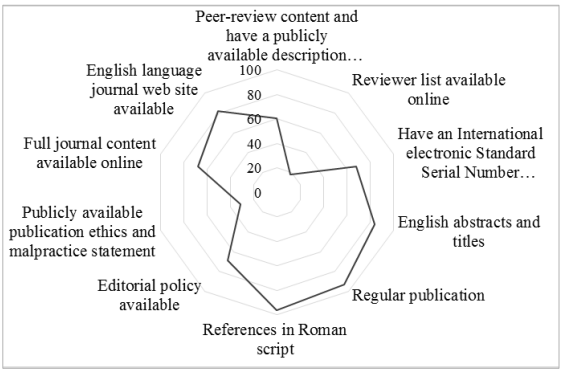


Figure 1. Radar graphic presentation of journals' fulfilment of evaluation criteria.

The evaluation criteria were divided into five classes in this study. These classes are accessibility, peer-review process, policy issues, internationalization and citation levels of journals. The detailed evaluation of each criterion is found in the following sections of this study.

We decided that accessibility on the web, regular publication and references in Roman script consist of the main components of the accessibility criteria in our study. Fifty-one percent of journals in our sample have had all the issues since the launch of their websites and had websites that included full contents of the issues (titles, abstracts, full texts, etc.). Almost all journals had references in Roman script (97%) and most of the journals had English titles/abstracts (84%) and English websites (82%).

The criteria of peer-review process consists of a journal having detailed information about how it is managed and its peer-review board list being available online. We found that 40% of the journals did not have any information on their websites about the peer-review process. Those that did, 73% did not have any information about how their peer-review processes were managed (e.g., double blind, single blind and so on). Only 18% of journals published a list of their reviewers. Under these circumstances, it was hard to determine the diversity of reviewers.

Having accessible publication policies and publicly available publication ethics and malpractice statements were regarded as policy issues. We found that 32% of the journals did not have any editorial policy on their websites. In addition, 68% of the journals did not have any publicly available publication ethics and malpractice statements. Because policy issues were parts of Scopus's minimum criteria, it was expected that journals without these policies would not have passed the preliminary evaluation. However, all these journals have been indexed in Scopus over the years.

The diversity of authors and the editorial board were important for Scopus' evaluation team. We evaluated the diversities as part of this study. Twenty-nine percent of the journals did not have a list of editorial board on their websites. The median for geographic diversity of editors was about 6 within the rest of journals. Eight journals had editors from more than 20 countries. A journal had editors from 53 different countries, while 21% had editors from only one country.

Author diversity is also important for internationalization of journals. We calculated the number of countries by using author affiliations of the last 10 published articles/reviews of each journal. Nine journals did not give any country information for their authors. The median for geographic diversity of authors was 4 within the rest of the journals. Authors were from only one country in 26% of the journals.

Citations are essential for indexed journals within citation databases, as almost all the performance evaluations rely on citations. We evaluated the citation levels of journals by using total cites (three years) indicator of SCImago database. The median number of citations was calculated as 26. Fourteen journals did not have any citations during the three-year period. Six journals had over 1,000 citations.

Conclusions

Citation databases are important for authors, decision-makers, institutions, countries and others. Therefore, it is vital to index high-quality journals for them. Citation databases have strict selection criteria to evaluate journals before indexing to achieve their aims. The criteria of databases are generally based on journal policy, regularity of publication, diversity and so on. We evaluated the journal selection criteria of Scopus and checked the extent of their implementation within this study.

According to the results of our study, the publishers, editors and Scopus should strive to enhance quality. On Scopus' side, Scopus must put the selection criteria into practice strictly and control indexed journals on the aspects of these criteria. Because of the huge competitive environment in the journal market recently, Scopus as well as other publishers of commercial citation databases should consider quality issues more importantly than commercial concerns. A comparative study on journal selection of citation databases may be the continuation of this study.

On the Correction of "Old" Omitted Citations by Bibliometric Databases

Fiorenzo Franceschini[1], Domenico Maisano[2] and Luca Mastrogiacomo[3]

[1] *fiorenzo.franceschini@polito.it,* [2]*domenico.maisano@polito.it,* [3]*luca.mastrogiacomo@polito.it*
Politecnico di Torino, DIGEP (Department of Management and Production Engineering),
Corso Duca degli Abruzzi 24, 10129, Torino (Italy)

Abstract
Omitted citations – i.e., missing links between a cited paper and the corresponding citing papers – are the main consequence of several bibliometric-database errors. To reduce these errors, databases may undertake two actions: (i) improving the control of the (new) papers to be indexed, i.e., limiting the introduction of "new" dirty data, and (ii) detecting and correcting errors in the papers already indexed by the database, i.e., cleaning "old" dirty data. The latter action is probably more complicated, as it requires the application of suitable error-detection procedures to a huge amount of data. Based on an extensive sample of scientific papers in the Engineering-Manufacturing field, this study focuses on old dirty data in the Scopus and WoS databases. To this purpose, a recent automated algorithm for estimating the omitted-citation rate of databases is applied to the same sample of papers, but in three different-time sessions. A database's ability to clean the old dirty data is evaluated considering the variations in the omitted-citation rate from session to session. The major outcomes of this study are that: (i) both databases slowly correct old omitted citations, and (ii) a small portion of initially corrected citations can surprisingly come off from databases over time.

Conference Topic
Data Accuracy and disambiguation

Introduction
An important branch of the bibliometric literature examines errors in bibliometric databases. Several studies show that the major consequence of database errors is represented by omitted citations, i.e., citations that should be ascribed to a certain (cited) paper but, for some reason, are lost (Moed, 2005; Buchanan, 2006; Jacsó, 2006, Li et al., 2010; Olensky, 2013).

Franceschini et al. (2013) proposed an automated algorithm for estimating the omitted-citation rate of bibliometric databases. This algorithm requires the combined use of two or more bibliometric databases and is based upon the hypothesis that the mismatch between the citations occurring in one database and another one is evidence of possible errors/omissions.

In a further study by Franceschini et al. (2014), this algorithm was applied to a relatively large set of publications, showing that, depending on the bibliometric database in use (Scopus or WoS), omitted citations are not distributed uniformly among publishers; e.g., regarding the publications in the Engineering-Manufacturing field, citations from papers published by Wiley-Blackwell are more likely to be omitted by Scopus, while those from papers published by ASME (American Society of Mechanical Engineers) are more likely to be omitted by WoS. A reason behind this result is that some editorial styles imposed by certain publishers can probably hamper the correct identification of the cited papers by some databases.

The presence of database errors, as well as journal coverage or author disambiguation, is probably one of the major concerns of database administrators. In the authors' opinion, database administrators may undertake two actions for reducing database errors:

1. Limiting the introduction of "new" dirty data in a database, i.e., errors concerning new papers to be indexed;
2. Cleaning "old" dirty data, i.e., errors concerning papers/journals already indexed by a database.

The recent effort by reviewers, publishers and database administrators in checking the cited article lists of new papers probably contributes to reducing "new" dirty data. This hypothesis is corroborated by a recent study by Franceschini et al. (2015), which shows that the databases' propensity to omit newer citations is generally lower than that to omit older citations.

Cleaning up old dirty data is certainly much more complicated because it requires the systematic application of suitable error-detection procedures to a huge amount of data. However, this effort would be essential for improving the quality of a database significantly.

This paper focuses on the ability of the major multidisciplinary bibliometric databases, i.e., Scopus and WoS, to clean up old dirty data. For this evaluation, we use a new procedure, derived from the automated algorithm by Franceschini et al. (2013). This procedure consists in (i) repeating the omitted-citation-rate analysis on the same sample of (cited and citing) articles, but in different-time sessions, and (ii) observing any variation in the results. A database's ability to clean old dirty data will be evaluated considering the variation in the omitted-citation rate from one session to another one.

The remainder of this paper is organized into four sections. The section "Automated algorithm for examining the omitted citations" briefly recalls the algorithm by Franceschini et al. (2013). The section "Methodology" describes the methodology used in our study, focusing on data collection and analysis. The section "Results" illustrates the results of the analysis, investigating similarities and differences between the two databases examined. Finally, the section "Conclusions" summarizes the original contributions of this paper, highlighting the major results, limitations and suggestions for future research.

Automated algorithm for analysing the omitted citations

Before recalling the algorithm, we present an introductory example to illustrate how it works. Let us consider a fictitious paper of interest, indexed by Scopus and WoS. The number of citations received by this paper is four in Scopus and six in WoS (see Table 1).

Table 1. Citation data relating to a fictitious article, according to Scopus and WoS. The union of the citations recorded by the two databases (see the first column) is a total of eight citations. Among the citations, only five come from sources officially covered by both databases (highlighted in grey).

Citation No.	Scopus	WoS
1	✓	
2		✓
3	Omitted	✓
4	✓	✓
5	✓	✓
6	Omitted	✓
7		✓
8	✓	Omitted
Total	4	6

The union of the citations recorded by the two databases is a total of eight citations. Among the citations, only five come from sources (i.e., journals or conference proceedings) officially covered by both databases (highlighted in grey in Table 1). Focusing on these five "theoretically overlapping" (TO) citations, two are omitted by Scopus (but not by WoS) and one is omitted by WoS (but not by Scopus). Therefore, from the perspective of the paper of interest, a rough estimate of the omitted-citation rate is $2/5 \approx 40\%$ in Scopus and $1/5 \approx 10\%$ in WoS. The same reasoning can be extended to multiple papers of interest and more than two bibliometric databases.

The automated algorithm, which is based on the combined use of two bibliometric databases (Scopus and WoS in this case), can be summarised in three steps:

1. Identify a set of (P) papers of interest, indexed by both the databases.
2. For each (i-th) paper of the set, identify the TO citations, defined as the portion of documents issued by journals officially covered by Scopus and WoS. The number of TO citations concerning the i-th paper of interest will be denoted as γ_i.
3. For each (i-th) paper of the set and for each database, determine the number (ω_i) of TO citations that do not occur in it and classify them as omitted citations. The omitted-citation rate (p) relating to the P papers of interest, according to a database, can be estimated as:

$$\hat{p} = \sum_{i=1}^{P} \omega_i / \sum_{i=1}^{P} \gamma_i \,. \qquad (1)$$

We emphasize that p is estimated on the basis of (i) a set of papers of interests and (ii) a portion of the total citations that they obtained (i.e., that ones related to citing articles purportedly covered by both the databases). For a more detailed description of the algorithm, we refer the reader to Franceschini et al. (2013).

The ability of bibliometric databases to clean old dirty data will be evaluated by applying this algorithm to the same sample of TO citations, in three different-time sessions.

Methodology

The study is based on the analysis of the citations obtained from a relatively large sample of papers of interest. The papers were issued by 33 scientific journals (i) included in the ISI Subject Category of Engineering-Manufacturing (by WoS) and (ii) covered by Scopus; Table 2 reports the list of these journals. For each journal, we considered the papers published in the time-window from 2006 to 2012 and the citations that they obtained from papers issued in the same period.

Data collection was repeated in three different-time sessions, spaced about seven months apart: i.e., session I on August 2013, session II on March 2014 and session III on September 2014. We remark that the duration of each data-collection session (i.e., a few days) is negligible with respect to the time period between two consecutive sessions.

To enable comparisons between data collected in different sessions, we adopted two measures:

1. Among the papers of interest (or cited papers) – i.e., those issued by the 33 Engineering-Manufacturing journals – we selected those indexed in each of the three sessions, by both the (Scopus and WoS) databases; in formal terms:

$$A = A^{(I)} \cap A^{(II)} \cap A^{(III)}, \qquad (2)$$

   A being the set of cited papers selected for our analysis and $A^{(I)}$, $A^{(II)}$ and $A^{(III)}$ the sets of papers indexed by both the databases, at the moment of session I, II and III respectively.

   Also, we excluded articles without DOI code or whose DOI code is not indexed by both databases, as they would be difficult to disambiguate.
2. Among the citations, we selected the so-called TO citations, i.e., those obtained from journals purportedly covered by both databases and issued in the 2006-to-2012 time-window. To avoid any misunderstanding, we excluded citations from journals covered in the 2006-to-2012 time-window, but later banned from the database[1]. The official lists of documents covered by the databases in use – which are essential for determining the TO

[1] A possible misunderstanding arises from the fact that, in some cases (mostly on Scopus), the expulsion of a journal from a database entails the entire removal of previously indexed papers, while in other cases (mostly on WoS), previously indexed papers are not necessarily removed.

citations – were retrieved from the databases' websites (Scopus Elsevier, 2015; Thomson Reuters, 2015).

Table 2. List of the Engineering-Manufacturing journals examined. For each journal, it is reported its title and ISSN code. Journals are sorted alphabetically according to their title

Journal title	ISSN
AI EDAM - Artificial Intelligence for Engineering Design Analysis and Manufacturing	0890-0604
Assembly Automation	0144-5154
CIRP Annals - Manufacturing Technology	0007-8506
Composites Part A - Applied Science and Manufacturing	1359-835X
Concurrent Engineering - Research and Applications	1063-293X
Design Studies	0142-694X
Flexible Services and Manufacturing Journal	1936-6582
Human Factors and Ergonomics in Manufacturing & Service Industries	1090-8471
IEEE Transaction on Components Packaging and Manufacturing Technology	2156-3950
IEEE Transactions on Semiconductor Manufacturing	0894-6507
IEEE-ASME Transactions on Mechatronics	1083-4435
International Journal of Advanced Manufacturing Technology	0268-3768
International Journal of Computer Integrated Manufacturing	0951-192X
International Journal of Crashworthiness	1358-8265
International Journal of Machine Tools & Manufacture	0890-6955
International Journal of Production Economics	0925-5273
Journal of Advances Mechanical Design Systems and Manufacturing	1881-3054
Journal of Computing and Information Science in Engineering - Transactions of the ASME	1530-9827
Journal of Intelligent Manufacturing	0956-5515
Journal of Manufacturing Science and Engineering - Transactions of the ASME	1087-1357
Journal of Manufacturing Systems	0278-6125
Journal of Materials Processing Technology	0924-0136
Journal of Scheduling	1094-6136
Machining Science and Technology	1091-0344
Materials and Manufacturing Processes	1042-6914
Proceedings of the Institution of Mechanical Engineers Part B - Journal of Engineering Manufacture	0954-4054
Packaging Technology and Science	0894-3214
Precision Engineering - Journal of the International Societies for Precision Engineering and Nanotechnology	0141-6359
Production and Operations Management	1059-1478
Production Planning & Control	0953-7287
Research in Engineering Design	0934-9839
Robotics and Computer-Integrated Manufacturing	0736-5845
Soldering & Surface Mount Technology	0954-0911

The sample of TO citations used in the analysis is the union of the TO citations (that meet the above requirements), collected in each of the three sessions. In formal terms, this sample of TO citations is:

$$B = B^{(I)} \cup B^{(II)} \cup B^{(III)}, \quad (3)$$

$B^{(I)}$, $B^{(II)}$ and $B^{(III)}$ being the TO citations collected during session I, II and III respectively.
This sample of TO citations will be used for estimating the omitted-citations rate of a certain database, in a certain session; the relationship in Eq. 1 can be used, being:

- $\hat{p}$ the estimate of the omitted-citation rate related to a certain session and a specific database;
- P the number of (cited) articles of interest;
- γ_i the number of TO citations relating to the i-th of the P articles of interest;
- ω_i the portion of the TO citations, collected in a certain session, which are omitted by a specific database.

Being $\hat{p}$ just an estimate of p – albeit the best possible – a relevant symmetrical $(1 - \alpha)$ confidence interval (*CI*) can be constructed as[2]:

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{\sum_{i=1}^{P} \gamma_i}}, \tag{4}$$

with:

α, the type-I error;

$z_{1-\alpha/2}$ the unit normal deviate corresponding to $1 - \alpha/2$.

In this case, we consider a symmetrical 95% *CI*, therefore $\alpha = 5\%$ and $z_{97.5\%} \approx 2$.

By adopting this procedure, we will obtain six different estimates of the omitted-citation rate, i.e., one for each of the three sessions and each of the two databases in use. The comparison of these estimates will tell us whether the databases examined are able to correct old omitted citations.

Results

The total number of papers of interest, i.e., those issued by the Engineering-Manufacturing journals examined, is $P = 23{,}806$. The corresponding TO citations are $\Sigma\gamma_i = 97{,}698$. Table 3 contains the $\hat{p}$ values and the relevant 95% *CI*s, relating to the three sessions and the two databases examined.

Table 3. Main results of the (repeated) analysis of the omitted-citation rate of databases. Citing and cited articles were issued from 2006 to 2012. Statistics concern each of the three sessions (i.e., session I, II and III) for Scopus and WoS respectively.

		(a) Scopus				(b) Wos			
Session	$\sum_{i=1}^{P} \gamma_i$	$\sum_{i=1}^{P} \omega_i$	$\hat{p}$	95% *CI*		$\sum_{i=1}^{P} \omega_i$	$\hat{p}$	95% *CI*	
I (August 2013)	97,698	5,183	5.3%	5.2%	5.4%	7,370	7.5%	7.4%	7.7%
II (March 2014)	97,698	4,607	4.7%	4.6%	4.8%	6,376	6.5%	6.4%	6.7%
III (October 2014)	97,698	4,473	4.6%	4.4%	4.7%	6,404	6.6%	6.4%	6.7%

$P = 97{,}698$ is the total number of (cited) articles, published by 33 Engineering-Manufacturing journals;

$\sum \gamma_i$ is the total number of TO citations (which is independent on the session);

$\sum \omega_i$ is the total number of omitted citations relating to each session and each database;

$\hat{p}$ is the estimate of the omitted-citation rate relating to each session and each database;

The 95% *CI* around $\hat{p}$ is obtained applying the approximated relationship in Eq. 4.

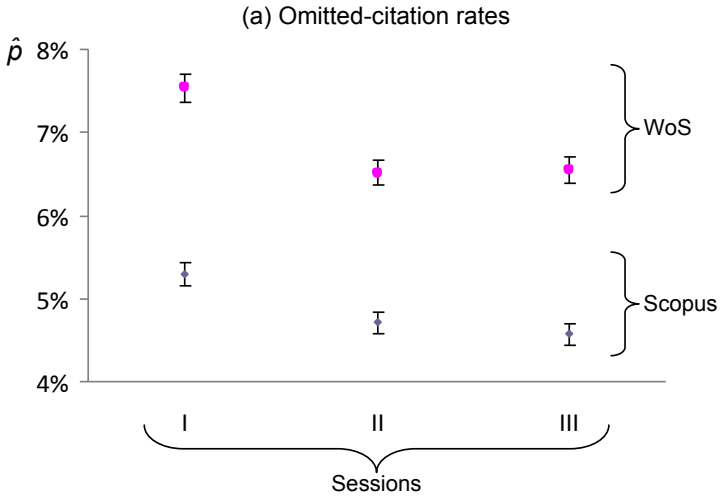
[2] The *CI* construction in Eq. 4 is grounded on the following considerations:

- For a generic sample consisting of $n = \Sigma\gamma_i$ TO citations, the number of omitted citations will be a binomially distributed variable with mean value $n \cdot p$ and variance $n \cdot p \cdot (1 - p)$;
- The aforesaid binomial distribution can be approximated by a normal distribution with the same mean value and variance. This approximation is acceptable in the case $n \cdot p \geq 5$ (Ross, 2009), which is generally satisfied when considering relatively large sets of TO citations.
- Based on the previous approximation, the percentage of omitted citations for a sample of n TO citations will be a normally distributed variable with mean value p and variance $p \cdot (1 - p)/n$. Since p is not known, it can be replaced by its best estimate $\hat{p}$.

In conclusion, Eq. 4 defines a symmetric *CI* around $\hat{p}$, which – with a probability $(1 - \alpha)$ – will include the "true" p value.

The $\hat{p}$ values of both databases tend to decrease over time, denoting that dirty data have been partially cleaned. Interestingly, the major reduction in the $\hat{p}$ values is between the session I and II for both databases; on the other hand, variations between session II and III are not significant, since the 95% *CI*s are partially overlapped (see Figure 1(a)); as regards WoS, we can even notice an imperceptible increase in the $\hat{p}$ value between session II and III.

The overall reduction in the number of omitted TO citations ($\Sigma\omega_i$) for WoS is greater than that for Scopus (i.e., 7,370 – 6,404 = 966 against 5,183 – 4,473 = 710); however, consistently with what observed in other studies (Franceschini et al., 2014; 2015), we note that the omitted-citation rates in Scopus are generally lower than those in WoS. Figure 1(b) shows that the overall percent variations in the $\hat{p}$ values between session I and III are very similar (i.e., -13.7% and -13.1%, for Scopus and WoS respectively).



(b) Relevant percent variations

$$\Delta p = \left(\frac{\hat{p}_{final} - \hat{p}_{initial}}{\hat{p}_{initial}} \right)$$

	Δp		
	I to II	II to III	I to III
Scopus	-11.1%	-2.9%	-13.7%
WoS	-13.5%	0.4%	-13.1%

Figure 1. (a) Graphical representation of the omitted-citation rate in the three sessions, for Scopus and WoS, and (b) relevant percent variations.

Having verified that both databases tend to slowly correct old omitted citations, we now investigate the possible differences in the indexing of individual TO citations, from one session to another one. Table 4 summarizes the eight possible events concerning the correct/missing indexing of individual TO citations. Since there are two possible indexing states (i.e., correct or missing indexing) for each of the three sessions, the total number of possible events is $2^3 = 8$; the file containing the complete list of individual TO citations, with the relevant cited papers, and their session-by-session indexing by the databases, is available under request to authors.

Not surprisingly, the most frequent events are those with no variation (i.e., the type 1 and 2 events in Table 4), in which the TO citations are indexed correctly ("✓") or incorrectly ("×") in all the three sessions; the portion of TO citations with no variation is 98.7% for Scopus and 98.5% for WoS). The type 3 and 4 events represent corrections in the TO-citation indexing, in session II and III respectively. The total number of corrections in WoS is basically larger to that in Scopus, probably due to the larger level of "initial dirt" in the former database, compared to that one in the latter. Moreover, we note that almost all of the corrections by WoS are concentrated in session II (i.e., 1193 out of 1215).

Despite these differences, the percentage of TO citations corrected by Scopus and WoS are pretty close to each other (i.e., roughly 1% and 1.2% respectively). This similarity is even more interesting if we consider the fact that, among the set of corrected TO citations, a relatively small subset is shared between the two databases (i.e., 392 citations out of (997 + 1,215 – 392) = 1,820, corresponding to about 21.5% of the set of corrected TO citations).

Table 4. Overall statistics concerning the indexing of the individual TO citations, in each session. Symbols "✓" and "×" respectively identify the TO citations correctly indexed or omitted in a certain session.

Type of event		Session			(a) Scopus				(b) Wos			
					Single event		Aggregated events		Single event		Aggregated events	
		I	II	III	TO citations	Percent	TO citations	Percent	TO citations	Percent	TO citations	Percent
No variation	1	✓	✓	✓	92,296	94.5%	96,411	98.7%	90,195	92.3%	96,214	98.5%
	2	×	×	×	4,115	4.2%			6,019	6.2%		
Correction	3	×	✓	✓	765	0.8%	997	1.0%	1,193	1.2%	1,215	1.2%
	4	×	×	✓	232	0.2%			22	0.0%		
Anomalous variation	5	✓	×	×	102	0.1%	290	0.3%	164	0.2%	269	0.3%
	6	✓	✓	×	112	0.1%			77	0.1%		
	7	×	✓	×	0	0.0%			0	0.0%		
	8	✓	×	✓	76	0.1%			28	0.0%		
				Total	97,698	100%	97,698	100%	97,698	100%	97,698	100%

The type 5 to 8 events are characterized by anomalous variations, in which some TO citations, which are correctly indexed in a certain session, are omitted in one (or more) subsequent sessions. It is surprising how citations, which were initially indexed correctly, can come off from a database over time; in other words, these events represent a form of generation of dirty data, which is independent of the introduction of new data in the database. Fortunately, the incidence of these abnormalities is rather low (coincidentally, about 0.3% for both Scopus and for WoS); in the future, we may conduct a thorough analysis of these anomalies, based on their manual examination.

Conclusions

The analysis presented in this paper shows that the two bibliometric database examined tend to gradually reduce the number of old omitted citations, although this reduction is relatively slow for both. It would be interesting to see to what extent these cleanings were due to error-correction campaigns structured by database administrators, or simply due to impromptu database-inaccuracy reports by authors and/or database users (even checking and cleaning up bibliometric data in personal research profiles, such as ResearcherID, Scopus Author ID, ORCID, etc.).

Results of this study show other interesting similarities/coincidences between the two databases examined:

1. Comparing the results related to session I and III (spaced about fourteen months apart), we noticed a 13-to-14% reduction in the *p* values for both Scopus and WoS.
2. For both databases, the greatest reduction in the omitted-citations rate was registered in session II and not in session III. This could be just a coincidence or it could denote a sort of "seasonality" of the two databases in cleaning up old dirty data.
3. The portion of TO citations whose indexing varies in the three sessions is roughly the same for both databases, i.e., roughly 1 to 1.5%. Apart from the previously omitted TO citations that have been justly corrected, they include a small portion of abnormal variations, i.e., TO citations correctly indexed in some session and subsequently omitted. Coincidentally, the percentage of abnormal variations is 0.3% for both databases.

The proposed analysis has several limitations. Even though the set of TO citations includes almost one-hundred thousand citations, the relevant cited papers are all confined within the Engineering-Manufacturing field. Also, the analysis was repeated in three sessions over a

total period of about 14 months; therefore, it reflects a database's ability to correct errors in short/middle-term period, but not in the long-term period.

In the future, we plan to extend the study to a longer time-scale (e.g., 2 or 3 years) and/or to scientific articles in other disciplines. Furthermore, the study will be expanded for investigating possible links between the omitted citations' propensity to be corrected and the publishers of the relevant citing papers.

References

Buchanan, R.A. (2006). Accuracy of Cited References: The Role of Citation Databases. *College & Research Libraries*, 67(4), 292-303.

Franceschini, F., Maisano & D., Mastrogiacomo, L. (2013). A novel approach for estimating the omitted-citation rate of bibliometric databases. *Journal of the American Society for Information Science and Technology*, 64(10), 2149-2156.

Franceschini, F., Maisano, & D., Mastrogiacomo, L. (2014). Scientific journal publishers and omitted citations in bibliometric databases: Any relationship? *Journal of Informetrics*, 8(3), 751-765.

Franceschini, F., Maisano, & D., Mastrogiacomo, L. (2015). Influence of omitted citations on the bibliometric statistics of the major Manufacturing journals. To appear in *Scientometrics*. A draft version is available at http://staff.polito.it/fiorenzo.franceschini/Pubblicazioni/Revised_IJPE-D-13-01272.pdf.

Jacsó, P. (2006). Deflated, inflated and phantom citation counts. *Online Information Review*, 30(3), 297-309.

Li, J., Burnham, J.F., Lemley, T., & Britton, R.M. (2010). Citation analysis: comparison of Web of Science, Scopus, Scifinder, and Google Scholar. *Journal of Electronic Resources in Medical Libraries* 7(3), 196-217.

Moed, H.F. (2006). *Citation analysis in research evaluation* (Vol. 9). Springer.

Olensky, M. (2013). Accuracy Assessment for Bibliographic Data. *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI)*, vol. 2, pp. 1850-1851, Vienna, Austria.

Ross, S.M. (2009). *Introduction to probability and statistics for engineers and scientists*. Academic Press.

Schenker, N., & Gentleman, J.F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55(3), 182-186.

Scopus Elsevier (2015). *Scopus Content Coverage*. Available at http://www.scopus.com [retrieved on August 2013, March 2014 and October 2014].

Thomson Reuters (2015). *Master Journal List*, http://ip-science.thomsonreuters.com/mjl/ [retrieved on August 2013, March 2014 and October 2014].

Can We Track the Geography of Surnames Based on Bibliographic Data?

Nicolas Robinson-Garcia[1], Ed Noyons [2] and Rodrigo Costas [2]

[1] *elrobin@ugr.es*
EC3Metrics spin-off and EC3 Research Group, Universidad de Granada, Granada (Spain)

[2]*noyons@cwts.leidenuniv.nl* [3] *rcostas@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (The Netherlands)

Abstract
In this paper we explore the possibility of using bibliographic databases for tracking the geographic origin of surnames. Surnames are used as a proxy to determine the ethnic, genetic or geographic origin of individuals in many fields such as Genetics or Demography; however they could also be used for bibliometric purposes such as the analysis of scientific migration flows. Here we present two relevant methodologies for determining the most probable country to which a surname could be assigned. The first methodology assigns surnames based on the most common country that can be assigned to a surname and the Kullback-Leibler divergence measure. The second method uses the Gini Index to evaluate the assignment of surnames to countries. We test both methodologies with control groups and conclude that, despite needing further analysis on its validity; these methodologies already show promising results.

Conference Topic
Data Accuracy and disambiguation

Introduction
Tracking the geographical origin of individuals has multiple applications and is of interest to many fields. For instance, in biomedical research it is used for racial and ethnic classification as this information is useful for identifying risk factors in epidemiological and clinical research (Burchard et al., 2003). It is also of interest in the field of Demography to analyse migration movements (e.g. Chen & Cavalli-Sforza, 1983) or migratory influences in a given country (Hatton & Wheatley Price, 1999). In the field of bibliometrics, scientific migration flows between countries has been a subject of study as they are considered beneficial for the exchange of new ideas and scientific knowledge between countries (Moed & Halevi, 2014) as well as to analyse case studies to identify the spread of researchers of a given nationality around the world (Costas & Noyons, 2013).

Surnames have been used as a proxy of geographic, ethnic and even genetic origin for some time now. According to Kissin (2011) "the use of surnames in human population biology dates back to 1875, when George Darwin used frequency of occurrences of the same surname in married couples to study in-breeding". Geographic information related to surnames may also be of use in the field of bibliometrics, especially with regard to collaboration and mobility studies. So far only few papers have been found using surname data for bibliometric purposes. Kissin and colleagues (Kissin & Bradley, 2013; Kissin, 2011) have performed several studies focused on the analysis of Jewish surnames in the database MEDLINE. Also Freeman and Huan (2014) recently analysed the effect of diversity of authorship in the impact of scientific publications.

Until recently, these studies relied on manually curated lists of surnames related to ethnic groups, languages or countries. In the last few years, surname research has been developed and many methodologies have been proposed to discern statistical approaches to geographically classify surnames (a good review on the subject can be found in Cheshire, 2014). In this regard, two types of approaches can be found: 1) probability and Bayesian

methods and 2) clustering techniques. For this, we can focus either on the concentration of surnames by areas or on tracking surnames to their original region (Cheshire, 2014).

So far the results reported are quite satisfactory (Mateos, 2007). While regional studies with large data sets offer relatively accurate results due to the skewness of the surnames distribution (Cheshire, 2014), there are still problems when applying these methodologies at a global level. Such limitations are due to migratory movements and data restrictions. For instance, the surname 'Lee' is considered in many studies as British. However, it is most common in the United States and at the same time in Asia. Also data availability may be an issue as most of it comes from census data and demography studies which usually come from different sources and present differences between them.

In this paper we suggest the use of a single data source to develop a methodology to track the geography of surnames worldwide. We propose using the authors' affiliation data from a scientific bibliographic database. For this purpose we analyse two different useful methodologies: one based on the application of information theoretic measures, and a second one based on the use of inequality indexes.

This paper is structured as follows. First we describe the data collection and processing. Then we describe each of the two methodologies proposed for assigning countries to names: one based on the Kullback-Leibler divergence (Kullback & Leibler, 1951) and a second one using the Gini Index, usually used in the field of Economics. In order to test the validity of each methodology, we compared our results with those from a list of surnames based on language origin for 11 different languages. Finally we conclude discussing the limitations of our methodologies, further developments and the potential use of this type of studies for the field of bibliometrics.

Data collection and processing

The goal of this paper is to develop a methodology to assign surnames to countries based on the bibliographic data offered by authors from a scientific database. For this we used the in-house CWTS version of the Web of Science database (not including the Conference Proceedings Citation Index or the Book Citation Index). This database covers all publications and authors for the 1980-2013 time period. The next step needed was to identify authors and relate them with their country of origin. Such approach assumes certain limitations:

- *Reliance on a single data source*. This means that errors or misrepresentations by countries derived from the Web of Science database will reflect on the quality of the result findings reported. Also, the surname information is restricted to the time period employed in the analysis, meaning that migration flows which have taken place before 1980 are not considered. This means that the origin of the surname is tracked according to a fixed image.
- *Limitations in the data*. We are working with a bibliographic database, implying that scholarly related patterns (e.g. migrations of scholars, mobility programs, issues related on how scholars use their name in publications, etc.) as well as database-coverage related problems (e.g. orientation of the database towards Anglo-Saxon countries, the lack of coverage of surnames that have never published, etc.) can play a role. Also, possible mistakes from the database (e.g., wrong linkage of authors to addresses, typos, transcription problems, lack of information, etc.) should be taken into account when interpreting the results.

In Figure 1 we offer an overview of the methodology followed. For all the surnames in all the publications covered in the Web of Science we detected all the 'trusted' linkages between authors and countries. By a trusted linkage we mean a surname-country relationship that is

unambiguously registered in a publication[1] based on linkages between authors and countries according to bibliographic data. This implies that only in those cases where there is strong evidence that an author is linked to a country, the link is created and the combination (surname-country) is taken into consideration for the statistical analysis. These trusted linkages were created based on the following author-country combinations:

- Authors and countries from the *reprint address* field in the Web of Science are directly linked to their affiliation (Costas & Iribarren-Maestro, 2007).
- *Registered combinations of author and affiliations* recorded in the Web of Science, as from 2008 onwards WoS registers the linkage between authors and countries as they appear in the publications.
- *First authors* are assigned to the *first address* in the publication. As Calero and colleagues (2006) show the linkage of the first author with the first address of the publication is quite reliable.
- *One country publications*. For all publications with only one address or only national collaboration all their authors can be assigned to this country.

As a result, a matrix distribution of surnames by countries was created. Based on this matrix, two approaches were considered to assign surnames to countries. The first one consisted on assigning surnames to the countries with the highest frequency (in terms of publications containing the surname-country trusted linkage) which complied certain levels of assurance. This level of assurance was obtained by means of the Kullback-Liebler divergence or information gain measure. The second approach was to assign surnames according to their relative concentration by countries. This was done by using the Gini Index. In the next two subsections we detail each of the two methods proposed and the results obtained for each of them.

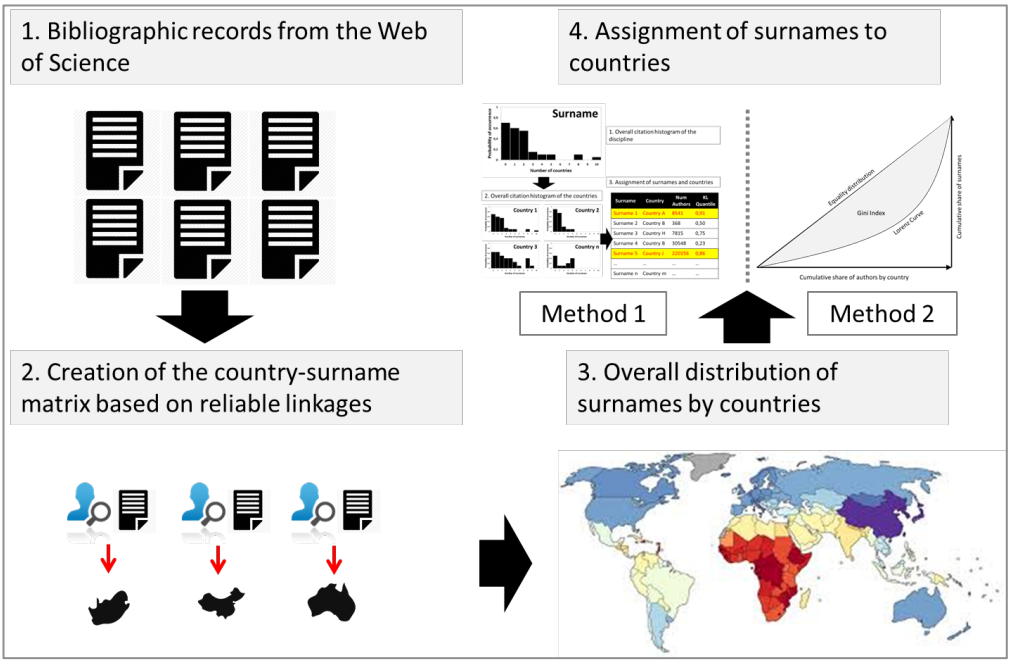


Figure 1. Overview of the methodology followed for assigning countries to surnames.

[1] For many publications in the Web of Science, not all the authors are directly linked to their affiliations in the paper, therefore sometimes it is very difficult to establish to which affiliation (and country) belongs every author.

Method 1: Kullback-Leibler divergence and distribution by country

When identifying the geographic origin of a surname one plausible approach is to consider that a surname will belong to the country with the largest number of occurrences. However, this assumption entails two problems that have to be solved. Firstly, while using raw data will benefit countries with a large presence in the database (e.g. Western and Anglo-Saxon countries), relative indicators will benefit smaller countries, preventing from a balance between countries. Secondly, some surnames may show similar numbers in various countries.

In order to overcome such limitations, we need a reasonable method to characterize the belonging of surnames to each country; and secondly, we have to be able to measure what is the amount of relative information between such characterizations. Here we propose the use of the information gain or Kullback-Leibler divergence measure (Kullback-Leibler, 1951). This measure allows us to select the country that contributes with more information to a given surname. It compares two distributions: a true probability distribution $p(x)$ and an arbitrary probability distribution $q(x)$, and indicates the difference between the probability of X if $q(x)$ is followed, and the probability of X if $p(x)$ is followed. Although it is sometimes used as a distance metric, information gain is not a true metric since it is not symmetric and does not satisfy the triangle inequality (making it a semi-quasimetric) (García et al., 2013).

In this paper, the true probability distribution $p(x)$ is represented by the authors' distribution of a given surname in the country with the highest number of such surname, while the arbitrary probability distribution $q(x)$ is represented by the frequency distribution of the surname in the rest of the countries. The objective is, on the one hand, to characterize the information gain between two probability distributions with a minimal number of properties, which are natural and thus desirable. Second, it aims to determine the form of all error functions satisfying these properties, which we have stated to be desirable for predicting surname-country dissimilarity. This analysis allows identifying similar and dissimilar distributions from a given one, but it does not explain the reasons for such dissimilarity. Such an approach has been previously used in the field of bibliometrics for very different purposes. For instance, Waltman and van Eck (2013) use it to identify national journals from international journals. García and colleagues (2013) use the Kullblack-Leibler divergence measure to determine similar academic institutions (García, et al., 2013). Finally, Torres-Salinas and colleagues (2013) apply it to characterize the field-specialization of publishers based on the citation patterns of book chapters (Torres-Salinas et al., 2013). In Figure 2 we summarize the main steps followed for assigning countries to surnames.

If we predict the similarity between the given surname and the country based on their information gain, then we can set a minimum value of information gain that should be reached in order to ensure that the assignment made is correct, thus relating the surname with the country that leads to the most alike assignment to the frequency distribution. In this case we have established a minimum value up to the percentile 0.8[2] of the overall distribution of surnames and main country by the Kullback-Leibler divergence measure in order to determine a good assurance in the surname-country association.

[2] In other words, we consider that up to 80% of the surname-country linkages based on the highest KL divergence measures are informative, and we disregard 20% of the combinations in which the surname and the country cannot be considered as a reliable linkage (as the surname could also reasonably belong to another country, based on the overall distribution of the surname across countries).

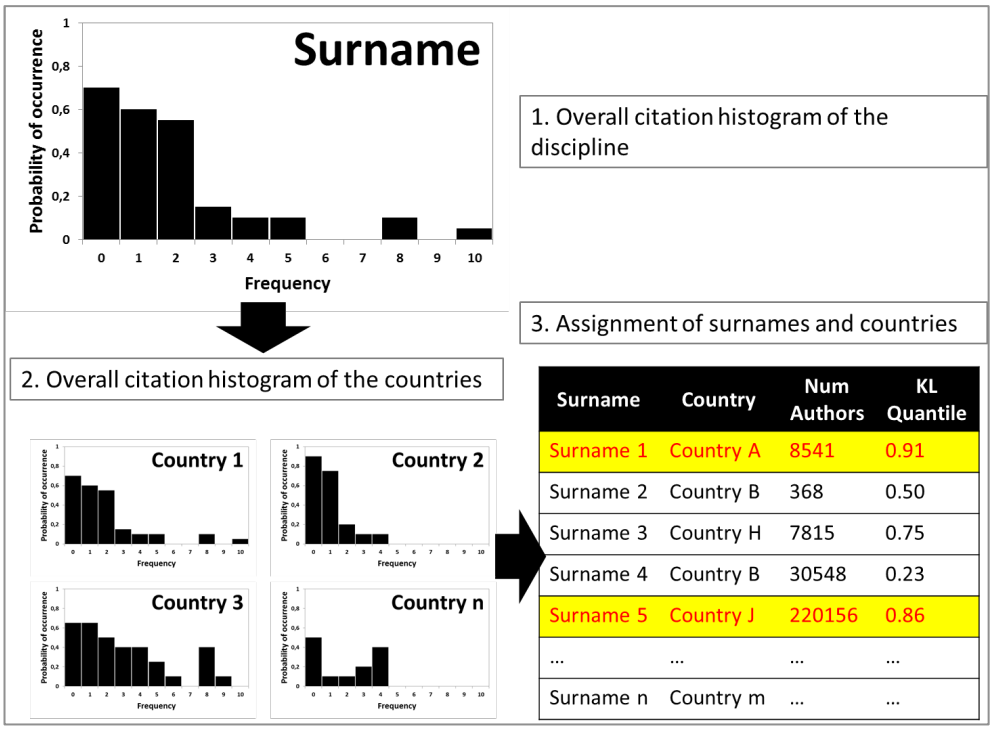


Figure 2. Overview of Method 1 employing the Kullback-Leibler divergence measure.

Table 1. Distribution of top 36 countries with the highest number of surnames according to Method 1. Kullback-Leibler Divergence.

Country	Surnames	Country	Surnames	Country	Surnames
FRANCE	138349	MEXICO	38367	FINLAND	15160
GERMANY	112445	BRAZIL	37198	UKRAINE	14582
RUSSIA	111716	GREECE	34917	CZECH REPUBLIC	14427
SPAIN	83529	IRAN	34235	NORWAY	12892
USA	76219	THAILAND	32426	DENMARK	12861
ITALY	69637	TURKEY	27671	ARGENTINA	11714
ENGLAND	63885	SWEDEN	26134	HUNGARY	10541
JAPAN	56345	ISRAEL	24482	PEOPLES R CHINA	10472
CANADA	49775	AUSTRALIA	24259	ROMANIA	9976
NETHERLANDS	41306	BELGIUM	22203	SOUTH AFRICA	9504
INDIA	41198	SWITZERLAND	21402	NIGERIA	9313
POLAND	40446	AUSTRIA	18048	EGYPT	8682

Results

A total of 1,568,052 surnames were assigned to 119 different countries. Table 1 shows the distribution by surnames of the 36 countries with the higher number of surnames assigned. As observed, the largest number of surnames is assigned to France (8.8%), followed by Germany (8.0%), Russia (7.1%) and Spain (4.9%).

As observed, some countries with the same language appear in this list, such as England and United States for English language or Spain and Mexico for Spanish language. Also some manual normalization of countries was required due to changes in the name of countries (i.e., USSR and Russia or Germany and Federal Republic of Germany).

Method 2: Gini inequality index and concentration by country

Another plausible approach to assigning countries to surnames is to consider the right country as the one where a given surname is more concentrated. For this, we suggest the use of inequality indexes such as the Gini Index. This indicator has already been used in the field of bibliometrics. For example, Torres-Salinas and colleagues (2014) employ it to determine the level of specialization of academic publishers indexed in the Book Citation Index. It is a measure of statistical dispersion. It is defined based on the Lorenz Curve, which plots the proportion of population (y axis, surnames in our case) that is cumulatively concentrated by the bottom *x*% of the population. In Figure 3 we represent its interpretation. The equality distribution is represented by a 45 degrees line. The Gini Index is defined as the ratio of the area that lies between the line of equality and the Lorenz Curve. Its value ranges between 0 and 1, 0 meaning total equality (or dispersion) and 1, total inequality (or concentration). The hypothesis we pose is that a surname can be assigned with certain levels of reliability to the country which shows a higher concentration of such surname, hence relativizing the presence of a given country in the database.

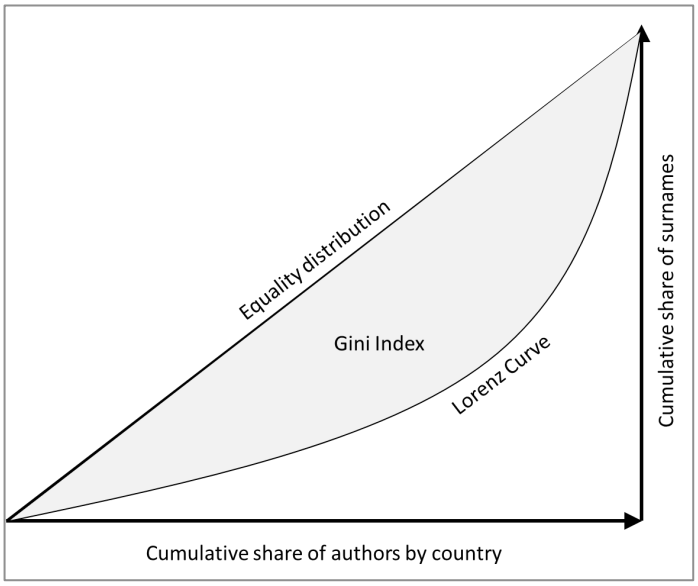


Figure 3. Interpretation of the Gini Index.

Table 2. Distribution of top 36 countries with the highest number of surnames according to Method 2. Gini Index

Country	Surnames	Country	Surnames	Country	Surnames
USA	310739	NETHERLANDS	40528	UKRAINE	17580
FRANCE	117938	BRAZIL	38386	ARGENTINA	16275
GERMANY	111375	GREECE	38034	FINLAND	16060
RUSSIA	94369	IRAN	37162	CZECH REPUBLIC	15166
SPAIN	77387	THAILAND	35090	NORWAY	15074
ITALY	65699	TURKEY	28473	DENMARK	14347
JAPAN	52399	ISRAEL	28360	HUNGARY	12291
ENGLAND	47521	SWEDEN	26051	ROMANIA	11767
CANADA	46146	SWITZERLAND	25029	SOUTH AFRICA	11018
POLAND	44087	BELGIUM	23863	NIGERIA	10619
INDIA	42897	AUSTRALIA	23396	CHINA	9531
MEXICO	41066	AUSTRIA	21609	EGYPT	9158

In Table 2 we show the distribution of surnames by countries for the top 36 countries with the highest number of surnames. A total of 1,885,782 surnames were matched to a list of 343 countries. The country with the largest number of surnames assigned is the United States, representing 16.5% of the total share, and followed by France (6.25%) and Germany (5.9%). In general terms we observe that this methodology distributes surnames among a larger number of countries, showing a less skewed distribution.

Validation

In order to validate the results of each method and determine their performance, we tried to compare them with a 'valid' list of surnames by countries. However, identifying such a list entails certain limitations. First, there is no 'perfect' and unique linkage between countries and surnames. Secondly, these linkages are not usually done for countries but rather for languages, cultures, ethnicities, etc. We decided to use a list of surnames by language provided from Wikipedia[3] and select a sample of languages.

Table 3. Control table of correspondences between countries and languages.

Normalized country	Languages	Countries
Denmark	Danish	Denmark; Greenland
England	Celtic; Anglo-Cornish; English; Scottish; Irish	Antigua & Barbuda; Australia; Bahamas; Barbados; Belize; Bermuda; Canada, England, Ghana; Gibraltar; Grenade; Guyana; Ireland; Jamaica; Liberia; Malawi; Mauritius; Micronesia; N Wales; Namibia, New Zealand, Nigeria; Scotland; Sierra Leone; Solomon Islands; South Africa, St. Kitts & Nevis; St. Lucia; St. Vincent; Trinidad & Tobago; USA; Wales; Zambia
Finland	Finnish	Finland
France	Breton; French	Benin; Burkina Faso; Congo; Côte Ivoire; Polynesia; France; French Guayana; Gabon; Guadeloupe; Guinea; Haiti; Ivory Coast; Mali; Martinique; Monaco; New Caledonia; Niger; Reunion; Senegal; Togo; Upper Volta
Germany	German	Austria; Germany; Liechtenstein
Greece	Greek	Greece
Iceland	Icelandic	Iceland
Italy	Italian	Italy; San Marino; Vatican
Japan	Japanese	Japan
Netherlands	Afrikaans; Dutch	Holland; Netherlands; Surinam
Portugal	Portuguese	Angola; Brazil; Cape Verde; Guinea Bissau; Mozambique; Portugal
Spain	Basque; Catalan; Galician;	Andorra; Argentina; Bolivia; Chile; Colombia; Costa Rica; Cuba; Dominican Republic; Ecuador; El Salvador; Guatemala; Honduras; Mexico; Nicaragua; Panama; Paraguay; Peru; Spain; Uruguay; Venezuela

We chose 20 different languages grouped in what we called 12 'normalized' countries, that is, the most representative countries of these 20 languages. Then we crossed our sample table with the surnames obtained from Web of Science and identified the countries to which each of the two methods proposed assigned these surnames. The list of countries was then processed in order to identify the 20 languages selected. We assigned to each retrieved country one of the selected language if one of the following premises was given (Table 3):

1. It was the official language of the country. For instance, French is the official language of countries such as Gabon, Haiti or Martinique.

[3] http://en.wikipedia.org/wiki/Category:Surnames_by_language

2. It is not the main language but it is only spoken in a given area. For instance, Galician, Basque and Catalan surnames were assigned to Spain, or Breton to France.
3. There is more than one official language (which is also used in other countries). This is the most important limitation noted from our validation method, as it excludes countries such as Switzerland, Belgium or Luxembourg (which have several languages spoken in more than one country). The only exception noted is Canada, which has been attributed to English language, acknowledging the important limitation towards French language.

Our validation list from Wikipedia contains a total of 8,239 surnames. After crossing this list with our list of surnames retrieved from Method 1, a total of 7,625 surnames were matched. In Table 4 we include the distribution of surnames by normalized countries according to our control list (Table 3), the coverage of 'valid' assignments made, that is, those surnames which could be assigned with certain levels of assurance according to their information gain; and the share of correct assignments.

Table 4. Distribution of surnames by countries of the control sample for 12 normalized countries according to their language, valid assignments and correct assignments according to the two methods proposed.

	METHOD 1*			METHOD 2**		
Countries	**Surnames**	**% coverage**	**% correct**	**Surnames**	**% coverage**	**% correct**
DENMARK	123	91.06%	68.75%	123	100%	60.16%
ENGLAND	932	28.76%	80.97%	929	100%	58.56%
FINLAND	225	99.11%	94.62%	224	100%	91.96%
FRANCE	562	88.08%	68.28%	560	100%	50.54%
GERMANY	2186	52.24%	69.00%	2170	100%	43.78%
GREECE	170	84.12%	78.32%	168	100%	78.57%
ICELAND	29	100.00%	65.52%	28	100%	100.00%
ITALY	972	87.65%	86.97%	968	100%	64.77%
JAPAN	1349	98.74%	98.95%	1347	100%	91.39%
NETHERLANDS	471	88.11%	60.96%	468	100%	41.67%
PORTUGAL	137	98.54%	92.59%	136	100%	91.91%
SPAIN	469	93.18%	48.74%	464	100%	54.74%
Total	**7625**	**73.22%**	**79.03%**	**7585**	**100%**	**61.29%**

*** Method 1:** Kullback-Leibler divergence; **** Method 2:** Gini Index

As observed, in general terms the coverage of 'reliable' assignments made was of 73.2% of the sample list. However, significant differences can be found by country. While in the case of Iceland all surnames were assigned with certain levels of assurance (>80 quartile of the Kullback-Leibler divergence distribution), in the case of England only 28.8% of the surnames were considered valid. Also the coverage figures are quite low for Germany (52.2%). From these surnames covered, around 80% of them were assigned to the correct country. The highest figures of correct assignments are observed for Japan (98.9%, also with a coverage of 98.5%), while the lowest figures go to Spanish surnames (48.7% of correct assignments with a coverage of 93.2%). In the case of England, although the coverage is low, 80.1% of the assignments were correct. In the case of Germany the share is lower (69%).

Regarding the methodology based on the Gini Index, a total of 7585 surnames were retrieved after crossing the list of surnames obtained with the control list. As observed, the coverage of 'reliable' assignments with this methodology is much higher (100%), however, many differences are observed on the share of correct assignments. In general terms this

methodology performs not as well as the first one, with 61.2% of all assignment correct. However, in some cases its share of correct assignments is higher. This is the case of Iceland where the 29 surnames of the control list were correctly assigned. Also the share of correct assignment for Spain increases (54.7%) but still shows low values.

Discussion and conclusions

In this paper we propose the identification of the geographic origin of surnames for bibliometric purposes. For this, we propose the use of scientific databases in order to work with data worldwide. In this way we overcome a major restriction of this type of studies regarding data availability (Cheshire, 2014). We propose two methodologies to assign countries to surnames. The first method is based on the number of surnames found in a given country when its Kullback-Leibler divergence measure is below the 80th percentile of all the combinations with the lowest Kullback-Leibler values. The second methodology is based on the concentration of a given surname in a country, using the Gini Index to calculate such concentration.

In this regard, a preliminary validation has been done comparing the coverage and correct assignments made with a sample list of 20 languages grouped into 12 'normalized countries'. The results reported are promising, especially for the first methodology. In fact, this has already been applied successfully elsewhere (Costas & Noyons, 2013). But the second methodology ensures a 100% coverage of all surnames. However, much research is still needed and further refinements in both methodologies. First, we believe that thresholds of minimum publications of a surname by country should be established in order to improve the methodology based on the Gini Index. Regarding the Kullback-Leibler divergence methodology, we considered reliable assignments those which were below the 80th percentile, however, different thresholds should be also tested. Finally, we will consider other validation lists as some questionable assignments were found in this control list (e.g., Pinto is assigned to Italian language, but it could also be assigned to Spanish or even Portuguese) which may blur the evaluation of the actual performance of each method. These methods should also be compared with those developed elsewhere.

The use of surnames to track demographic movements or analyse diversity in collaboration shows interesting opportunities for implementing these methodologies in bibliometric analyses. One example of such application is the recent work of Freeman and Huan (2014). However, frequently little attention to the methodology employed for assigning countries, languages or ethnicities to surnames is paid, something that may represent a challenge to results based on these data. Thus, understanding better the limitations and possibilities of these data is critical for a proper use. Although further research is still needed, we believe that applying methodologies such as the ones suggested here using bibliographic databases will lead to more reliable results.

References

Burchard, E., Elad, Z., Coyle, N., Gomez, S.L., Tang, H., Karter, A.J., Mountain, J.L., Pérez-Stable, E.J., Sheppard, D. & Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine*, 348(12), 1170-1175.

Calero, C., Buter, R., Cabello Valdés, C. & Noyons, E. (2006). How to identify research groups using publication analysis: An example in the field of nanotechnology. *Scientometrics*, 66(2), 365-376.

Chen, K.-H. & Cavalli-Sforza, L.L. (1983). Surnames in Taiwan: Interpretations based on geography and history. *Human Biology*, 55(2), 367-374.

Chesire, J. (2014). Analysing surnames as geographic data. *Journal of Anthropological Sciences*, 92, 99-117.

Costas, R. & Iribarren-Maestro, I. (2007). Variations in content and format of ISI databases in their different versions: The case of the Science Citation Index in CD-ROM and the Web of Science. *Scientometrics*, 72(2), 167-183.

Costas, R. & Noyons, E. (2013). "Detection of different types of 'talented' researchers in the Life Sciences through bibliometric indicators: Methodological outline". Retrieved from: http://hdl.handle.net/1887/22165

Freeman R.B. & Huang, W. (2014). Collaboration: Strength in diversity. *Nature*, 513(7518), 305.

García, J., Rodriguez-Sánchez, R., Fdez-Valdivia, J., Robinson-García, N. & Torres-Salinas, D. (2013). Benchmarking research performance at the university level with information theoretic measures. *Scientometrics*, 95(1), 438-452.

Hatton, T.J. & Wheatley Price, S. (2005). Migration, migrants and policy in the United Kingdom. In: Zimmermann, K.F. (ed.), *European migration: What do we know?* (pp. 113-172). Oxford University Press.

Kissin, I. (2011). A surname-based bibliometric indicator:publications in biomedical journal. *Scientometrics*, 89(1), 273-280.

Kissin, I. & Bradley, E.L.J. (2013). A surname-based patent-related indicator: The contribution of Jewish inventors to US patents. *Scientometrics*, 97(2), 357-368.

Kullback, S. & Leibler, R.A. (1951). On the information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79-86

Mateos, P. (2007). A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space and Place*, 13(4), 243-263.

Moed, H.F. & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. *Scientometrics*, 101(3), 1987-2001.

Torres-Salinas, D., Robinson-García, N., Campanario, J.M. & Delgado López-Cózar, E. (2014). Coverage, field specialisation and the impact of scientific publishers indexed in the Book Citation Index. *Online Information Review*, 38(1), 24-42.

Torres-Salinas, D., Rodriguez-Sánchez, R., Robinson-García, N., Fdez-Valdivia, J. & García, J.A. (2013). Mapping citation patterns of book chapters in the Book Citation Index. *Journal of Informetrics*, 7(2), 412-424.

Waltman, L. & van Eck, N.J. (2013). Source normalized indicators of citation impact: an overview of different approaches and an empirical comparison. *Scientometrics*, 96(3), 699-716.

An 80/20 Data Quality Law for Professional Scientometrics?

Andreas Strotmann[1] and Dangzhi Zhao[2]

[1] *andreas.strotmann@gmail.com*
ScienceXplore, D-01814 Bad Schandau (Germany)

[2] *dzhao@ualberta.ca*
University of Alberta, School of Library and Information Studies, Edmonton, Alberta (Canada)

Scientometric network error consequences

Only very recently have researchers begun looking at what concrete effect the errors in a network model caused by name ambiguities in the data sources may have on the results of popular types of network analysis. The results that they report are quite alarming in the aggregate: not only do typical evaluative analyses of individuals (e.g., citation rankings) suffer significantly from these errors, but there is mounting evidence that even the most basic statistical features of realistic large-scale networks are hugely distorted by ambiguities. Strotmann et al. (2009), for example, document significant distortions in co-authorship network visualizations, and Diesner and Carley (2013) report that "minor changes in accuracy rates of [name disambiguation] lead to comparatively huge changes in network metrics, while the set [of] top-scoring key entities is highly robust. Co-occurrence based link formation entails a small chance of false negatives, but the rate of false positives is alarmingly high."

In fact, Fegley and Torvik (2013) go so far as to dismiss one of the most famous recent results in large-scale social network analysis, the exact power-law distribution from preferential attachment (Barabási & Albert, 1999), at least in the case of scientific collaboration networks (Barabási et al., 2002), as a mere artefact produced by a lack of name disambiguation in the underlying dataset! The ultimate irony here is that Fegley and Torvik's (2013) data are consistent with an interpretation that Barabási's cooperation network power may have been induced by a power law distribution of name ambiguities rather than co-authorships.

Similarly, Strotmann and Zhao (2013) find that even highly stable statistical analysis methods of author co-citation analysis fail in the face of large-scale ambiguity errors in the underlying dataset.

While for evaluative bibliometrics the most serious problem is generally the "splitting" of individuals, i.e., the failure to recognize each and every one of an individual's contributions correctly (especially of high-performing individuals), Fegley and Torvik (2013) find that splitting is not the main concern in relational network analysis. Instead, they and Strotmann and Zhao (2013) both find that it is the erroneous "merging" of individuals, i.e., the failure to separate the contributions of multiple individuals correctly because their names are too similar, that causes major distortions of large-scale network analysis results in relational network analysis. Especially East Asian names are prone to extreme amounts of merging. While in European cultures there are relatively few common given names but a large variety of family names, in Chinese, Korean and other East Asian cultures the opposite is the case—a small number of surnames is shared by half their populations, but given names are much more varied. The old tradition in scientific publishing to list authors by their surnames and initials works, sort-of, when science is done in European-origin cultures, but all bibliographic databases have in recent years had to move to a full-name model as research boomed in the Asian Tiger nations (e.g., PubMed/MEDLINE in 2002).

When is a scientometric network sufficiently complete and clean?

As Torvik and Smalheiser (2009) make abundantly clear, it is for all intents and purposes impossible to disambiguate the names of all the individuals in a large dataset completely and fully correctly. With absolute perfection thus out of the question, what remains is to ask when a disambiguation is "good enough", and if (and how) it is possible for a typical researcher to go about disambiguating the dataset well enough. Unfortunately, there is very little research, if indeed any, into what constitutes "good enough" for a scientometric study. The few studies that have looked into what goes wrong when individuals are not recognized correctly do give us a hint, though.

First of all, "good enough" usually means that the most important contributions of the top-ranked individuals must be absolutely correctly attributed. Whatever other good methods (e.g., name disambiguation algorithms or author registries) we may find to disambiguate our data, in the end it will therefore be necessary to manually double-check, and where necessary fix, the highest-impact individuals' data. Secondly, some statistical procedures or network measures are more vulnerable than others to name ambiguities. Local network measures (e.g., node degree) are less affected than global ones (e.g., size of connected component), and evaluative studies (e.g., ranking) are more affected than relational ones (e.g.,

correlations) (Diesner & Carley, 2013; Strotmann & Zhao, 2012).

An 80/20 scientometric data quality rule?

For ranking studies, absolute correctness is paramount, and huge efforts need to be expended to get all the top-ranked individuals just right. When the "individuals" are research institutions, this can be a daunting task. For correlative studies, on the other hand, a study by Albert, Jeong, and Barabási (2000) warns us that, while global measures of power-law distributed networks may be quite resilient to *uniformly* distributed random errors, they are also quite vulnerable to the kind of *highly skewed* error distributions that we observe for name ambiguities, for example. In the case of an extremely skewed error distribution, they observed that an error rate as low as 10%-20% completely changed the measured values for a fundamental global network metric, namely, connectivity.

We can take this as a warning that, as a rule of thumb, we generally need to aim for a roughly 90% (but definitely 80% or better) complete and correct dataset when error distributions are skewed. Note that the requirement of 80% completeness or better applies, in particular, to the underlying citation index's coverage of the field being studied: a focus on high-impact literature implies a highly skewed error distribution! On the plus side, studies on the life sciences can thus be relied upon to yield reliable results as long as their disambiguations are good. Results from *any* scientometric study on the social sciences, however, are suspect as long as they rely on these databases and these databases cover much less than 80% of the literature in those fields.

Note that an 80% data correctness requirement for a professional scientometric study would apply to the data as it is used for network statistics. When both data collection *and* cleaning are subject to highly skewed error distributions, this means that we need 90% correct data collection *and* 90% correct data cleaning to guarantee 80% correct data for analysis.

Conclusions: the bad news and the good

This, then, is the bad news for those who aim to provide a truly professional scientometric service to their community: power-law-like data *and* error distributions may mean that only nearly-complete *and* nearly-clean datasets can be trusted to serve as a reliable basis for nearly *any* type of network or statistical analysis.

The good news is that there are plenty of successful bibliometric studies that imply that this level of correctness is also usually quite sufficient for meaningful studies, as long as only "local" measures or relational statistics are required. There *are* fields that are covered to 90%+ in citation databases, e.g., the citable literature of the life sciences, and there *are* disambiguation methods (e.g., some of those reviewed in Smalheiser & Torvik, 2009 or that of Strotmann et al., 2009) that do make reliable scientometric studies possible. However, scientometric professionalism may well require that these methods be utilized in nearly *all* future studies, and thus, that they be applied to, and adopted by, the citation databases themselves.

Acknowledgments

This Ignite Talk extends, with permission of the publisher, Section 4.4, "Disambiguation in Citation Network Analysis: Ambiguity and Power Laws," of Zhao & Strotmann (2015).

References

Albert, R., Jeong, H.W. & Barabási, A.L. (2000). Error and attack tolerance of complex networks. *Nature 406*, p.378

Barabási, A.L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*, p.509

Barabási, A.L., Jeong, H., Neda, Z, Ravasz, E, Schubert, A. & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A, 311*, p. 590

Diesner, J. & Carley, K.M. (2013). Error propagation and robustness of relation extraction methods. *XXXIII International Sunbelt Social Network Conference*, Hamburg, Germany, May 2013.

Fegley, B.D. & Torvik, V.I. (2013). Has large-scale named-entity network analysis been resting on a flawed assumption? *PLoS ONE 8* (7): e70299.

Smalheiser, N. R. & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology* 43, 287.

Strotmann, A., Zhao, D. & Bubela, T. (2009). Author name disambiguation for collaboration network analysis and visualization. *Proceedings of the American Society for Information Science and Technology 2009 Annual Meeting*, November 6–11, 2009, Vancouver, BC, Canada

Strotmann, A. & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63* (9), p.1820

Torvik, V. I. & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data, 3* (3)

Zhao, D. & Strotmann, A. (2015). *Analysis and Visualization of Citation Networks.* Morgan & Claypool.

Some Features of the Citation Counts from Journals Indexed in Web of Science to Publications from Russian Translation Journals

Maria Aksenteva

ms@ufn.ru
"Uspekhi Fizicheskikh Nauk" Editorial Office, Lebedev Physical Institute, Russian Academy of Sciences, Leninskii prosp. 53, 119991 Moscow (Russia)

Introduction

As it was emphasized by Moed, H.F., Glänzel W. & Schmoch U. (2005) in their editors' introduction to the *Handbook of Quantitative Science and Technology Research*: "A most important data source for analysis of the science system is the *Science Citation Index* (SCI) and related Citation Indexes published by the *Institute for Scientific Information* (ISI–Thomson Scientific, Philadelphia, PA, USA), or, in a more recent version, ISI's *Web of Science*." Due to this very competent opinion (supported of course by major part of scientists all over the world) it is very important for proper evaluation of the science and its development in Russia to investigate how publications in Russian journals indexed in SCI and how citations to these publications were counted and recorded in SCI in previous decades and is counted and recorded now in Web of Science (WoS).

Some systematic problems with proper indexing and correct counting of citations to publications in Russian journals in SCI was revealed by brilliant founder of modern bibliometrics ("statistical bibliography") Eugene Garfield long time ago in 1974. The greatest problems (according to Garfield) occurred with so-called "translation journals": "The term Russian journals is used here as it is daily used in libraries in the United States. We are aware of its inadequacy and inaccuracy, but plead its convenience. A few of the journals are Slavic, but not Russian. The term Soviet journals might seem more appropriate, but it would not be. An important group of the journals considered is published outside the Soviet Union the so-called translation journals. Neither Russian nor Soviet, they are nevertheless the product of Russian and Soviet research. They also present, as we learned in this study, a formidable stumbling block in journal citation analysis of this type. I speak here only in terms of statistical bibliography as regards the translation journals." (Garfield, 1974).

What was (and is now) the biggest problem with indexing and counting of citations of the "translation journals"? It was (and is now) the adopted by SCI (now Web of Science) policy of the counting of citations to original publications (articles, published in Russian) and to the English version of the same article, published in "translation journals". As it was found in the present research this policy were changed several times during the period of SCI existence and this policy can significantly affect the conclusions, which could be made about Russian science in many analytical reports and investigations, based on *Web of Science* data (see, for example, Albarrán et al., 2013).

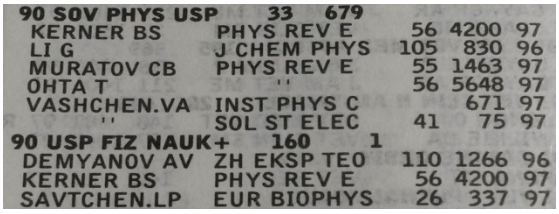
In this research we studied the style (the policy) of records for publications from Russian (and translation) journals and counting of citations to them in printed volumes of SCI in 1960-1998 years and compared these styles with the policy, adopted in the internet version of the successor of SCI (WoS) in 1990-es and now. It is possible to say after this investigation, that significant (sometimes huge) amount of citations (from the journals indexed in WoS) to Russian publications are not possible to find in WoS now without some complicated additional search. All these citations are not taken into account in many analytical reports about Russian science (especially about natural science such as physics, chemistry, biology etc.). At the same time it is not very difficult now to return back to the Garfield's old policy of records and calculations of the citations to Russian publications in translation journals, which could collect properly all citation using new possibilities of Internet linking of publications. (See, for example, UFN journal's web-site www.ufn.ru on which the citing articles are collected using CrossRef system (using Digital Objects Identifier - DOI) or www.mathnet.ru site for more precise and elegant citations collecting (Zhizhchenko & Izaak, 2009; Chebukov et al, 2013)).

Methodology and data

We compared the number of citations to an article published in "Uspekhi Fizicheskikh Nauk" (UFN) journal (or to the English translation to the same article published in "Physics-Uspekhi" (former "Soviet Physics-Uspekhi" journal until 1992 year) ─ cover-to-cover English translation of UFN journal) presented in printed volumes of SCI with the number of citations to the same article presented in Web of Science (on-line version) and with the number of citations, which could be found using CrossRef links (DOI) on www.mathnet.ru and/or www.ufn.ru web-sites (see details in Aksenteva, Kirillova & Moskaleva, 2013).

Results and discussion

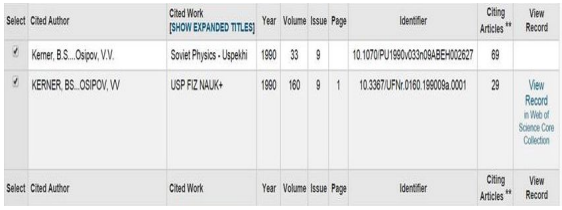
Let's consider (as a typical example) an article (Kerner & Osipov, 1990). First of all we have found that in printed volume of SCI (see Fig. 1) both Russian original article and its English translated version were indexed (citations to them were collected separately, but all citations were displayed, see Figure 1):



90 SOV PHYS USP	**33**	**679**		
KERNER BS	PHYS REV E	56	4200	97
LI G	J CHEM PHYS	105	830	96
MURATOV CB	PHYS REV E	55	1463	97
OHTA T	''	56	5648	97
VASHCHEN.VA	INST PHYS C		671	97
''	SOL ST ELEC	41	75	97
90 USP FIZ NAUK+	**160**	**1**		
DEMYANOV AV	ZH EKSP TEO	110	1266	96
KERNER BS	PHYS REV E	56	4200	97
SAVTCHEN.LP	EUR BIOPHYS	26	337	97

Figure 1. Copy from SCI (1997) for Kerner B.S.

But now in WoS (internet version) we cannot find citations to the English version of this article. It is possible to find them only by using the WoS's option "Cited References Search" (see Figure 2).



Select	Cited Author	Cited Work [SHOW EXPANDED TITLES]	Year	Volume	Issue	Page	Identifier	Citing Articles **	View Record
✓	Kerner, B.S....Osipov, V.V.	Soviet Physics - Uspekhi	1990	33	9		10.1070/PU1990v033n09ABEH002627	69	
✓	KERNER, BS...OSIPOV, VV	USP FIZ NAUK+	1990	160	9	1	10.3367/UFNr.0160.199009a.0001	29	View Record in Web of Science Core Collection
Select	Cited Author	Cited Work	Year	Volume	Issue	Page	Identifier	Citing Articles **	View Record

Figure 2. Cited references search in WoS core collection for article Kerner B.S. &Osipov, 1990.

It is possible to see on this figure, that there are 29 citations to the Russian version of this article and 69 citations to the English version of the article, but (unfortunately for the Russian journal) it is possible to view citing articles for the Russian version only (only 29 citing articles). 69 citations to the English version of this article are not taken into account in Prof. Kerner's (and of course for Prof. Osipov too) citation report, are not included into their Hirsh's indexes, are not taken into account for his laboratory and his institute bibliometrics etc. (and for Russian physics and science in general). On our web-site using CrossRef links it is possible to find 70 citing article: http://ufn.ru/ru/articles/1990/9/a/.

It is necessary to mention that for publications in UFN journal until September 2001 only citations to the Russian version are presented in WoS (but citations to the English version are not taken into account). We have checked more than one thousand articles (published in 1990-2000 years in UFN) and have found that about 67% of citations (in average) to these articles were not presented now directly in WoS (and so do not taken into account for any analytical scientometric report). According to WoS in 1990-2000 years 1190 articles were published in UFN (and indexed in WoS) and they have only 9002 citations (on April 25, 2015). Using DOI on our website we have found 14973 citations to 1167 articles, published in UFN in the same period.

Conclusions

It was found that now WoS show less than half of citations (from journals indexed in WoS) to described above article (Kerner, Osipov, 1990), but this is not an exceptional example. So all publications in Russian translated journals (indexed in WoS) lose a lot of their absolutely correct citations (about 60% in average) from journals indexed in WoS and therefore scientometrics, based on WoS direct data, underestimates the real impact of Russian scientists and science in general.

Acknowledgments

I am grateful to Prof. M.Yu. Romanovsky who has encouraged me to make this investigation.
The work was supported by the Russian Foundation for Basic Research (project No. 13-07-00672 a).

References

Aksenteva, M.S., Kirillova, O.V., & Moskaleva, O.V., (2013) On Paper Citation by Web of Science and Scopus From Translated Russian Journals (in Russian), *Nauchnaya periodika: problemy i resheniya* [Scientific periodical press: problems & solutions], No. 4(16), 4-18. Retrieved January 18, 2015 from http://ufn.ru/tribune/trib124.pdf

Albarrán, P., Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2013). Differences in citation impact across countries. In *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference, Vienna, Austria*. Vol. 1, p. 536. Retrieved January 18, 2015 from http://www.issi2013.org/Images/ISSI_Proceedings_Volume_I.pdf

Chebukov, D., Izaak, A., Misurina, O., Pupyrev, Yu. & Zhizhchenko, A., (2013) "Math-Net.Ru as a digital archive of the Russian mathematical knowledge from the XIX century to today", *Lecture Notes in Computer Science*, 7961 (Ed. J. Carette et al.) 344–348, arXiv: 1305.5655.

Garfield, E. (1974) "Russian Journal References and Citations in the Science Citation Index Databank". In *Journal Citation Studies*, 22. Philadelphia: ISI. Retrieved 01/18/ 2015 from: http://www.garfield.library.upenn.edu/papers/244.html

Kerner, B.S., Osipov, V.V. (1990) *Usp. Fiz. Nauk* 160 (9) 1–73 [*Sov.Phys.Usp.* 33 (9) 679–719]

Moed, H.F., Glänzel W. & Schmoch U. (Eds.) (2005) Handbook of Quantitative Science and Technology Research. The Use of Publication and Patent Statistics in Studies of S&T Systems, Dortrecht: Kluwer Academic Publishers.

Zhizhchenko A.B. & Izaak A.D. (2009) "The information system Math-Net.Ru. Current state and prospects. The impact factors of Russian mathematics journals", *Russian Math. Surveys*, 64**,** 4.

Semantics, a Key Concept in Interoperability of Research Information -The Flanders Research Funding Semantics Case

Sadia Vancauwenbergh

Sadia.Vancauwenbergh@uhasselt.be
ECOOM-UHasselt, Hasselt University, Research Coordination Office, Martelarenlaan 42, BE-3500 Hasselt (Belgium)

Introduction

In a knowledge-based economy, a good overview of the scientific and technological portfolio is essential for policy formation and driving knowledge transfer to the industry and the broad public. In order to enhance open innovation, the Flemish public administration has created a Flanders research information portal (FRIS, http://www.researchportal.be) that integrates information available from its data providers (research institutions, funding organizations…) using the CERIF (The Common European Research Information Format) standard. Although this standard allows for almost unlimited flexibility for modelling the research information, it has limitations when it comes down to communication to end-users, in terms of semantics. However, interoperability of research information is only meaningful when a well-defined semantics is used. This paper describes the implementation of a business semantics tool on data concepts and classifications for research funding as a means to unambiguously exchange and interpret these data.

The need of semantics

A couple of decades ago, the demands on the research community to report on research data were rather low. Results were published in preferably highly-rated journals and rather limited research reports were written. Over the years, more research data became available and the need for research databases grew. Unfortunately, these databases were predominantly developed per organization without consultation of other organizations. Moreover, because of the rather low data volume and people involved, there seemed no explicit need for defining an accompanying semantics.

However, as the research system expanded, there has been a massive increase in the amount and nature of the information stored as well as its information consumers. These changes are not only due to the advancements made in the research field itself, but are also explained by the global efforts undertaken to transfer the obtained knowledge to industry and the broad public. In Flanders, this resulted in the creation of the FRIS-portal which makes Flemish research information publicly available. This information is provided via a multitude of data providers that often use a different terminology for a similar concept or alternatively, use a similar terminology for a different concept. The correct interpretation of the information at the FRIS portal is realized by the addition of a semantic layer on top of the data by the data providers, which later on is translated to a general FRIS semantics resulting in data communication in the same language. The focus on the explicit semantic alignment with the data providers, adds further to existing initiatives like VIVO and CERIF based CMS (Guéret et al., 2013). Data unambiguity is increasingly important, in an era where many initiatives have seen light to measure and benchmark research and where public research reporting obligations are vastly growing. Obviously, the lack or incomplete definition of semantics puts large constraints on the interoperability of research information, and in extension on the policies drawn out of these data.

The Flanders research information landscape

In Flanders, research institutions receive funding from a broad range of research funding providers going from the regional to national and international level. Obviously, each funding provider has its own requirements with regards to the formats or classifications used for reporting on the resulting research output, thereby creating a multitude of largely similar research reports. Obviously, this places a large burden on the research community. Until now, the data providers tried to keep pace with this vast expansion of research reporting by improving or even creating databases, unfortunately without generally agreed upon semantics. At the same time, the data providers were feeding their information to the FRIS-portal in order to increase the visibility of the research in Flanders to third parties (i.e. companies, research institutions and individual researchers).

In line with the growing concern on the administrative burden put on the research community, a report was published by Peters et al. (2011) providing guidelines for the reduction of redundant research information reporting. Following these advices, the Flemish Department of Economy, Science and Innovation (EWI) is currently improving the FRIS-portal in order to be used as a virtual research information space, for information retrieval in a transparent and automated manner that can be used for research reporting (Figure 1) (Debruyne et al., 2011). This implicates

the use of unambiguous data concepts and research funding classifications. Until recently, funding organizations were using their own funding classification schemes which were semantically poorly defined and lacked concordance mappings to other (inter)national classifications.

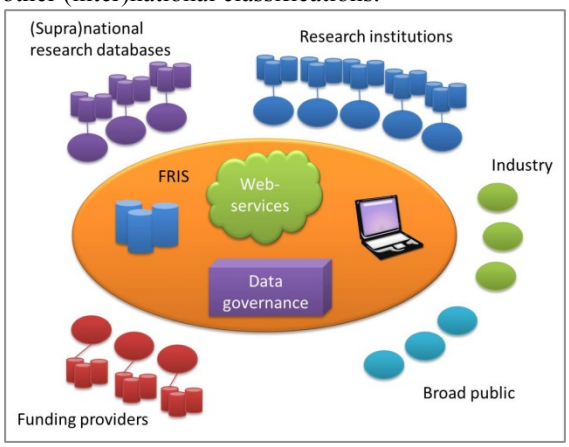


Figure 1: Representation of the FRIS design.

Funding data and classification governance

In order to add a semantic layer on top of the FRIS database layer, the Data Governance Centre® (DGC) platform of Collibra has been used. This platform allows data suppliers to manage their own data models used to describe, i.e. research funding together with the corresponding institution specific semantics (Figure 2).

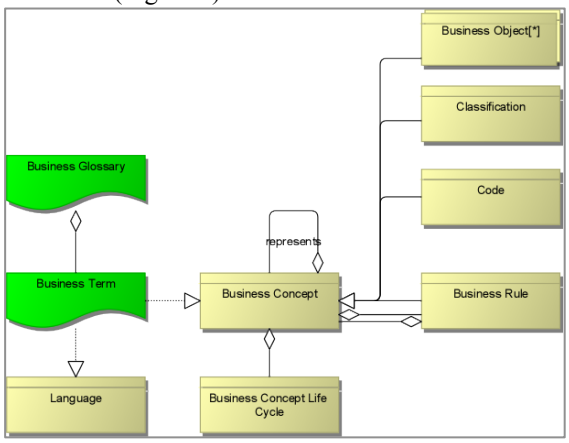


Figure 2. Incorporation of a business semantics glossary on the research funding model.

At the same time, the DGC platform has been used for the description of each individual component of the FRIS research funding model using definitions (Figure 3). By explicitly defining all concepts, the governance tool assists in the swift identification of semantic inter-organizational misalignments when mapping corresponding concepts by the stakeholders. The resulting ontologies can be exported and used to annotate data in relational databases, and hence render data meaningful. Furthermore, the DGC tool has been used for defining the semantics of classifications and code sets on research funding, which is essential when it comes down to consistent and unambiguous reporting on research funding to third parties. Obviously, the research community at large will benefit from this, as the information retrieved via FRIS will be much more reliable and accurate.

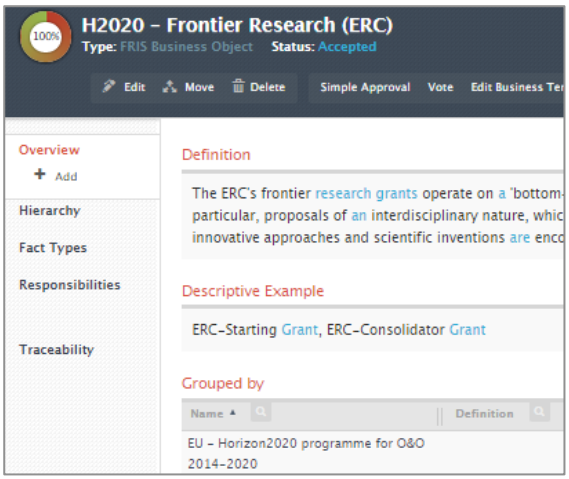


Figure 3: DGC as a governance tool for research funding classifications.

Altogether, the use of a data governance tool focused on semantics opens new avenues in terms of efficiency of the research ecosystem. Not only will governments be able to delineate better founded policies, also research administrations and researchers themselves can gain tremendously as research reporting could be automated from the FRIS-portal in a reliable manner, thereby reducing the administrative burden at the benefit of scientific discovery and innovation.

Acknowledgement

This work is part of the Classification Governance project carried out for the Expertise Centre for Research & Development Monitoring (ECOOM) in Flanders, which is supported by the Department of Economy, Science and Innovation, Flanders.

References

Debruyne, C., De Leenheer, P., Spyns, P., Van Grootel, G., & Christiaens, S. (2011). Publishing open data and services for the Flemish research information space. In *Advances in Conceptual Modeling. Recent Developments and New Directions,* Springer.

Guéret, C., Chambers, T., Reijnhoudt, L., et al. (2013). Genericity versus expressivity – an exercise in semantic interoperable research information systems for Web Science. *In: http://arxiv.org/pdf/1304.5743.pdf*

Peters, A., & Lambrechts, L. (2011). De vereenvoudiging van onderzoeksverslaggeving, een analysetraject uitgevoerd door de Vlaamse universiteiten en hogescholen en de VLIR, in opdracht van de Vlaamse Overheid (EWI).

The Information Retrieval Process of the Scientific Production at Departmental-level of Universities: A New Approach.

César David Loaiza Quintana[1] and Víctor Andrés Bucheli Guerrero[2]

[1] *cesar.loaiza@correounivalle.edu.co*
Universidad del Valle, Escuela de Ingeniería de Sistemas y Computación, Facultad de Ingeniería. Universidad del Valle, Sede Meléndez. Calle 13 # 100-00. Cali, Valle Del Cauca (Colombia).

[2] *victor.bucheli@correounivalle.edu.co*
Universidad del Valle, Escuela de Ingeniería de Sistemas y Computación, Facultad de Ingeniería. Universidad del Valle, Sede Meléndez. Edificio 331 oficina 2113, Calle 13 # 100-00. Cali, Valle Del Cauca (Colombia)

Introduction

Our work was focused on document retrieval from Scopus databases of the Escuela de Ingeniería de Sistemas y Computación (EISC) of the Universidad del Valle (Cali - Colombia).

The databases systems as WoS (Web of Science) or Scopus contain the knowledge produced by engineer schools. However, this information is ambiguous and the retrieving of the specific documents of one school is identity uncertainly (Pasula et al., 2003). Thus, the design of machines (search engines) to retrieve the relevant documents of engineer schools is a complex process.

After the work of Bucheli et al. (2013); Cuxac, Lamirel, & Bonvallot (2013) proposed a semi-supervised approach, mixing soft-clustering and Bayesian learning. Additionally, Huang et al. (2014) proposed a rule-based algorithm. Both approaches were for affiliation disambiguation.

We reproduced the model proposed by Bucheli et al. (2013). The results show that the model can be used to information retrieval of department-level. In addition, we proposed a new approach addressing the problem of classification using network science. The future work will be related with building a model according to the network science approach.

Methodology

Model of Bucheli et al. (2013)

We followed the methodology specified by Bucheli et al. (2013) shown in Figure 1(a).

1) The configuration of the initial search strategy proposed by Bucheli et al. (2013) was applied using the Scopus search engine to get a set **I** composed by documents that contains all the documents that belong to EISC and others that not belong to it.

2) The initial search strategy was based on a review of the research activity of the School and it proposes recovering a set of documents $\mathbf{I} = \mathbf{A} \cup \mathbf{J} \cup \mathbf{S} \cup \mathbf{O}$. The staff **S** set is made up by papers which are related to a list of school professors names explicitly. The journal set **J** is the bunch of documents published in the journals where the school has previously published. The address set **A** is related to the documents that have in their affiliation the name of the school explicitly. Finally, socio-semantic set $\mathbf{O} = \mathbf{S} \cup \mathbf{C}$, where the concepts set **C** is made up by the documents related to a bunch of research areas from a school. Every set mentioned before has an additional restriction; his documents must belong to the university that hosts the internal-level unit, in our case to the Universidad del Valle.

3) An Expert from EISC classified all the documents from the initial search and we built a relevant set **R** with **I** elements that belong to EISC.

4) We built a dataset where one paper or instance is characterized by a vector (with five positions). Each position is a binary variable, related to sets **A**, **S**, **J**, **O** and **R**, that tell us if the paper belongs or not to the corresponding set. Thus, the instance class is determined by the variable **R**.

5) Afterwards, we made the classification using the Naïve Bayes model of information retrieval illustrated in (1). It was evaluated based on all instances of the dataset. We used standard measurements over cross validation test 10 fold (Witten, 2005; Baeza-Yates, 1999). On the other hand, the publication year was taken into account as parameter of evaluation. Thus, we train the model with paper published between two specific years, for instance 1989-2010 and testing the model with papers published in the following years. This procedure was evaluated by the following years of training 1989-2011, 1989-2012 and 1989-2013.

$$p(R|J,S,O,A) = \frac{p(R)p(J,S,O,A|R)}{p(J,S,O,A)} \quad (1)$$

Proposed model based on network science

The machine learning process follows five phases: Selecting data, expert validation, co-author network building, feature extraction from network and classification, as shows the Figure 1(b).

The data selection trough the initial search strategy and the expert validation have be taken into account similarly to the review model of Bucheli et al. (2013). Here, the document corpus used is the same of evaluation model applied to the EISC, however the feature extraction changes and the features are related with network measurements.

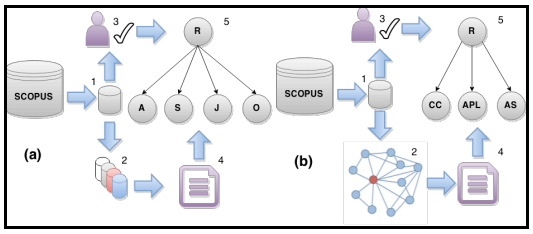


Figure 1. The five phases of the evaluated and proposed methodologies.

The document corpus contains information about co-authorship relations. Each author is identified by an ID that Scopus assigns. We build a co-authorship network, where, the network is traduced as a weighted and undirected graph in which the weight of the edges designates the number of documents where whichever two authors have participated. The new dataset is built as follows: one document or instance is a vector of values where each position is a variable related with one measurement of co-author network, in which, the specific paper was subtracted. Thus, for each instance, the authors that participated in the specific document are deleted and the measures are computed again. Additionally, the last variable **R** shows if the paper belongs or not belongs to the School. The measurements of networks are:

1. The Cluster Coefficient (**CC**): The local clustering coefficient captures the degree to which the neighbours of a given node link to each other. We use the average of all local clustering coefficients.
2. The average path length (**APL**) is the average distance between all pairs of nodes in the network.
3. The average strength (**AS**), is the average of the sum of the edge weights of each node. (Barabasi. 2012).

Finally, we develop a supervised learning environment through a Naïve Bayes Classifier and the proposed model is evaluated and compared with the model proposed by Bucheli et al (2013).

Results, discussion and future work

Table 1 shows standard evaluation measurements. Here, we introduce the cross validation fold 10 test, the measurements show in Bucheli. et al. (2013), and the evaluation for different publication years 1989-2011, 1989-2012 and 1989-2013. The results show that the model was applied to other School with similar performance measurements, in this sense the model is consistent and allows to build one search engine of department-level. Additionally, we evaluated the practical utility of the model, verifying that it is capable of doing an acceptable prediction of EISC's documents published after a specific date when it is trained with a set of documents published until that date.

In this work, we found the finger prints of department-level of universities that allow us to design search engines that retrieve relevant documents of department-level.

Table 1. Evaluation measurements of the model.

	Recall	Precision	ROC curve
EISC Univalle			
Cross Validation fold 10	0,932	1,000	0,989
Bucheli et al. (2013)			
Department of Industrial Engineering –University of Pittsburgh	0,494	0,997	0,984
Faculty of Engineering – Universidad de los Andes (Colombia)	0,954	0,992	0,965
EISC Univalle			
Training:1989-2011 Evaluation: 2012-2014	0.833	1.000	0.974
Training 1989-2012 Evaluation: 2013-2014	0.826	1,000	0.964
Training 1989-2013 Evaluation: 2014	0,786	1,000	0,939

The networks science approach is an opportunity to propose a mathematical model able to learn the structure of co-authorship network from a particular school. Then, we can design a classifier of relevant documents at department-level based on co-authorship relations. This allows making a classification with little a priori information about an organization, which turns into a more general model than Bucheli et al. (2013). We proposed a model, namely (2).

$$p(R|CC, APL, AS) = \frac{p(R)p(CC, APL, AS|R)}{p(CC, APL, AS)} \quad (2)$$

We suggest as future work to evaluate the model based on network measurements at the same school and other 3 schools of engineering from different universities.

Acknowledgments

Thanks to Convocatoria Interna, Universidad del valle 2014; Facultad de Ingeniería, Universidad del valle; and EISC.

References

Baeza-Yates, R. (1999). *Modern information retrieval*. New York: Addison-Wesley.

Barabási, A.L. (2012). *Network science book*. Center for Complex Network Research, Northeastern.

Bucheli, V., Calderón, J., Gonzales, F., Bidanda, B., Valdivia, J., & Zarama, R. (2013). Model to support the information retrieval process of the scientific production at departmental-level or faculty-level of universities. *Proc. ISSI*.

Cuxac, P., Lamirel, J. C., & Bonvallot, V. (2013). Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics*, *97*(1), 47-58.

Huang, S., Yang, B., Yan, S., & Rousseau, R. (2014). Institution name disambiguation for research assessment. *Scientometrics*, *99*(3), 823-838.

Pasula, et al. (2003). Identity Uncertainty and Citation Matching, *NIPS,* MIT Press.

Witten, I. (2005). *Data mining: practical machine learning tools and techniques*. 2nd ed., Amsterdam: Morgan Kaufman.

Efficiency, Effectiveness and Impact of Research and Innovation: a framework for the analysis

Cinzia Daraio

daraio@dis.uniroma1.it
Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, via Ariosto, 25 00185 Rome (Italy)

Introduction, motivation and policy relevance

The main objective of this paper is to provide a framework for the assessment of the research activity and its impacts. This is a difficult task. First of all, because of the heterogeneity, partial overlapping and fragmentation of the different streams of literature. Secondly, due to the need of applying a systemic approach to account for the complexity of the research activity and its complementarities and interrelationships with teaching, third mission activities and other relevant dimensions of performance, including the inputs.

This work originated from Daraio (2015) which pointed out the unavailability of a best evidence on the "efficiency, effectiveness and impact of research and innovation" due to the lack of a suitable framework for a comprehensive analysis.

Two recent policy initiatives witness the need and call for the proposal of a general framework for assessing research and its impact. We refer to the STAR metrics in US and to the EC (2014) "Expert Group to support the development of tailor-made impact assessment methodologies for ERA" in Europe.

We discuss in the following the main dimensions of our framework which are: 1. Theory, 2. Methods, 3. Data.

Research and innovation in the theory

In theory, the following streams of literature have considered research and innovation as the main link of Science and Society interplay:

• *Economics of science and technology* as an emerging field, which draws on the fields of economics, public policy, sociology and management (Audretsch et al., 2002).

• *Growth theory* (Aghion & Howitt, 2009), within which «the residual» is considered as technology advance over time (Solow, 1957); or as our ignorance (Abramovitz, 1956). The old growth theory (Nelson & Phelps, 1966) considers as additional inputs investments in R&D and education while the new growth theory (Romer, 1986; 1994) emphasizes the influence of other factors such as technologies or efficiencies, spill-overs and incentive of agents.

• *Quantitative science and technology research*, organized as quantitative studies of science system, of technology system and of science-technology interface. The focus here is -though not exclusively- on scholarly publications and patents, it embraces bibliometrics, scientometrics (Moed, Glanzel & Schmoch, 2004) and informetrics (Egghe & Rousseau, 1990), more recently starting to consider also other non-scholarly and societal «altmetrics» dimensions (Cronin & Sugimoto, 2014).

• *Economics of innovation*, which is at the core of several different economic fields, including macroeconomics, industrial organization (strategies and interactions of innovative firms), public finance, policies for encouraging private sector innovation, and economic development (innovation systems and technology transfer) (Hall & Rosenberg, 2010).

• *Science of Science policy* (Fealing et al., 2011; National Academy of Science, 2014; Lane, 2011, 2014).

• *Science and Society interplay* (Etzkowitz & Leydesdorff, 2000; Aghion et al., 2009; Helbing & Carbone, 2012).

A neglected aspect within these streams of work is the building block of education. From the economics of education (Johnes & Johnes, 2004; Hanushek et al., 2011) we know that education is an investment in human capital analogous to an investment in physical capital. The missing link with previous streams of literature is people. People in fact carry out research and innovation activities; attend schools and higher education institutions, acquiring competences and skills. Here another link could be added with Dosi (2014).

Methods for the assessment of Research

The assessment of the performance of an activity can be carried out on its output, on its outcome (indirect output), on its productivity (partial or total factor productivity), on its efficiency, on its effectiveness, on its impact.

From a methodological point of view, a distinction between productivity and efficiency has to be done. Productivity is the ratio of the output/input. Efficiency, in the broad sense, is defined as the distance with respect to the frontier of the best performers (Daraio & Simar, 2007). The econometrics of production functions is different than that of production frontiers as the main objective of their analysis differs: production

functions look at average behaviour whilst production frontiers analyse best performers behaviour (Bonaccorsi & Daraio, 2004). Obviously, assessing the impact on the average performance is different than assessing the impact on the best performance. This distinction has been considered also recently in the theory of growth and in the managerial literature. From a methodological perspective, different approaches, both parametric and nonparametric (Badin, Daraio & Simar, 2012; Daraio & Simar, 2014) have been proposed.
On the other hand, classical methods of impact assessment (Bozeman & Melkers, 1993; Khandker et al., 2010) proved inadequate to the checklist of "sensitivity auditing" (Saltelli & Guimarães Pereira; Saltelli & Funtowicz, 2014).

Important role of data

The data dimension is characterized by a kind of "data paradox". On the one hand, we are in a "big data" world, with open data and open repositories that are exponentially increasing. On the other hand, in empirical applications «data constraints» are almost the same as those described in Griliches (1989, 1994).
We believe that a great improvement could come by the adoption of an Ontology-Based-Data-Management (OBDM) Approach (Calvanese et al. 2010; Lenzerini, 2011; Poggi et al., 2008) to integrate the heterogeneous sources of data on which the empirical analysis has to be carried out.

A framework for the analysis

A general framework to investigate and empirically assess the research activity and its impacts is derived integrating existing approaches according to three dimensions. The main building blocks of these dimensions are reported in Figure 1.

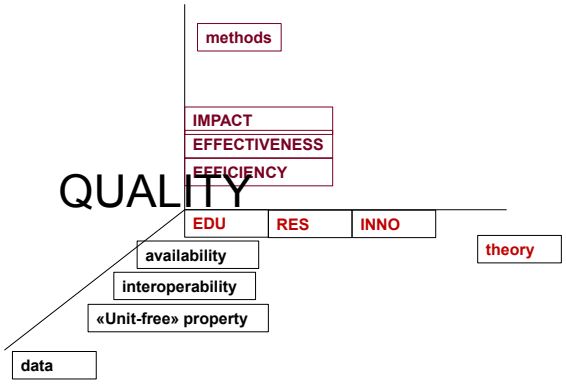


Figure 1. A framework for the analysis of research assessment and its impacts.

We propose "quality" as the overarching concept, which links together all the three dimensions. Quality should be declined along the three dimensions and by each building block. In theory, in education, a lot of progresses have been done. Much more work is needed for research and innovation. If we include quality indicators in the analysis we can move from efficiency to effectiveness. Moreover, it is the quality of education, research and innovation, which has an "impact" on the growth and development of the society. Finally, it is on the data dimension that the quality issues are of primary importance in all the three main building blocks proposed.
If we are not able to conceptualize and formalize in an unambiguous way the different meanings of «quality» for each building block proposed, we will not be able to make a real step forward in the empirical evaluation of the Efficiency, Effectiveness and Impact of Education, Research and Innovation. Third mission indicators (see Bornmann, 2013 for a survey) have a crucial role in this respect. It is indeed the role played by third mission indicators formally conceptualized as a measure of quality of higher education/research institutions, which can be used to investigate the Science-Society interplay.
For the conceptualization and formalization of the «quality» dimensions we suggest to adopt a very different approach based on: 1. Knowledge infrastructure (Edwards et al., 2013); 2. Convergence as «the coming together of insights and approaches from originally distinct fields», which «provides power to think beyond usual paradigms and to approach issues informed by many perspectives instead of few» (National Research Council, 2014).
We need to develop a knowledge infrastructure to model research and innovation and all the activities related to their (economical and societal) impacts in a systemic way. To advance towards an "open science" we have to build a common platform that has to be able to show us which data is relevant for assessing the model we selected for the analysis. In this way, the data could be analysed under different perspectives while sharing the same common conceptual characterization.

Selected References[1]

Aghion, P., David, P. A., & Foray, D. (2009). Science, technology and innovation for economic growth: linking policy research and practice in 'STIG Systems'. *Research Policy*, *38*(4), 681-693.

Bornmann L. (2013), What Is Societal Impact of Research and How Can It Be Assessed? A Literature Survey, JASIST, *64*(2), 217–233.

Daraio C. (2015), What do we know about Efficiency, Effectiveness and Impact of Research and Innovation? *Pro.of Workshop, 20th February DIAG Sapienza University of Rome*, edited by C. Daraio, Efesto Edizioni, Rome, pag. 13-25

Fealing K. H., Lane J. I., Marburger J. H. JIII, & Shipp S. S. (Eds.) (2011), *The Science of Science Policy, A Handbook.* Stanford, USA, Stanford University Press.

[1] The full list of references can be found at the author website.

Integrating Microdata on Higher Education Institutions (HEIs) with Bibliometric and Contextual Variables: A Data Quality Approach

Cinzia Daraio[1], Angelo Gentili[1] and Monica Scannapieco[2]

[1] *daraio@dis.uniroma1.it, angelo_gentili@hotmail.it*
Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, via Ariosto, 25 00185 Rome (Italy)

[2] *scannapi@istat.it*
Italian National Institute for Statistics (Istat), Rome (Italy)

An introduction on data quality

Data quality has been addressed in different research areas, mainly including statistics, management and computer science. The statistics researchers were the first to investigate some of the problems related to data quality by proposing a mathematical theory for considering duplicates in statistical data sets, in the late 60s. The management research began at the beginning of the 80s; the focus was on how to control data manufacturing systems in order to detect and eliminate data quality problems. Only at the beginning of the 90s, computer science researchers began considering the data quality problem, specifically how to define measure and improve the quality of electronic data, stored in databases, data warehouses and legacy systems. Data quality has been defined as "fitness for use", with a specific emphasis on its subjective nature. Another definition for data quality is "the distance between the data views presented by an information system and the same data in the real world"; such a definition can be seen as an operational definition, although evaluating data quality on the basis of comparison with the real world is a very difficult task.

Data quality is well-recognized as a multidimensional concept including several distinct dimensions (Batini & Scannapieco, 2006) proposed in various contexts (Catarci & Scannapieco, 2002). A crucial dimension of data quality is data accuracy: it measures the closeness between a value v and a value v', considered as the correct representation of the real-life phenomenon that v is intended to represent. However, quality is more than simply data accuracy. Other significant dimensions play a role in the definition of the Data Quality concept, including completeness, consistency, and timeliness (i.e. degree of up-to-dateness), just to cite some significant ones.

Data Quality issues in data integration processes

In a data integration system, sources are typically characterized by various kinds of heterogeneities that can be generally classified into:

(i) Technological heterogeneities.
(ii) Schema-level heterogeneities.
(iii) Instance level heterogeneities.

Technological heterogeneities are due to the use of products by different providers, employed at various layers of an information and communication infrastructure.

Schema heterogeneities are principally caused by the use of (a) different data models, such as one source that adopts a relational data model and a different source that adopts a graph-based data model, and (b) different data representations, such as one source that stores addresses as one single field and another source that stores addresses with separate fields for street, civic number, and city. Schema level heterogeneities can be solved according to well-defined methods that harmonize data collected by the different sources with respect to a schema global to the whole data integration system. However, from a practical perspective, in order to make such harmonization possible it is also necessary to solve (iii) instance level heterogeneities, namely:

For overlapping data sources, same objects can be represented as different due to data quality errors. Hence, in order to resolve such conflicting representations, an object matching activity must be performed. Such activity should be as much automated as possible, especially in complex data integration systems (Zardetto, Scannapieco, Catarci, 2010).

For all sources, also those that are not overlapping, a quality control at instance-level is very useful in order to prevent the possible population of the data integration system with erroneous data. Depending on the specific types of data integration systems, such a quality control can be performed in different ways.

A Data Quality Approach to integrate HEIs microdata in a platform

For a platform supporting European Universities for Education, Research and Technology Studies, on the one hand, the lower level of disaggregation of data makes them more sensible and increases the chances of instance-level errors. On the other hand,

data collection is performed by integrating data already collected by statistical institutions by means of different statistical surveys or administrative data.

Hence, the quality control activity should have the following features:

1. It has to be applied on the overall collected data and cannot be applied to single processes producing data. Monitoring and control of processes producing data can be very useful to prevent quality problems, however, it cannot be applied to our case, due to the different nature of production processes and to the practical impossibility to revise such processes in a preventive fashion. This does not exclude of course the fact that feedbacks deriving from quality analysis could be used by organizations that produce data to revise their production processes.
2. A specific quality activity of outlier detection could be applied, by comparing data provided by "similar" sources on the same subject. Here, "similar" could mean, for instance, belonging to the same country and with analogous features such as the number of personnel. Data that are recognized as outlier by automated procedures should subsequently undergo a human analysis. This analysis can either explain the outlier on the basis of available context information, or it can recognize that the outlier is actually caused by quality problems. In this latter case, quality improvement actions must be engaged.

The following Table 1 illustrates the main sources of data which have been integrated to test the data quality approach proposed in the paper.

Figure 1 instead shows an example of outliers detection carried out through a systematic check against different distributions. The check has been done on the ratios given by number of publications divided by the number of academic staff, for all European universities in the sample.

Table 1. Main sources of data integrated

Source (link)	Description
ETER (www.eter.joanneum.at/ imdas-eter/) integrated with data from HESA for UK	Microdata on inputs outputs of higher education institutions in Europe.
Scimago Institutions Rankings (www.scimagoir.com)	Bibliometric data on scientific production and impact.
Eurostat (http://ec.europa.eu/eurostat)	Contextual factors, data at territorial level on economic and social development.

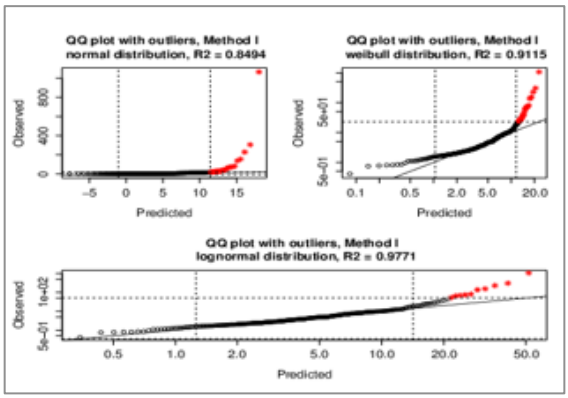


Figure 1. An example of outliers detection. Outliers are reported as stars in red: the graph top left shows outliers with respect to the normal distribution (worst fit, r-square=0.85), the one top right with respect to the Weibull distribution (r-square=0.91), the one below with respect to the lognormal distribution with the highest fit (r-square=0.98).

References

Batini, C., & Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies, and Techniques,* Springer, The Netherlands.

Catarci, T. & Scannapieco, M. (2002). Data Quality under the Computer Science Perspective. Journal of Archivi & Computer, 2.

Lepori, B., Daraio, C., Bonaccorsi, A., Daraio, A., Scannapieco, M., Gunnes, H., Hovdhaugen, E., Ploder, M. & D. Wagner-Schuster (2014), 'ETER Project, Handbook for Data Collection', Brussels, June.

Luwel, M. (2005). The use of input data in the performance analysis of R&D systems. In *Handbook of Quantitative Science and Technology Research* (pp. 315-338). Springer Netherlands.

Luwel, M. (2015), Heterogeneity of data in research assessment, in *Efficiency, Effectiveness and Impact of Research and Innovation*, Proceedings of the Workshop of the 20th February 2015, C. Daraio (ed), DIAG Sapienza University of Rome, Efesto Edizioni Rome.

Zardetto, D., Scannapieco, M., & Catarci, T. (2010). Effective Automated Object Matching. *Proceedings of the International Conference on Data Engineering (ICDE 2010)*.

Is the Humboldtian university model an engine of local development? New empirical evidence from the ETER database

Teresa Ciorciaro[1], Libero Cornacchione[1], Cinzia Daraio[1], Giulia Dionisio[1]

[1]*teresa.ciorciaro@gmail.com,* [1]*lillo-1991@libero.it,* [1]*daraio@dis.uniroma1.it,* [1]*giulia.dionisio@hotmail.it*
Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, via Ariosto, 25 00185 Rome (Italy)

Introduction

The higher education system, in advanced countries, has reached the point of massification (i.e. enrolment rates exceeding 50% of the relevant age cohort), while the public budget has not grown correspondingly. Universities are put under pressure to use existing resources, namely staff and funding, in the most efficient way. At the same time there is an increased pressure from the research side: the expectations of society and policy makers on the contribution of research to societal problems have grown significantly, there are new entrants in scientific arena (particularly from Asia) and the competition for funding has increased sharply. This situation creates a classical issue in public policy: we have two valuable goals (serving better mass educational needs and producing good research) between which there is tension or trade-off.

Do universities benefit from having inputs (staff and funding) that can produce jointly teaching and research, or there are efficiency-enhancing specialization effects that suggest to keep these activities under separate institutions? What is the impact of the environmental context of the universities? We focus here on the complementarity between teaching and research, which is at the core of the Humboldtian model of university (Schimank & Winnes, 2000). Is the traditional Humboldtian model of university, in which teaching and research are produced jointly by the same academic staff able to foster the economic development of the area in which the university is located? What are the main contextual factors which affect the performance of the European Humboldtian universities?

Purpose of the analysis and method

The main objective of this paper is to investigate the determinants of the efficiency scores of European universities, whose production is characterized by teaching and research outputs.

In efficiency analysis, nonparametric estimators are particularly attractive because they do not rely on restrictive parametric assumptions on the process that generates the data.

We apply a nonparametric approach, DEA (Data Envelopment Analysis, Charnes et al., 1978), which allows for multi-input - multi-output analyses, followed by a bootstrap analysis to estimate bias corrected efficiency scores and to provide confidence intervals on the efficiency scores. Given that universities in Europe face heterogeneous conditions, in a second step, we applied a semiparametric bootstrap-based approach (Simar & Wilson, 2007) to assess the statistical significance of external contextual factors on their performance.

Data and variables

Our sample is composed by 753 HEIs (Higher Education Institutions) belonging to 22 different European countries.

In the following tables we present the data analysed, the inputs, the outputs and the external factors investigated in the paper.

Table1. Data.

Data Source	Description
SCIMAGO INSTITUTION RANKING	The SIR purpose is a characterization of institutions, based on three different ranges: research, innovation and web visibility. This source uses normalized indicators, in a scale from 0 to 100, to facilitate the comparison between the institutions. The SIR database provides some bibliometric indicators for each institution, like number of publications, high quality publications, normalized impact, international collaboration and specialization index.
ETER	The European Tertiary Education Register wants to build a complete register of higher education institutions. Its database gives various information, like number of students, professors, graduates, doctorates, total incomes and expenditures. This register is developed by the Directorate General for Education and Culture of the European Commission.
EUROSTAT database	The EUROSTAT database wants to be the leading provider of high quality statistics on Europe. It contains regional data at a very disaggregated level.

Table2. Selected inputs

Input	Formula
Teaching	$\frac{\#\text{ of academic staff}}{\#\text{ of students}} * 100$
Structural	$\frac{\#\text{ of administrative staff}}{\#\text{ of students} + \#\text{ of academic staff}}$
Research	$\frac{\#\text{ of graduates at ISCED 8}}{\#\text{ of undergraduates enrolled}}$

Table 3. Selected outputs.

Output	Formula
Teaching	$\frac{\#\text{ of graduates}}{\#\text{ of students enrolled}}$
Research	$\frac{\text{output (pub)} * \text{HQP(\% high quality pub)}}{100 * (\#\text{ of academic staff} + \#\text{of graduates at ISCED 8})}$
Third mission	Percentage of third party funding.

Table 4. Selected External factors.

External factor	Description
GDP	Gross domestic product at current market prices
PAT	Patent applications
HOSP	Hospital yes/no
ER	Employment rates- age group 20-64
GERD	Total intramural R&D expenditure (GERD) at NUTS 2 level
SIZE	Size
AGE	No. of years from foundation

Modelling strategy

We estimate several partial models, i.e. models of single output production (teaching model, research model, third mission model) as well as complete models (of joint production of teaching and research, including also the third mission dimension) to analyse how the evaluation of the impact of external factors affects the production of the considered universities.

A correlation analysis is carried out to analyse the degree of association of the obtained efficiency scores with the degree of internationalization of the considered universities to account for recent results that show that is the quality of the academic staff that plays an important role to facilitate and faster third stream activities as complement of teaching and research missions.

Preliminary results and next steps

Figure 1 reports some illustrative preliminary results of the two-stage analysis conducted on the dataset.

We are going to extend the analysis in the following directions:

1. Inclusion of other third mission indicators in the input-output characterization (Geuna & Rossi, 2015), to investigate how their inclusion affects the impact of the considered external factors.
2. Apply robust nonparametric approaches (Daraio & Simar, 2007) which do not rely on the separability condition assumed by the two stage approach applied in this paper, and are more robust to outliers and extremes in the dataset as well as more flexible directional distance models (Daraio & Simar, 2014; Daraio et al., 2015a,b).

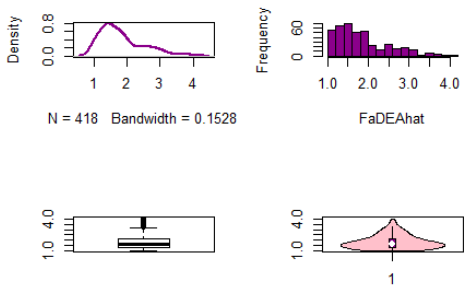


Figure1. Distribution of the European efficiency scores. Top left panel: nonparametric kernel density distribution, top right panel: histogram, bottom left panel: box plot and bottom right panel: violin plot.

Some Selected References[1]

Daraio, C., et al. (2011). The European university landscape: A micro characterization based on evidence from the Aquameth project. *Research Policy*, *40*, 148

Daraio, C., Bonaccorsi, A., & Simar, L. (2015a). Efficiency and economies of scale and specialization in European universities: A directional distance approach, *Journal of Informetrics*, *9*, 430-448.

Daraio, C., Bonaccorsi, A., & Simar, L. (2015b). Rankings and University Performance: a Conditional Multidimensional Approach, *European Journal of Operational Research, 244*, 918-930.

Daraio, C., & Simar, L. (2007). *Advanced robust and nonparametric methods in efficiency analysis. Methodology and applications*, Springer, New York.

Daraio, C., & Simar, L. (2014). Directional distances and their robust versions: Computational and testing issues. *European Journal of Operational Research*, 237, 358–369.

[1] See authors' webpage for a full list of references, which are removed due to space limitations.

Connecting Big Scholarly Data with Science of Science Policy: An Ontology-Based-Data-Management (OBDM) Approach

Cinzia Daraio[1], Maurizio Lenzerini[1], Claudio Leporelli[1], Henk F. Moed[1], Paolo Naggar[2], Andrea Bonaccorsi[3], Alessandro Bartolucci[2]

daraio@dis.uniroma1.it, lenzerini@dis.uniroma1.it, leporelli@dis.uniroma1.it, henk.moed@uniroma1.it
[1] DIAG, Sapienza University of Rome, via Ariosto, 25, Rome (Italy)

paolo.naggar@gmail.com,alessandro_bartolucci@fastwebnet.it
[2] Studiare Ltd.

a.bonaccorsi@gmail.com
[3] DESTEC, University of Pisa (Italy)

The OBDM approach in a nutshell

The key idea of OBDM is to resort to a three-level architecture, constituted by the ontology, the sources, and the mapping between the two. The ontology is a conceptual, formal description of the domain of interest to a given organization (or, a community of users), expressed in terms of relevant concepts, attributes of concepts, relationships between concepts, and logical assertions characterizing the domain knowledge. The data sources are the repositories accessible by the organization where data concerning the domain are stored. In the general case, such repositories are numerous, heterogeneous, each one managed and maintained independently from the others. The mapping is a precise specification of the correspondence between the data contained in the data sources and the elements of the ontology.

The main purpose of an OBDM system is to allow information consumers to query the data using the elements in the ontology as predicates. In this sense, OBDM is a form of information integration, where the conceptual model of the application domain, formulated as an ontology expressed in a logic-based language, replaces the usual global schema. The integrated view that the system provides to information consumers is not merely a data structure accommodating the various data at the sources, but becomes a semantically rich description of the relevant concepts in the domain of interest, as well as the relationships between such concepts.

Sapientia: a Platform for Developing Science of Science's Policy Models

We consider the building of descriptive, interpretative, and policy models of our domain as a distinct step with respect to the building of the domain ontology. The ontology will intermediate the use of data in the modelling step, and should be rich enough to allow the analyst the freedom to define any model she considers useful to pursue her analytic goal.

Obviously, the actual availability of relevant data will constrain both the mapping of data sources on the ontology, and the actual computation of model variables and indicators of the conceptual model. However, the analyst should not refrain from proposing the models that she considers the best suited for her purposes, and to express, using the ontology, the quality requirements, the logical, and the functional specification for her ideal model variables and indicators. This approach has many merits, and in particular:

- it permits the use of a common and stable ontology as a platform for different models;
- it addresses the efforts to enrich data sources, and verify their quality;
- it makes transparent and traceable the process of approximation of variables and models when the available data are less than ideal;
- it makes use of every source at the best level of aggregation, usually the atomic one (see examples in the following).

In this framework, exploratory data analysis, and the building of synthetic indicators, are only an intermediate step of the modelling effort that aims to the interpretation of behaviours, the explanation of differences in performance, the identification of causal chains of phenomena. That leads to the development of a policy-design model, whose inputs are policy instruments, and whose outputs are performance indicators for research activities and economic welfare.

The learning and theory building process requires feedbacks that could also concern the ontology level: the addition of new concepts and data, through the specialization of general concepts or the enlargement of the ontology commitment, could reflect the intermediate achievements of the

learning process such as the necessity of improvement of the theories submitted to test.

More often, however, a well-conceived ontology will resist to the competency test implied by new model and theories, and the most serious constraint to model development will be the impossibility of a complete mapping between the ontology and the sources, i.e. the lack of data. This is a negative result only for the short-term. In the medium and long term, the dialogue within the community of researchers that use the ontology as a workbench will result in a joint effort towards other stakeholders in order to improve detail, quality, and scope of data collection. Moreover, the shared use of logically sound definition for indicators increase the ability of the analysts to compare their studies and to test old and new theories.

Consider as an example the important issue of the assessment of the effects of scale economies on the performance of a research institution and of its affiliates. The results can widely differ if you set the analysis at different levels of aggregation: all the public research and education institutions of single countries, single universities, faculties, let's say, of Science and Technology, departments of Computer Science, research groups, or individuals within these groups.

Moreover, at different aggregation levels, the possible moderating variables or causes of different performances can widely differ. Legislation and regulation, public funding, teaching fees and duties matter at national level. Geography, characteristics of the local economic and cultural system, effectiveness of research and recruiting strategy, budgeting, infrastructures matter at the university or department level. Intellectual ability of researchers, history and stability of the group, ability to recruit doctoral students, worldwide network of contacts matter at the research groups and individuals level.

Time is a crucial dimension of research modelling. We pursue a modelling approach based on processes, i.e. collections of activities performed by agents through time. To represent the knowledge production activities, at an atomic level, we consider both stock inputs such as the cumulated results of previous research activities (those available in relevant publications, and those embodied in the authors' competences and potential), the infrastructure assets, and flow inputs as the time devoted by the group of authors to current research projects. Similarly, we can analyze the output of teaching activities, considering the joint effect of resources such as the competence of teachers, the skills and the initial education of students, and educational infrastructures and resources. Thirdly, service activities of research and teaching institutions provide infrastructural and knowledge assets that act as resources in the assessment of the impact of those institutions on the innovation of the economic system. The perimeter of our domain should allow us to consider the different channels of transmission of that impact: mobility of researchers, career of alumni, applied research contracts, joint use of infrastructures, and so on. In this context, different theories and models of the system of knowledge production could be developed and tested.

Conclusions

To bridge the gaps existing in the literature, and to integrate existing bottom-up initiatives in a coherent theoretical-based platform, we suggest an OBDM approach.

We need a change in the overall approach to the assessment of science and technology: metrics and indicators can have negative effects on the scientific community because they encourage a reductionist philosophy; on the contrary, we propose using well-defined concepts and data to build interpretative models, in order to compare and discuss theories. That can be useful both to promote a pluralistic community of analysts, and to build consensus on less superficial evaluation procedures of researchers and institutions. Moreover, indicators are often produced in closed circles, collecting ad hoc databases, with no built-in interoperability, updating and scalability features. We have to move towards an environment in which data are publicly available, collected and maintained on stable platforms, where ontologies give confidence on the precise meaning of data to people that propose models and to those that evaluate them. These repositories of knowledge can evolve following the analytical needs of the research community and the policy institutions, instead of starting from scratch each time a new research project starts. We propose our *Sapientia* ontology as a starting point to be opened, shared with the community and further developed and integrated with existing bottom-up initiatives as well as with new theories and paradigms.

Acknowledgments

Research support from the Progetto di Ateneo 2013 of the Sapienza University of Rome is gratefully acknowledged.

References

Daraio C., Lenzerini M., Leporelli C., Moed H.F., Naggar P., Bonaccorsi A. & Bartolucci A. (2015). Sapientia the Ontology of Multi-Dimensional Research Assessment*, Proc. ISSI*.

Fealing K. H., Lane J. I., Marburger J. H. JIII, & Shipp S. S. (Eds.) (2011), *The Science of Science Policy, A Handbook.* Stanford University Press.

Lenzerini M. 2011. Ontology-based data management, *CIKM 2011*: 5-6.

Poggi A., D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, & R. Rosati. (2008). Linking data to ontologies. *Journal on Data Semantics*, X: 133–173.

Incomplete Data and Technological Progress in Energy Storage Technologies

Sertaç Oruç[1], Scott W. Cunningham[1], Christopher Davis[2], Bert van Dorp[1]

[1] *s.oruc@tudelft.nl, s.cunningham@tudelft.nl, bertvandorp@gmail.com*
[1]Delft University of Technology, Faculty of Technology, Policy and Management, Jaffalaan 5 C2.010, 2628 BX Delft (The Netherlands)

[2] *c.b.davis@rug.nl*
University of Groningen, Center for Energy and Environmental Sciences (IVEM), Nijenborgh 4, 9747AG Groningen (The Netherlands)

Abstract
Energy storage is an important topic as many countries are seeking to increase the amount of electricity generation from renewable sources. An open and accessible online database on energy storage technologies was created, incorporating a total of 18 energy storage technologies and 134 technology pages with a total of over 1,800 properties. In this database information on technical maturity, technology readiness level and forecasting is included for a number of technologies. However, since the data depends on various sources, it is far from complete and fairly unstructured. The chief challenge in managing unstructured data is understanding similarities between technologies. This in turn requires techniques for analyzing local structures in high dimensional data. This paper approaches the problem through the use and extension of t-stochastic neighborhood embedding (t-SNE). t-SNE embeds data that originally lies in a high dimensional space in a lower dimensional space, while preserving characteristic properties. In this paper, the authors extend the t-SNE technique with an expectation-maximization method to manage incompleteness in the data. Furthermore, the authors identify some technology frontiers and demonstrate and discuss design trade-offs and design voids in the progress of energy storage technologies.

Conference Topic
Mapping and visualization

Introduction
High dimensional datasets are difficult to visualize contrary to two or three dimensional data, which can be plotted comparatively easily to demonstrate the inherent structure of the data. To aid visualization of the structure of a dataset, a family of algorithms have been devised in the literature, which are collectively referred as dimensionality reduction algorithms, of which an extensive review can be found in (van der Maaten, Postma, & van den Herik, 2009).

Among these algorithms *t-stochastic neighborhood embedding* (t-SNE) is a novel machine learning technique that has burgeoning applications. t-SNE maps each data point in a given high-dimensional space to a low-dimensional space, typically to a two or three dimensional one, for visualization purposes. The algorithm does a non-linear mapping such that similar points in the high-dimensional space situated nearby each other in the low-dimensional space as well.

In its first stage, the algorithm constructs a probability distribution over pairs of high-dimensional points in such a way that similar points have a high probability of being picked. In the second stage, it constructs the same probabilities between these points in the low-dimensional space. Finally the algorithm minimizes the difference between these probabilities by minimizing Kullback-Leibler divergence between these two distributions (Van der Maaten & Hinton, 2008).

Inherently, the algorithm preserves the manifold that possibly exist in the high-dimensional data and represents this manifold in low-dimensional space. Indeed, this class of dimensionality reduction algorithms is called "manifold learning". In comparison to the more

conventional, linear dimensionality reduction techniques such as *principal component analysis* (PCA), which finds a linear mapping with an objective to find a subspace where the projection of each data point lies as close to the original point as possible, manifold learning algorithms preserve the distance between pairs of points. Because of this the manifolds are preserved as well, whereas with PCA, clusters that are far from each other in high-dimensional space might be merged in low dimensional space.

t-SNE also proves to be useful for technology analysts in monitoring target technologies. Technologies such as batteries and storage, which is the target technology in this article, have multiple characteristics that develop over time. The problem facing the analysts is that most modern data sources are unstructured in character. Unstructured data often indicates that the data is of mixed provenance and quality. Furthermore, readily available data is often a mix of actual performance results, and forecasts of potential future results. Even when performance data is available the data is rarely standardized, and therefore contains incomplete and uncertain data.

Table 1. List of technologies in the database.

Compressed Air Energy Storage (CAES)	Nickel–cadmium (NiCd) battery
Edison (NiFe) battery	Nickel-metal hydride (NiMh) battery
Flow batteries	Nickel–zinc (NiZn) battery
Flywheels	Pumped Hydro
Hydrogen storage	Saltwater (sodium-ion) batteries
Lead-acid battery	Sodium-sulfur (NaS) battery
Lithium–air (Li-air) battery	Supercapacitors
Lithium-ion (Li-ion) battery	Superconducting magnetic energy storage
Lithium–sulfur (Li-S) battery	Zinc-air battery

Table Table 1 shows typical sources used in appraising technological development. The data varies by provenance – it is provided through a mix of academic, commercial, government, non-profit and media organizations. Furthermore, the data itself pertains to technologies at different stages of development, and in different modes of deployment or development. An exemplary data source, discussed in the next section, compiles research and development data concerning storage and battery technologies.

Despite the mixed quality of the data sources, such data is useful and should be incorporated into quantitative analyses. In this paper we are primarily concerned with technometric approaches to modelling technology (Coccia, 2005). In particular we are concerned with utilizing such data to produce technological frontiers. Such frontiers are useful for anticipating the future rate of growth, and can be used for developing coordination mechanisms such as technology roadmaps (Phaal, Farrukh, & Probert, 2004).

Evidence and belief need not be mutually incompatible. Bayesian statistical techniques acknowledge that data is often collected in an open, rather than controlled, experimental framework (Gill, 2004). As a result the necessity for belief prevails in the collection of data. There are beliefs concerning the quality of data, the underlying system relationships, and the nature and number of underlying cases to be measured. What is significant then is that prior beliefs concerning the data are acknowledged, that these beliefs actually encompass the true state of the world, and that these beliefs are consistently updated in light of new data. These are requirements which are achievable given the appropriate collection, treatment, and handling of mixed data.

What is required therefore is a technique for handling complexly structured data, for judging cases and similarities, and for managing incomplete data. This paper approaches the problem through the use and extension of *t-stochastic neighborhood embedding* (t-SNE). The technique is used to develop a non-linear manifold of technological performance, and to use this manifold to manage incompleteness in the data. This builds on a long-established technique for handing missing data known as the expectation-maximization procedure (Dempster, Laird, & Rubin, 1977). In the next section, the paper details a database of storage and battery technologies. In the subsequent section, a method is proposed for dealing with this semi-structured data, and in specific, for dealing with uncertain and incomplete technological information.

Data Sources

This work builds upon data collected from Enipedia,[1] a website that collects, organizes and visualizes open data related to energy systems. One of the initiatives on the website has focused on gathering information related to energy storage technologies.

Energy storage is an important topic as many countries are seeking to increase the amount of electricity generation from renewable sources. An issue with renewable energy is that the amount of generation is often variable and can exceed or fall short of the amount that is demanded. If there is an excess of production, then not all of the electricity can be fed into the grid. If there is an undersupply, then power plants relying on fossil-fuels must often be relied on in order to help meet demand. To address this variability, large-scale energy storage could be used to store energy during periods of excess renewable electricity production, and then supply this energy during periods of increased demand.

A key problem is that large-scale energy storage does not currently exist, aside from pumped-storage hydroelectricity plants which can only be built in locations with suitable geography. The development of economically feasible large-scale energy storage technologies will be a major game changer in the energy sector as it can support a larger integration of renewables and decrease reliability on electricity generation from fossil sources.

The research indicated that a number of energy scenarios and simulations fail to include models on energy storage, and lack accurate data on technologies. Also, forecasting is often not included, while battery technologies and costs are rapidly evolving. By these needs, an accessible and open technology database was created, incorporating a total of 18 energy storage technologies and 134 facilities or technology pages with a total of over 1,800 properties. In this database,[2] information on technical maturity, technology readiness level and forecasting is included for a number of technologies.

An overview of sources of technology information on the potential and future demand for energy storage indicates that a number of technologies and solutions focus on applications with small time-scales, such as frequency and voltage control, load shifting, diurnal storage, output smoothing, mobility and reserve grid capacity. Far few technologies and facilities focus on providing seasonal and large-scale grid storage. For a number of these technologies, installations with a lower technology readiness level have been included to provide some numbers on feasibility.

Developing metrics on comparing these technologies was done through an iterative design scheme, incorporating metrics relevant to a range of applications. It was observed that a number of technologies cannot be described fully, as information is missing or the ranges in which information sources report the information are exceptionally wide. Also, the definitions found for some technologies, such as Li-ion, are weaker than those found for other

[1] http://enipedia.tudelft.nl

[2] http://enipedia.tudelft.nl/wiki/Electricity_Storage

technologies. Furthermore, metrics are often made available on a systems level, and information on other levels needs to be translated to this system level.

Table 2. Variable number, name and description

No.	*Variable Name*	*Description*
1	Case	Case number
2	Product	Product name
3	Technology	Technology type
4	Year	Reference year
5	Institutional Data	Indicator whether observation is institutional
6	Technology Readiness Level[3]	Technology maturity level
7	Investment per Unit Power	Investment unit power (EUR/KW)
8	Investment per Unit Energy	Investment cost per unit energy (EUR/KWh)
9	Efficiency	Energy efficiency
10	Cycles	Life span in cycle times
11	Energy Density	Energy density (WH/L)
12	Power Density	Power density (WH/Kg)
13	LCoE[4]	Levelized cost of energy

Method

The chief challenge in managing unstructured data is understanding similarities between technologies. This in turn requires techniques for analysing local structures in high dimensional data. The technique of choice for this is t-stochastic neighborhood embedding (van der Maaten & Hinton, 2008). Finding a manifold which represents the data is useful for developing lower dimensional representations of the data. Such a manifold is inherently non-linear, and by necessity it preserves the local structures in the data at the expense of finding any global structures which might be present. For this analysis we adopt an implementation of the algorithm created in Matlab (van der Maaten, 2007).

The t-SNE technique has previously been used in technometrics. Cunningham and Kwakkel (2014) investigate a case of electric vehicle and hybrid electric vehicle designs and technologies. The case benefitted from the use of a non-linear fitting technique since the designs considered differ substantially in fundaments. As a result different designs highlight fundamentally distinct kinds of engineering trade-offs. The case also demonstrated a potential convergence across multiple technologies. Other patterns of technological evolution on a manifold, in addition to convergence, are identified in the paper.

Other technometric approaches utilize a linear, or quasi-linear technological frontier. Many of these approaches also assume a constant rate of technological change as the frontier advances over time. These alternative approaches are useful for single technologies with well-understood morphologies. Such techniques are also suitable for technologies where there are suitable indicators of performance, outcome, or merit. The techniques are less useful for analyzing broader fields with a heterogeneous base of technology. In such fields different technological trade-offs may be at work, and the pace of technological change may be discontinuous or punctuated. Indeed, the technologies themselves each may be valued for different purposes and outcomes.

[3] http://en.wikipedia.org/wiki/Technology_readiness_level
[4] http://en.wikipedia.org/wiki/Cost_of_electricity_by_source

A desirable method must be suitable for use with diverse data types. Before applying t-SNE to the data set of Table 2, the data is first transformed and normalized. Transforming the data eases a search for locally similar data points. Furthermore, the normalization of the data helps address difficulties associated with variables being measured in different units, potentially highly discrepant in scale. The choice is made to take the logarithm of the data whenever the data is right skewed. Logistic transformation is used to create more normal-like distributions than the actual.

As previously noted, a major challenge in addressing such data sets is the presence of missing data. The principle technique for handling missing data in the statistical literature is known as the expectation-maximization procedure. This powerful technique has been extended to address the estimation of missing model parameters, as well as missing data, and later become a mainstay of machine learning techniques. Modern machine learning procedures are now availed of much faster algorithms than expectation-maximization procedures; nonetheless the technique has had a powerful effect on the field.

The expectation-maximization procedure consists of two steps. In the first, or expectation step, the missing data is replaced with an expected value. Initially the expected value can be set by the mean of the data, or even by replacing the missing data with random values. Then in the maximization step, a model of the data is selected and applied. After an initial modeling step, further estimates of expected values derived from the model can be derived. These expected values become new expected values for additional rounds of the modelling procedure. After repeated cycles of expectation and maximization the estimated values converge, and the full model of the data is derived. The technique has the benefit of replacing missing values with neutral values consistent with an assumed model of the data. The technique therefore makes the best use of available data that is possible, rather than excluding whole variables or cases because they are incomplete.

Unstructured data in this domain is not just incomplete, but also uncertain. This is expressed with reported ranges of expected performance data. In order to treat this data, an upper bound and a lower bound on the data is reported, using two distinct model variables. When the data is certain, the upper and lower bound of the variable is identical. In subsequent model runs a constraint is imposed on the expectation maximization procedure – the maximum estimated upper bound on missing data must be greater than the lower bound. When estimated variables do not satisfy this criteria they are either not updated, or both the upper and lower bounds are replaced with averages.

Every point on the manifold estimated by t-SNE is associated with a potential technological design. Thus the t-SNE model is generative – it reports the expected best fit to the data, and also anticipates new cases or designs which have not yet been reported. Nonetheless, technological constraints or other factors may mean that parts of the manifold are not populated with new designs. Interpolation using the manifold can proceed following two directions. A locally linear direction of change can be interpolated from the data given specific examples or cases. Or, a weighted average of surrounding points can be used given their relative proximity on the technological manifold.

Analysis

The following section details a complete procedure for analysis, as depicted in Figure 1Figure. The procedure begins with preprocessing the data. The raw data includes lower and upper bounds for various attributes. Thus, we made a choice to create two different features for each of such variables, e.g., both “Energy density lower bound” and “Energy density upper bound” features.

The next step identifies and masks out the missing data. The process is facilitated by the use of data structures (for instance in Python or Matlab) where the missing data is identified using

indicator values. A data matrix therefore contains two layers – the first layer stores the data itself, and the second layer contains a bit matrix for masking. The bit matrix indicates where the data is complete or non-missing, or incomplete and missing.

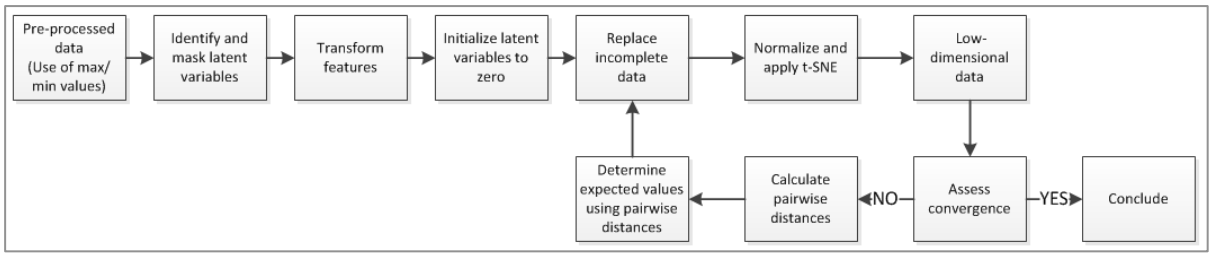


Figure 1. A Flow Chart of the Analysis Procedure.

Then the features are transformed and normalized to normal-like distributions. The following state initializes the missing variables to zero, which is in effect the mean of the normalized features. In subsequent iterations of the algorithm more refined estimates of the missing data are made. This brings us through the initialization and the first maximization step of the algorithm.

The data is complete, and can now be fitted using the t-SNE algorithm. The major output of the algorithm is a set of coordinates for all the cases – in this example there were 118 points. Intermediate outputs, such as data coordinates and scatter plots are produced.

Next, convergence of the algorithm is tested by comparing the current imputed high dimensional representation to the high dimensional representation of the previous iteration. Obviously this step is skipped for the first iteration.

If the algorithm has not converged, then pair-wise similarities between the points are evaluated as the next procedure. The purpose of this comparison is to determine the closest peers of any given technology. The basis for this comparison is the Euclidean distance between two points in the three-dimensional space as output from the t-SNE algorithm. The distance is then scaled according to the negative exponential of the squared distance between the two points. The total distance is then re-scaled to sum to 100% percent to create weightings for updating the originally missing variables in the data. The idea here is to calculate the new values for the missing data such that these values are closer to the related data points implied by the low dimensional data. Using pair-wise distances, a new expected set of values is established and finally the high dimensional representation is updated. The model converges when there is negligible differences between the consecutive imputed high dimensional representations.

Results and Visualization

This section discusses some results of the t-SNE analysis, visualizes and interprets some of the results, instead of all, due to space limitations, and displays the technologies according to their respective dates of introduction or their forecasted date of introduction. These colors suggest that the frontier of technological performance is gradually moving outward (to the upper right) over time. This is further illustrated in Figure 3.

Technological development, at least as measured by year of introduction is a somewhat noisy variable. Nonetheless, in Figure 3, we can qualitatively place three frontier lines. The first is dated 10 1985, the second to 2010, and the third to 2035. It seems plausible given the figure that the rate of technological change is higher among battery technologies than it is among storage technologies. This is demonstrated by the comparative “fanning out” of the battery technologies over time.

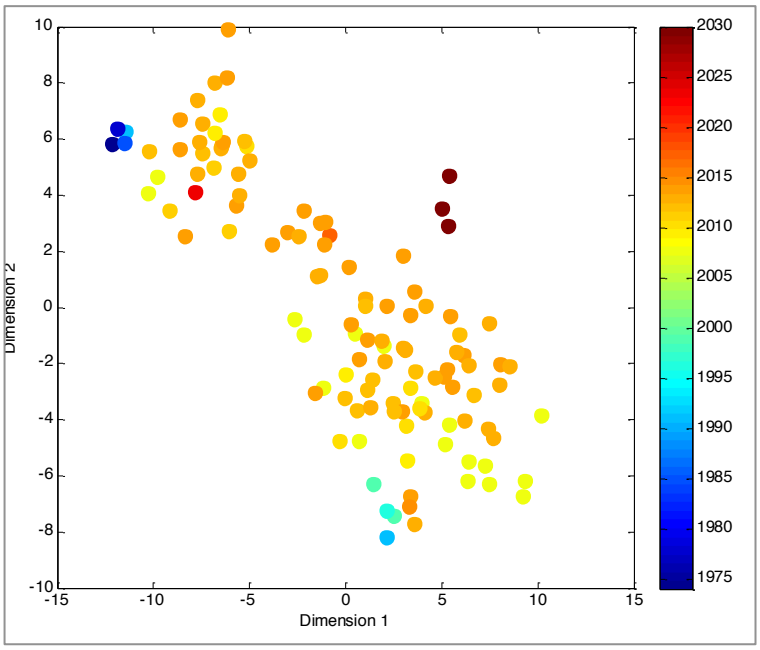


Figure 2. Technologies Positioned by t-SNE and Colored by Date of Introduction

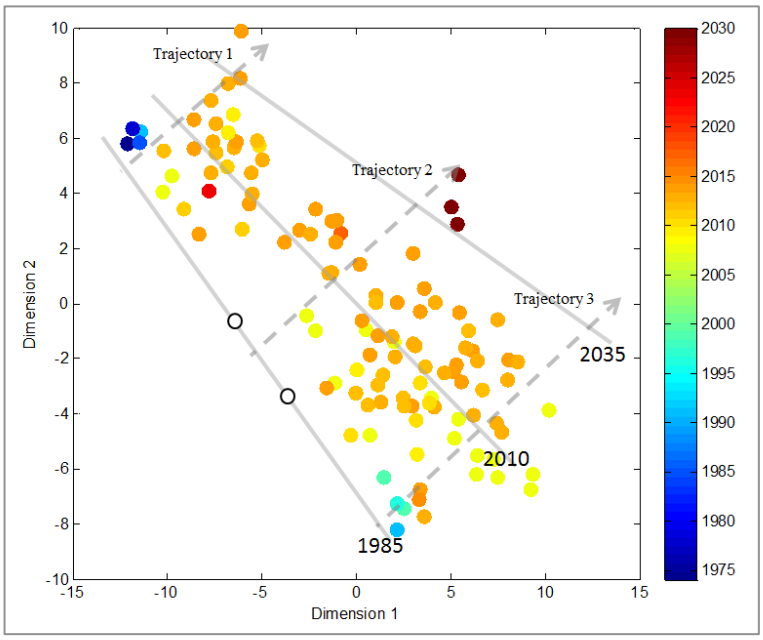


Figure 1. Technological Trajectories

In Figure 3 three technological trajectories are displayed. Changes in technological performance, based on benchmark technologies on or near the trajectory are calibrated. Then the three trajectories are compared with one another to determine whether there are common elements of change across the trajectories.

Figure 4 describes a potential trade-off in the design and selection of battery and storage technologies. In general the trade-off is between the respective cost and advantages of storage technologies versus batteries. Storage technologies are more robust, providing more cycles of operation at a lower levelized cost of energy. This comes at the cost of having a lower energy density, a lower technology readiness level, and a lower efficiency. In contrast battery technologies offer more energy density, are more readily available on the market, and operate

at a higher level of efficiency. In consequence, batteries are less robust, operating for fewer cycles, and requires a higher levelized cost of energy to be paid out.

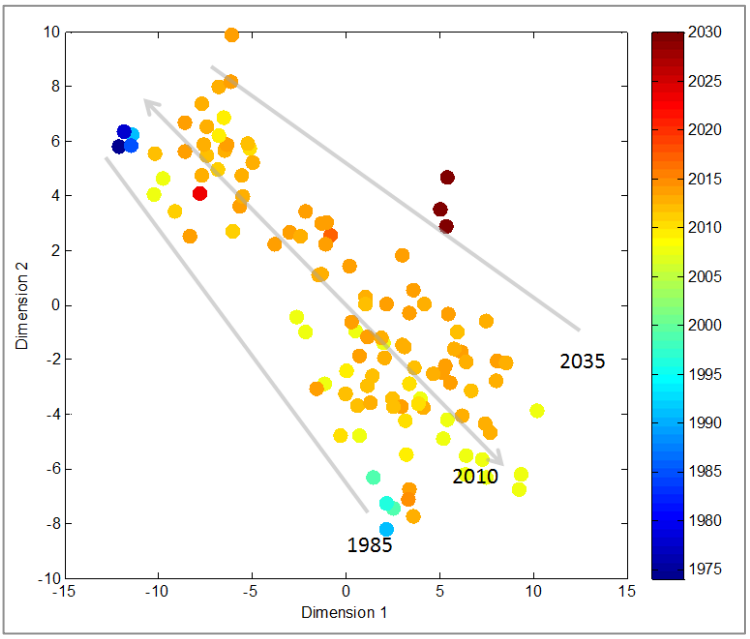


Figure 4. Design Trade-Offs.

There are three design voids on the manifold as shown in Figure 5. These are areas in the space of potential design which have not been explored. One space, design void 1, occurs along the 1985 technological frontier. The space is sparsely explored, although by 2010 a flywheel technology has emerged to occupy the space. The next two voids lie along the 2035 frontier. Because we are not yet on the 2035 frontier, these voids may be unanticipated breakthroughs. Design void 2 is in the space of high performing storage systems, and design void 3 is in the space of high performing batteries. One organization, EASE, anticipates a number of 2030 battery technologies on or beyond this frontier.

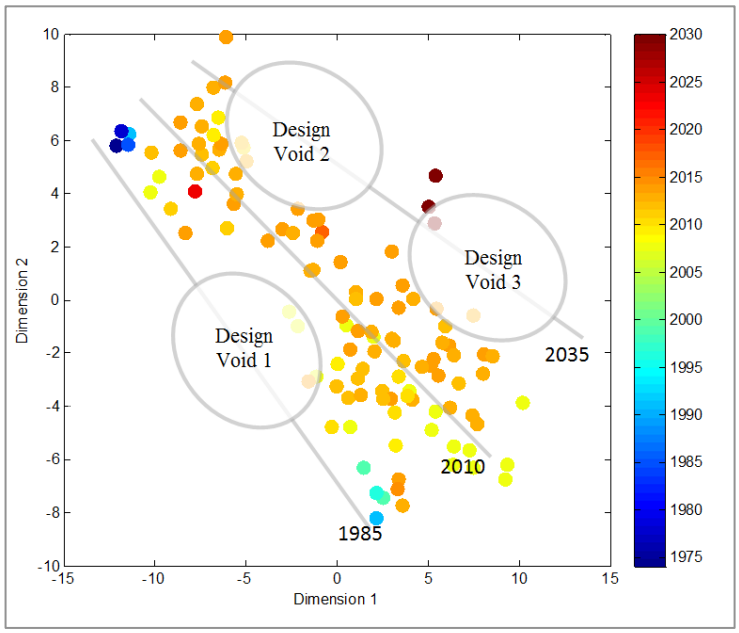


Figure 5. Design Voids.

Table 1. Historical and Emerging Designs.

	Void 1	*Void 2*	*Void 3*
Year	2013	2012	2030
InstitutionalData	0.01	0.79	0.99
TRL	8	6	9
Investment lowerbound	1,093	69	103
Investment upperbound	1,149	131	147
InvestmentEURperKW lowerbound	1,244	729	574
InvestmentEURperKW upperbound	1,262	1549	898
Efficiency lowerbound	0.767	0.709	0.785
Efficiency upperbound	0.849	0.809	0.847
Cycles lowerbound	4,265	11,306	3456
Cycles upperound	4,554	70,551	9804
EnergyDensity lowerbound	40	5	105
EnergyDensity upperbound	60	11	186
Power Density lowerbound	131	82	158
PowerDensity upperbound	220	210	295
LCoE lowerbound	0.149	0.074	0.056
LCoE upperbound	0.506	0.224	0.123

Table 3 provides, by interpolation, the performance characteristics of the technologies in the three voids mentioned previously. The exemplary void 1 technology is most likely a battery. The year of introduction suggests that there have been too few lower technology exemplars, so that the performance here is likely highly overstated. There should likely be a lower power and energy densities, and a lower levelized cost of energy. The closest existing technology is the "Wemag AG Li-Mn storage plant."

The void 2 technology, likely a storage device, should afford dramatically reduced investment and investment per kilowatt hour over previous technologies. The cycle times should be up to an order of magnitude higher than the void 1 exempla. While the power density may not be affected much from its 1985 peer, the energy density is likely to be reduced. The levelized cost of energy may be half of the previous levels of the void 1 technology. The year of introduction is too early, suggesting still higher energy and power densities over those listed. The closest existing technology is an advanced compressed air energy storage device.

The exemplary void 3 technology is most likely a battery. It will require an order of magnitude less unit investment, although the investment in terms of euros per kilowatt may be up to one half of previous levels. Cycle times will be improved, and energy densities may be doubled or even tripled over previous technologies. Power densities will also be somewhat improved. The levelized cost of energy will be three or four times lower than the equivalent technologies from 1985. The technology as anticipated is closest to some of the forecasted lead-acid battery advances for the year 2030.

Conclusions

In this paper, a database of energy storage technologies with various corresponding attributes is examined. The authors described a method to manage incompleteness of the data. The described method synthesizes t-SNE technique, which is a novel dimensionality reduction technique, with long-established expectation-maximization technique. The completed database later used for building a technology frontier that shows the progress of technology in

time, discussing the design trade-offs in the technology and finally identifying some design voids in the progress of the technology.

The technique described in this paper can be complementary to wide variety of technometrics or evolutionary technology dynamics approaches which make use of high dimensional technology data.

The technique performs better especially in visualization than other dimensionality reduction applications such as feature selection or feature extraction for two reasons. Firstly, it uses expectation maximization to impute the missing variables, which manages the incomplete data in such a way that the imputed variables have minimal weighting in producing the low dimensional map. Hence, it has least effect on the derivation of the lower dimensional map. Secondly, the t-SNE technique itself is a more suitable approach compared to other dimensionality reduction algorithms such as incumbent Principal Component Analysis (PCA). PCA aims to keep variation in the data and does not care about the pairwise relationships between data points, whereas manifold learning techniques such as t-SNE performs better in keeping similarities.

As a follow up to this work, more applications of this techniques next to the technology trajectories and design voids, as showcased in this paper, are yet to be explored. The promise of this technique is its complementary position in various technometrics analysis, which is yet to be fulfilled.

Furthermore, a methodological study regarding the validation of the technique using controlled experiments on a complete data set is on the research agenda of the authors.

Acknowledgement

Authors thank *Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities (BYTE)* project for funding this research.

References

Coccia, M. (2005). Technometrics: Origins, historical evolution and new directions. *Technological Forecasting and Social Change, 72*(8), 944-979.

Cunningham, S. W. & Kwakkel, J. H. (2014). *Technological frontiers and embeddings: A visualization approach.* Paper presented at the International Conference on Management of Engineering & Technology (PICMET), 2014 Portland, Oregon.

Delft University of Technology. (2014). Enipedia. Retrieved June 20, 2015 from: http://enipedia.tudelft.nl.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B, 39*(1), 1-38.

Gill, J. (2004). *Bayesian Methods: A Social and Behavioral Sciences Approach* (Third ed.). Boca Raton, FL, USA: Chapman & Hall / CRC Press.

Phaal, R., Farrukh, C. J. P., & Probert, D. R. (2004). Technology roadmapping -- A planning framework for evolution and revolution. *Technological Forecasting and Social Change, 71*, 5-26.

van der Maaten, L. (2007). An introduction to dimensionality reduction using MatLab. *Report, 1201*(07-07), 62.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579-2605.

van der Maaten, L. J., Postma, E. O., & van den Herik, H. J. (2009). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research, 10*(1-41), 66-71.

Bibliometric Characteristics of a "Paradigm Shift": the 2012 Nobel Prize in Medicine

Andreas Strotmann[1] and Dangzhi Zhao[2]

[1] *andreas.strotmann@gmail.com*
ScienceXplore, F.-G.-Keller-Str. 10, D-01814 Bad Schandau (Germany)

[2] *dzhao@ualberta.ca*
University of Alberta, 3-20 Rutherford South, Edmonton, Alberta (Canada)

Abstract
This research-in-progress paper reports bibliometric characteristics that illustrate and give credence to the claim of the Nobel Prize committee that its 2012 Nobel Prize in Physiology or Medicine was awarded for a "paradigm shift". An all-author co-citation analysis (ACA) of stem cells research 2004-2009 provides an interesting characterization of this paradigm shift, which was triggered by a mid-2006 publication by the younger of the two 2012 laureates. In particular, while ACAs of 2-year time slices for the period consistently indicate the presence of a single cohesive subfield in which the "paradigm shift" occurred, with some fluctuation in membership throughout the period, an ACA of the entire six year period shows instead a closely interlinked pair of subfields, which on closer inspection turn out to represent the pre- and post-paradigm shift states of the same subfield. This bibliometric characterization also correctly identifies the name of the researcher primarily responsible for the paradigm shift, namely, Shinya Yamanaka, as that of the dominant post-shift cited author in that subfield. The relative lack of dominant figures in the subfield in the pre-shift period also underlines the area's pre-paradigmatic state of multiple conflicting and relatively unsuccessful research directions attempting to address a fundamental crisis in that field at that point.

Conference Topics
Mapping and Visualization; Citation and Co-citation Analysis; Methods and Techniques

Introduction
The 2012 Nobel Prize in physiology or medicine was awarded to John B. Gurdon and Shinya Yamanaka for having triggered, the latter with a discovery first reported in his mid-2006 publication (Takahashi & Yamanaka, 2006), "a paradigm shift in our understanding of cellular differentiation" (Nobel.org, 2012).

In the present paper, we report bibliometric evidence and characteristics for this paradigm shift. Results from this study may contribute to research that combines relational and evaluative citation analysis methods to extend the research problems that are addressed by citation analysis.

Methodology
We examined the evolution of the stem cell research during 2004-2009 through an author co-citation analysis (ACA) of three 2-year time slices using the same dataset as in Zhao and Strotmann (2011), which reported results from a study of the full 6-year time period. We adapted methods from that study.

The data set was constructed by retrieving about 60,000 full PubMed records of stem cell research articles published during 2004-2009 with MeSH heading "stem cells", enriched by their cited references from Scopus records corresponding to these PubMed records (Strotmann & Zhao, 2009). Automatic author name disambiguation was performed on this dataset (Strotmann, Zhao, & Bubela, 2009).

For each of the three 2-year time slices, the 200 most highly cited authors were identified by fractional author citation counting, and their exclusive all-author co-citation counts were

calculated (Zhao & Strotmann, 2008). An exploratory factor analysis with oblique rotation was performed on each of these co-citation matrices (SPSS Direct OBLIMIN) with the number of factors to extract determined by Kaiser's rule of eigenvalue greater than one. Only factor loadings greater than 0.3 were retained in the factor analysis results in order to focus on the most important relationships.

The visualization used here is similar to that in Strotmann and Zhao (2012), improving on the one introduced in Zhao and Strotmann (2008). It visualizes directly the results of a factor analysis, with authors as square, and factors (research specialties) as circular nodes. An author node is colored according to the factor that it loads most highly on in the pattern matrix result of the factor analysis. Node sizes are proportional to citations received (author nodes) or to the sum of member author citations weighted by each author's loading (factor nodes). The visualization merges information on both the pattern and the structure matrix results of the obliquely rotated factor model, using the latter for automatic layouting (Kamada-Kawai algorithm in Pajek) and the former for gray-scale values of lines that link authors to the factors that they load on. Interpretation of the factor nodes (i.e., research specialties identified) proceeded exactly as in earlier papers, by manually examining highly co-cited papers of authors that load highly on a factor.

Results

Figures 1-3 show the intellectual structure of the stem cell research field for three consecutive 2-year periods.

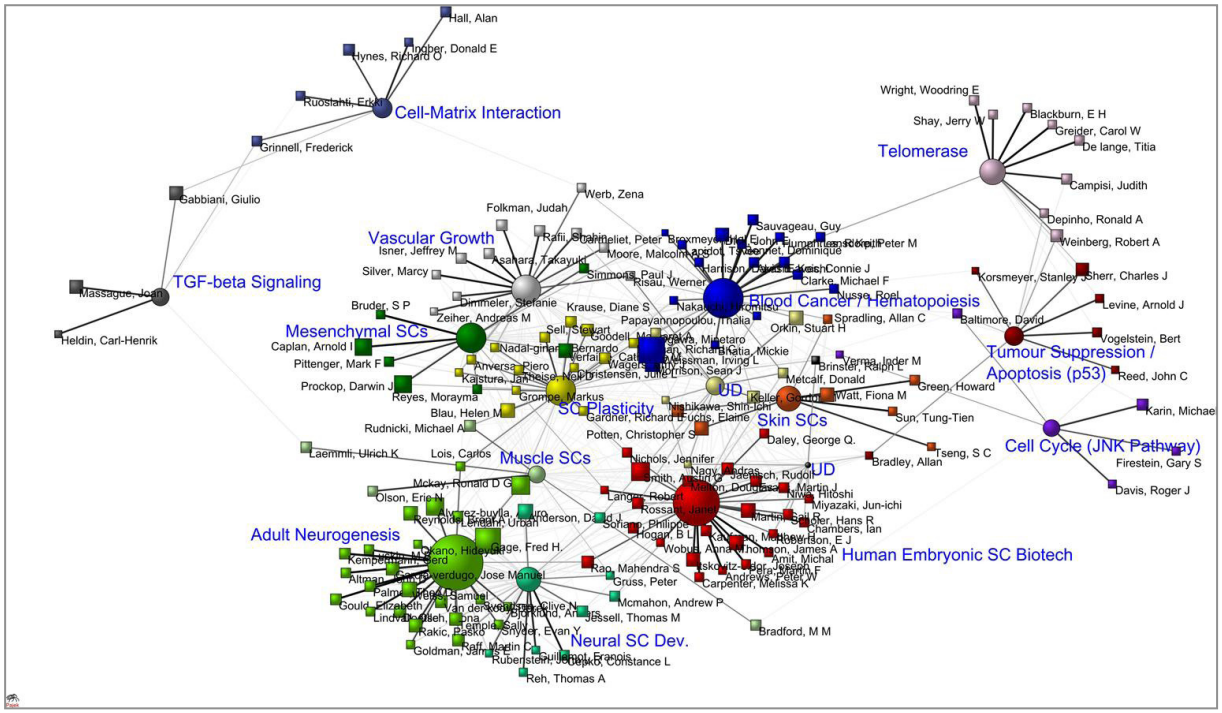


Figure 1. ACA of stem cell research 2004-05.

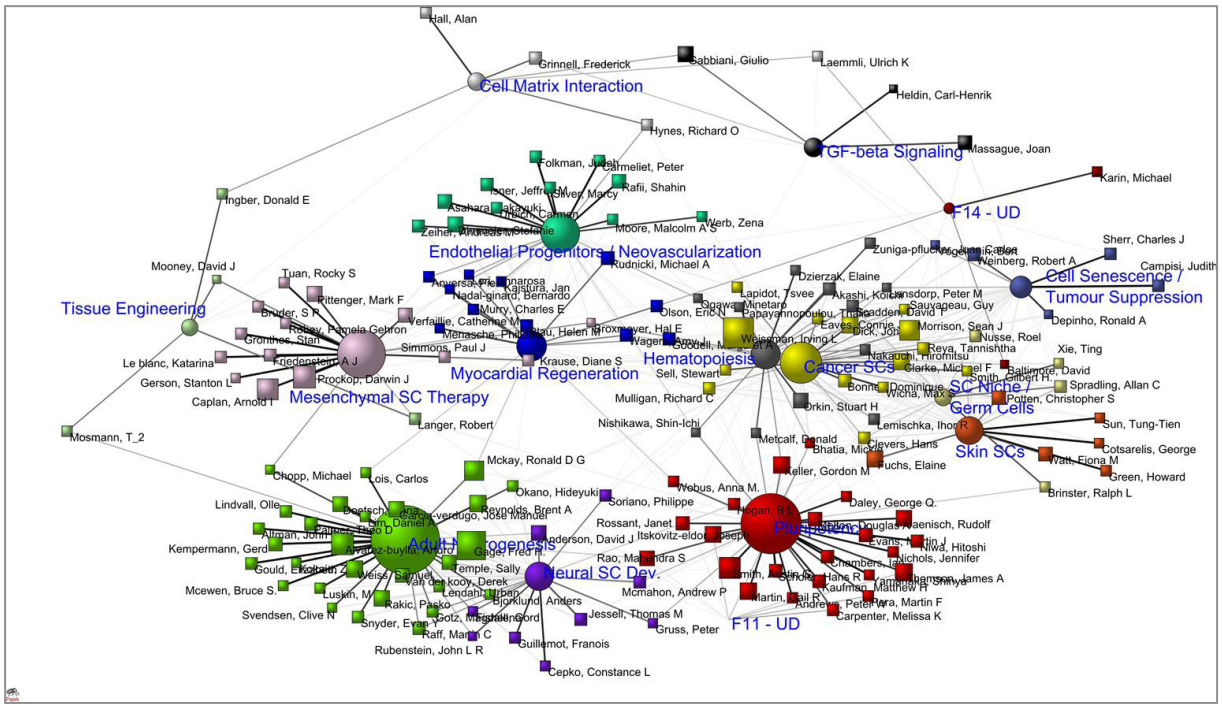


Figure 2. ACA of stem cell research 2006-2007.

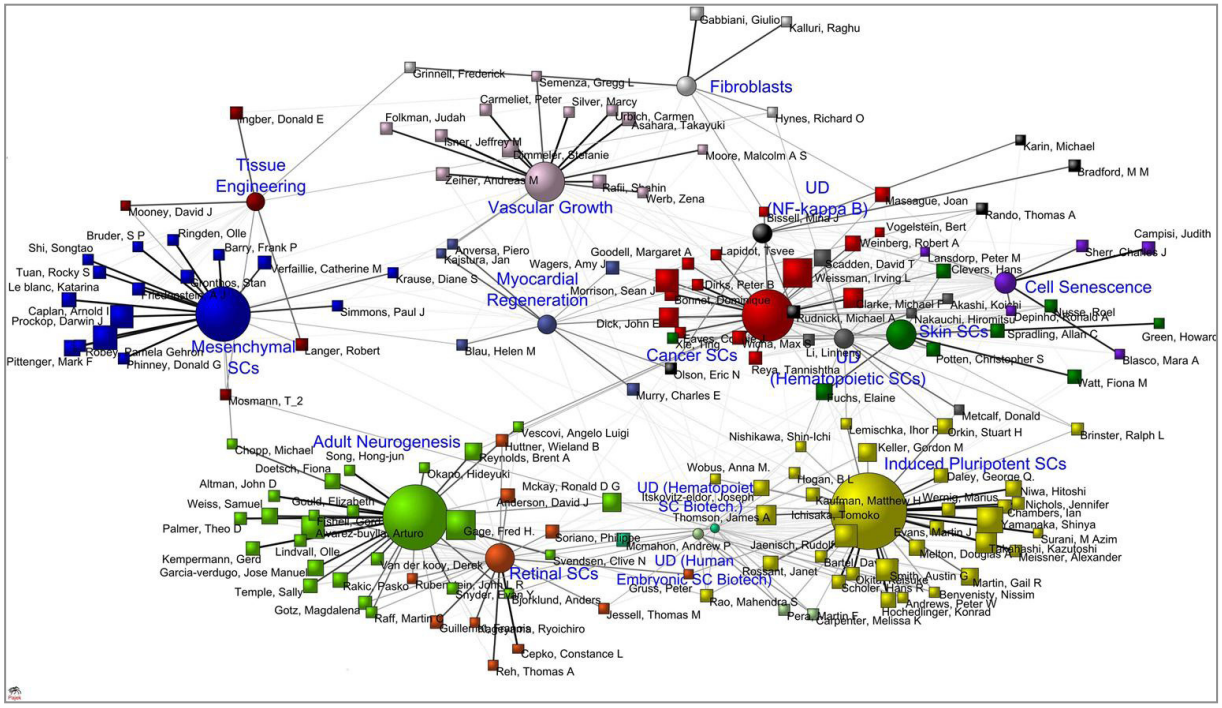


Figure 3. ACA of stem cell research 2008-2009.

While many interesting features of the international stem cell research field may be observed by examining these maps closely, we focus here on one particular major development in this field during the 2004-2009 time period as seen from changes over time. During the entire 2004-2009 time period, a subfield is shown prominently in the bottom right area of these maps as one of the two dominating specialties in stem cell research (the other being neural stem cells, bottom left). However, the entire focus appears to be shifting from (human) embryonic stem cell research in 2004-2005 (Fig. 1) through the study of pluripotency in

2006-2007 (Fig. 2) to the study of (human) induced pluripotent stem cells in 2008-2009 (Fig. 3). With this renewed focus on induced pluripotent stem cells, this subfield overtook the Neural stem cells specialty to become the most prominent specialty in the entire stem cell field in 2008-2009.

The transformation of this subfield is linked to the phenomenal rise of Shinya Yamanaka in these maps. Yamanaka was awarded the 2012 Nobel Prize in physiology or medicine for his discovery of induced pluripotent stem cells in mid-2006. He was not a highly influential researcher yet in 2004-05 as measured by citation impact (his name does not appear in Fig. 1); his name emerges in 2006-2007 (a small square in Fig. 2) and dominates this subfield by 2008-09 (the largest square in Fig. 3) with a citation impact reaching that of the two long-time most highly influential authors in the entire stem cell research field: Irving Weissman in the cancer stem cells specialty (red) and Fred Gage in the Neural stem cells area (green).

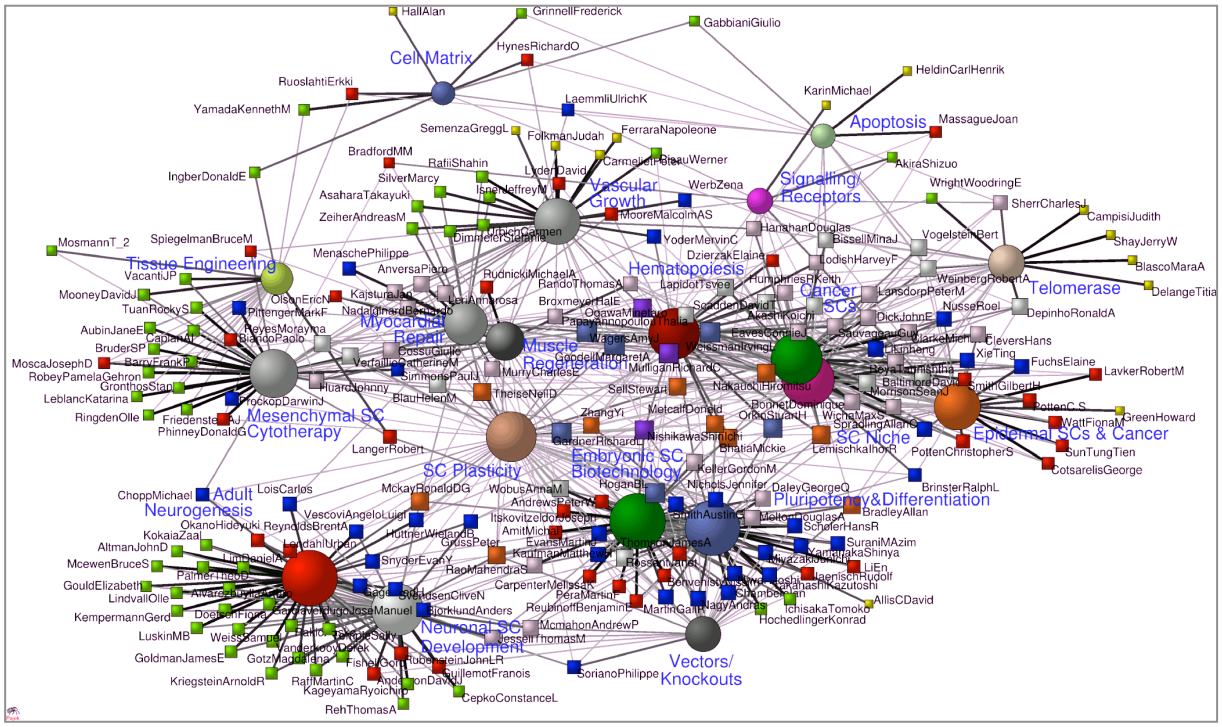


Figure 4. ACA of stem cell research 2004-2009.

By contrast, Figure 4, reproduced from (Zhao & Strotmann, 2011), which covered the entire 2004-2009 period in a single visualization, shows this subfield as consisting of two heavily interlinked research areas (bottom center), namely embryonic stem cell research (left, green) and (induced) pluripotent stem cell research (right, blue). This clarifies that what at first blush looks like it might have been a gradual change within this subfield when considering only Figures. 1-3 in fact constitutes a major in-place shift of research focus. Taken together with Figures 1-3, this confirms that the entire knowledge base for this subfield of stem cell research shifted from the former to the latter within just a couple of years of the publication of the key transformative paper – a true paradigm *shift* indeed. Most authors in this subfield co-loaded strongly on both these areas in the 6-year visualization, indicating a widespread realignment of researchers. A major paradigm shift becomes apparent.

Discussion

Kuhn's main criterion for a scientific revolution, or paradigm shift, is that something previously unthinkable becomes standard knowledge in a scientific field and a major crisis within the field is resolved as a result (Kuhn, 1970). In the case of stem cell research,

Yamanaka found that differentiated cells can be "reset" (induced) to undifferentiated (pluripotent) state, which essentially reverses the arrow of time in cell development biology, something previously unthinkable indeed.

It had been known in principle since Gurdon's 1960s paper (Yamanaka's co-laureate) that adult cells could be turned into even totipotent cells. For decades, stem cell research had been attempting to make this process feasible and controllable for therapeutic use, hoping someday to be able to regrow any type of damaged tissue (hence, the term regenerative medicine). The insurmountable research problem was a practical one: all methods for manipulating cells to this end produced stem cells that carried an unacceptably high risk of growing into malignant cancers rather than viable organs. Yamanaka's methods appear to have been the first (among uncountable failed attempts by others) to promise a fully viable resetting of cell development to the pluripotent or even totipotent state.

At the same time, Yamanaka's methods promised "safe", "natural", and abundant sources of pluripotent stem cells for research on early stages of cell development, which provided an immediate solution to a major social crisis that faced stem cell research in this subfield. This crisis came from the huge ethical and legal problems of obtaining and handling the embryonic stem cells that it required. By triggering a "natural" reset switch of much less problematic adult cells to the pluripotent state, as it were, the resulting stem cells not only side-stepped the ethically problematic use of embryos as a source, but did so without the kinds of major intervention such as genetic manipulation that had severely limited the usefulness of earlier versions of such cells for studying the "natural" biology of cell development.

As the Committee points out, Yamanaka's solution was also quite simple, so that human embryonic stem cell research was able to rapidly shift its entire focus to the study of induced pluripotent stem cells, in the remarkably short time of just a couple of years. Yamanaka's methods became standard knowledge very quickly – "textbooks were rewritten".

In the visualizations produced from an ACA of the type we performed here, this paradigm shift is characterized, somewhat paradoxically, by a stable visual appearance of the affected research subfield, accompanied by a shift in topic focus (factor labels). That a major topic shift took place can be confirmed through an analysis of a larger time slice spanning the triggering event, as we saw above. The initiator of the paradigm shift, Yamanaka, stands out as the author whose node shows explosive growth in citations received within the area as the shift occurs. The success of the paradigm shift is also seen from a rapid growth spurt of the shifting subfield relative to other subfields.

Interestingly, our visualization appears to also capture the "pre-paradigmatic" stage of this subfield, during which no single proposed solution managed to dominate the field (or subfield) that is undergoing a crisis (Kuhn, 1970). Unlike e.g. Gage in Neural stem cell biology or Weissman in bone marrow stem cell medicine research, whose citation impacts (indicated by relative node sizes) clearly dominated their respective subfields, no individual stood out in the embryonic stem cell research to that degree in Figure 1 (2004-2005). By 2008-2009, however, with the paradigm shift from embryonic to (induced) pluripotent stem cells as primary research tools completed, Yamanaka clearly plays that role in this area.

This ACA was actually performed, and Figures 1-4 were created, well before the 2012 Nobel Prize was announced (Strotmann & Zhao, 2011; Zhao & Strotmann, 2011). It appears that this paradigm shift could in principle have been identified and the 2012 Nobel Prize predicted through bibliometric studies of this kind (we did identify it as a "major development" of the field). Now that we have an idea what to look for, we could perhaps proactively look for patterns of this kind in bibliometric research in order to identify scientific breakthroughs and to make interesting predictions for major research awards. Research of this kind could enhance previous attempts to predict who among millions of scientists might qualify for the

honor of a Nobel Prize (Garfield & Malin, 1968) by combining relational and evaluative citation analysis methods to provide more convincing evidence.

Conclusions

This paper provides bibliometric evidence that the 2012 Nobel Prize in Physiology or Medicine was indeed awarded for a paradigm shift, through ACA of three consecutive 2-year time periods of stem cells research 2004-2009 compared to a single 6-year ACA for the same data. The success of this paradigm shift is seen on the ACA maps from the explosive growth in node size (citations received) of the researcher whose research initiated the shift, along with a complete shift of research focus in a subfield of stem cells research and a rapid growth spurt of this shifting subfield relative to other subfields. An ACA of the full period confirms that a major shift in the knowledge base of the subfield took place over this short time period; indeed, it shows signs of moving from a Kuhnian "pre-paradigmatic" to a "normal science" stage.

We hope that results from this study will contribute to research that combines relational and evaluative citation analysis methods to extend the research problems that are addressed by citation analysis.

Acknowledgments

This project was funded in part by the Social Sciences and Humanities Research Council of Canada.

References

Garfield, E., & Malin, M. (1968). Can Nobel Prize winners be predicted? *135th Annual Meeting, AAAS, Houston, Texas*.

Kuhn, T. (1970). *The structure of scientific revolutions*. Enlarged (2nd ed.). University of Chicago Press.

Nobelprize.org (2012). The 2012 Nobel Prize in Physiology or Medicine - Advanced Information. Retrieved June 2, 2015 from: http://www.nobelprize.org/nobel_prizes/medicine/laureates/2012/advanced.html

Strotmann, A. & Zhao, D. (2011). Evolution of stem cell research 2004-2009. A citation analysis perspective. *Stem Cells Europe. Edinburgh, 20.-21. July 2011*.

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63*(9), 1820-1833.

Takahashi, K. & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, *126*(4), 663-676.

Zhao, D. & Strotmann, A. (2008). Information Science during the first decade of the Web: An enriched author co-citation analysis. *Journal of the American Society for Information Science and Technology*, *59*(6), 916-937.

Zhao, D. & Strotmann, A. (2011). Intellectual structure of the stem cell research field. *Scientometrics*, *87*(1), 115-131.

Bibliometric Mapping: Eight Decades of Analytical Chemistry, with Special Focus on the Use of Mass Spectrometry

Cathelijn J. F. Waaijer[1] and Magnus Palmblad[2]

[1] *c.j.f.waaijer@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Faculty of Social and Behavioural Sciences, Leiden University, P.O. box 905, 2300 AX, Leiden (the Netherlands)

[2] *n.m.palmblad@lumc.nl*
Center for Proteomics and Metabolomics, Leiden University Medical Center, Postzone L4-Q, P.O. Box 9600, 2300 RC Leiden (the Netherlands)

Introduction[a]

Bibliometric mapping tools and other scientometrics analyses may be used to study the historical development of a research field. In our paper, we use automatic bibliometric mapping tools to visualize the history of analytical chemistry from the 1920s until the present, with special focus on the application of mass spectrometry (MS).

Data and methods

Co-word maps were based on noun phrases (nouns and preceding adjectives) parsed from titles and abstracts of all papers published between 1929 and 2012 by *Analytical Chemistry*, a key journal in the field of MS. Maps were constructed by determining the co-occurrence of noun phrases and visualized using VOSviewer software (Waltman & van Eck, 2010).

Results

Evolution of topics in analytical chemistry 1929-2012

Co-word maps were based on all texts published in *Analytical Chemistry* except for advertisements (1929-1995) or on all articles, letters and reviews published in *Analytical Chemistry* (1996-2012). Table 1 shows a summary of the different clusters in the co-word maps (due to space constraints, the maps themselves could not be included).

The maps show that inorganic chemistry has been an important topic within analytical chemistry for a long time; from 1929 until 1990 there were one or more clusters on inorganic chemistry. In the 1991-2000 period it was merged with the topics of electrochemistry and sensors. Much attention was given to (the development of) different apparatuses between 1929 and 198. A cluster on general and editorial issues can be found in almost every period. Topics that have developed over time include electrochemistry, chromatography and mass spectrometry. Electrochemistry shows up as its own cluster in the 1951-1960 period, but terms relating to the subject can also be found in the inorganic chemistry and metals cluster from 1941. This suggests the topic of electrochemistry has developed from inorganic chemistry and metals to form its own subfield. Chromatography is apparent in the maps from the 1951-1960 period onwards; mass spectrometry from the 1971-1980 period. The maps suggest the widespread use of mass spectrometry in analytical chemistry primarily developed through its coupling to chromatography; for the 1971-1980 period terms relating to mass spectrometry can be discerned in the maps, but the cluster is still dominated by chromatographic techniques and applications. However, from the 1981-1990 period, mass spectrometry broke off and formed its own subfield. Finally, from 2001 a cluster on separations and microfluidics emerged. This cluster also contains terms relating to theory and simulations (of such microfluidic systems).

Use of different techniques in analytical chemistry

Next, we analyzed the development and use of a number of techniques *within* analytical chemistry. As a proxy, we determined how many articles mentioned the technique in their titles during the 1929-2012 period. This shows that titration techniques reached their publication peak in the 1950s, gas chromatography in the 1960s, and liquid chromatography in the 1980s (Fig. 1). Of these techniques, only the latter was still mentioned in the titles of over 5% of papers published in the 2001-2012 period. On the other hand, microfluidics is an example of a technology not mentioned before 1990 that has really taken off in this 2001-2012 period. A technique not mentioned to a great extent in the titles of *Analytical Chemistry* papers is nuclear magnetic resonance (NMR). As the co-word maps already suggested, the mention of mass spectrometry increased throughout the entire period. Whereas in the 1929-1940 period none of the *Analytical Chemistry* papers mentioned mass spectrometry in their title, the percentage of papers that did increased to eighteen in the 2001-2012 period (Fig. 1). This indicates *Analytical Chemistry* has made a shift towards the publication of research using mass spectrometry instead of other techniques.

Table 1. Main topics in mass spectrometry within the field of analytical chemistry.

Clusters per period
1929-1940
Apparatuses
Inorganic chemistry
Gases
Industrial applications, hydrocarbons and food
1941-1950
Apparatuses
Inorganic chemistry: gases/halogens
Inorganic chemistry: metals
Industrial applications and hydrocarbons
Organic and food chemistry
General/editorial
1951-1960
Apparatuses
Inorganic chemistry: metals
Electrochemistry
Chromatography
General/editorial
1961-1970
Inorganic chemistry
Electrochemistry
Chromatography
General/editorial and "informatics"
1971-1980
Apparatuses
Inorganic chemistry
Gases
Electrochemistry
Chromatography
General/editorial
1981-1990
Inorganic chemistry
Electrochemistry
Chromatography
Mass spectrometry
General/editorial
1991-2000
Inorganic chemistry, electrochemistry and (bio)sensors
Chromatography
Mass spectrometry and proteomics
Electrophoresis
General/editorial
2001-2012
Mass spectrometry
Detection, electrochemistry and (bio)sensors
Small molecules and quantitation
Separations, microfluidics, and theory and simulations

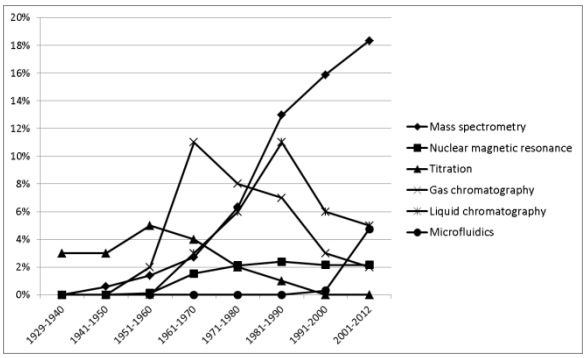


Figure 1. Use of different techniques in *Analytical Chemistry*. Search terms used were "mass spectro*", "nuclear magnetic resonance" or "NMR", "titration", "gas chromato*", "liquid chromato*", and "microfluid*", searched against the titles of *Analytical Chemistry* papers.

Additional work

Additional results, such as the trends in research topics in analytical chemistry research using MS, an assessment of which research fields use MS, and a citation network of research using MS, will be included on our poster.

Endnote

[a]A manuscript with the same title has been published in *Analytical Chemistry* as a Feature.

References

Waltman, L. & van Eck, N. J. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*, 523-538.

Introduction of "Kriging" to Scientometrics for Representing Quality Indicators in Maps of Science

Masashi Shirabe[1]

[1]*shirabe.m.aa@m.titech.ac.jp*
Tokyo Institute of Technology, Oookayama 2-12-1 S6-5, Meguro
152-8550, Tokyo (Japan)

Introduction

Maps of science are an effective technique, especially for non-experts, to facilitate intuitive understanding of science activities, even though they could be cut both ways. Among such maps, science overlay maps have received adequate attention from scientometrics researchers (Perianes-Rodríguez et al., 2011; Grauwin & Jensen, 2011; Klaine et al., 2012; Leydesdorff, Rotlo, & Rafols, 2012; Boyack & Klavans, 2013; Gorjiara & Baldock, 2014). Actually they are an attractive approach "to visually locate bodies of research within the sciences, both at each moment of time and dynamically." (Rafols, Porter, & Leydesdorff, 2010)

To produce science overlay maps, (1) we draw a basemap, which contains positional information of nodes from bibliographical data, then (2) we overlay other information on the basemap by assigning the information (i.e., indicators like publications and citations) to the nodes with such factors as colors and/or size of circles representing the nodes.

To think more abstractly, an essence of science overlay maps is "sharing" of positional information of nodes by different science maps, which are similar in concept to thematic maps in geography. What makes such "sharing" possible is the stability of global maps (Rafols, Porter, & Leydesdorff, 2010). This perspective could broaden choices of expressions in science overlay maps to improve our understandings. For example, VOSviewer (Van Eck & Waltman 2010) provides five different views, i.e., label view, density view, scatter view, cluster view, and cluster density view, for a fixed set of positional information of nodes. By switching these views, we can understand phenomena behind the maps deeply and multidimensionally. Therefore, introducing a new way to project bibliographical information on given maps is expected to expand availability of science overlay maps, just as a new method to produce thematic maps does in geography.

From this perspective, the author first pays attention to density view provided by VOSviewer. By mapping journals in the fields of Business, Business-Finance, Economics, Management, and Operations Research & Management Science, Van Eck and Waltman (2010, p. 529) explain functionality of the density view as follows: "The density view immediately reveals the general structure of the map. Especially the economics and management areas turn out to be important. These areas are very dense, which indicates that overall the journals in these areas receive a lot of citations." As they pointed out, this view is helpful to outline the macro structures of maps and to show which areas in the maps are important. Basically, however, density view can be used only for representing quantitative indicators, because "the item density of a point in a map depends both on the number of neighboring items and on the weights of these items." (p. 533) If citations were used as weights of items, the density map might be seen to show "quality" of areas. Actually, citation densities are only a representation of quantities. That is particularly evident in assuming to represent quality (impact) indicators like proportion of top 10% publications in the density view.

Judging from many scientometrics studies rely on density or heat maps (e.g., Pinto, Pulgarin, & Escalona, 2014), it would be reasonable to assume that graphical representations like the density view to represent quality indicators on science maps is very helpful to outline the structures of bibliographical data and to show which areas in maps of science are efficient, superior, or highly shared. Then, this paper introduces "kriging" to scientometrics for representing quality indicators.

Data

The author uses a data platform that consists of datasets from SCI Expanded, PubMed, and USPTO patent databases. By adopting matching methods developed in Shirabe (2014), records in PubMed are linked to those in SCI expanded, and non-patent references in the face sheets of US utility patents are also matched to records in SCI Expanded. As a result, three databases can be analyzed in an integrated fashion by using this platform.

This platform contains the product set (number of items is 8.5 millions) of SCI expanded (articles, reviews, letters, notes, and articles & proceedings papers; their database years are between 1992 and 2011) and PubMed (their publication years are between 1991 and 2012) as well as science citations of US utility patents registered between 1991 and 2012.

Method

First "macro and micro" basemaps are constructed by co-occurrence analysis of MeSH terms (Leydesdorff & Opthof 2013), where VOSviewer is used for mapping and clustering. For making the macro map, all the items of the product set are included in the analysis, and only third layer descriptors are treated as subjects of co-occurrence analysis. For that, lower layers' MeSH terms are replaced by their higher taxon. For making the micro map, only items containing mesenchymal stromal cells, mesenchymal stromal cell transplantation, totipotent stem cells, multipotent stem cell, induced pluripotent stem cells, pluripotent stem cells, and embryonic stem cells as their MeSH terms are included in analysis. Top 150 MeSH terms (except highly shared terms) are used in co-keyword analysis. Thus, this micro map is a map of pluripotent stem cell research.

Secondly, sets of data overlaying on the basemaps are produced. For that, positional data (i.e., two-dimensional position coordinate) of nodes produced by VOSviewer are transmitted to SAGA (Böhner, McCloy, & Strobl, 2006). Then, overlaying data for density maps (by Gaussian kernel function) or those for isograms (by kriging) are calculated from bibliographic indicators and overlaid on the basemaps.

Results

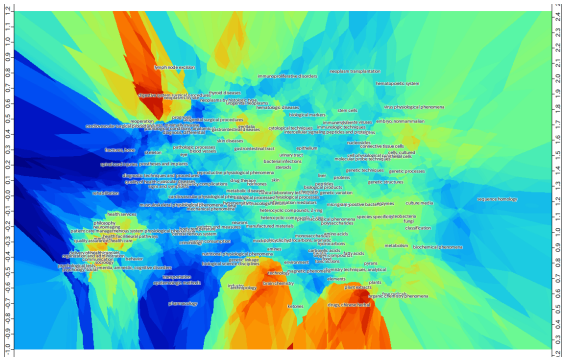


Figure 1. Japanese Share of Life-Science Papers cited by US Patents Registered between 2001-11.

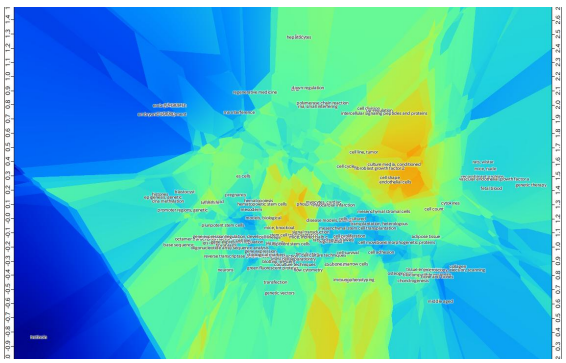


Figure 2. Japan's Relative Frequencies of Top 10% Cited Papers in Stem Cell Research.

The above figures show examples of overlay maps to represent quality indicators. They make it easier to understand the quality of Japanese research outputs intuitively and multidimensionally either at macro or micro level.

Acknowledgments

This research is partly supported by JST/RISTEX research funding program "Science of Science, Technology and Innovation Policy".

References

Böhner, J., McCloy, K.R., Strobl, J. (Eds.) (2006). *SAGA – Analysis and Modelling Applications*. Göttinger Geographische Abhandlungen, Vol. 115.

Boyack, K.W., & Klavans, R. (2013). Creation of a Highly Detailed, Dynamic, Global Model and Map of Science. *JASIS&T*, *65*(4), 670–685.

Gorjiara, T., & Baldock, C. (2014). Nanoscience and nanotechnology research publications: a comparison between Australia and the rest of the world. *Scientometrics*, *100*(1), 121-148.

Grauwin, S. & Jensen, P. (2011). Mapping scientific institutions. *Scientometrics*, *89*(3), 943-954.

Klaine, S. J., Koelmans, A. A., Horne, N., Carley, S., Handy, R. D., Kapustka, L., Nowack, B., & von der Kammer, F. (2012). Paradigms to Assess the Environmental Impact of Manufactured Nanomaterials. *Environmental Toxicology and Chemistry*, *31*(1), 3-14.

Leydesdorff, L., & Opthof, T. (2013). Citation analysis with Medical Subject Headings (MeSH) using the Web of Knowledge: A new routine. *JASIS&T*, *64*(5), 1076–1080.

Leydesdorff, L., Rotlo, D., & Rafols, I. (2012). Bibliometric Perspectives on Medical Innovation Using the Medical Subject Headings of PubMed. *JASIS&T*, *63*(11), 2239–2253.

Perianes-Rodríguez, A., O'Hare, A., Hopkins, M. M., Nightingale, P., & Rafols, I. (2011). Benchmarking and visualising the knowledge base of pharmaceutical firms (1995-2009), *Proceedings of ISSI 2011*, Vol. II, 656-661.

Pinto, M., Pulgarin, A., & Escalona, M. I. (2014). Viewing information literacy concepts: a comparison of two branches of knowledge. *Scientometrics*, *98*(3), 2311-2329.

Rafols, I., Porter, A. L. & Leydesdorff, L. (2010). Science Overlay Maps: A New Tool for Research Policy and Library Management. *JASIS&T*, *61*(9), 1871–1887.

Shirabe, M. (2014). Identifying SCI covered publications within non-patent references in U.S. utility patents. *Scientometrics*, *101*(2), 999-1014.

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for

bibliometric mapping. *Scientometrics*, *84*(2), 523-538.

The *Technology Roots spectrum*: a New Visualization Tool for Identifying the Roots of a Technology

Eduardo Perez-Molina[1]

[1]*eperezmolina@icloud.com*
CTB, Universidad Politecnica de Madrid (Spain)

Introduction

The purpose of this work is to present a new tool for identifying the technological foundations, or roots, of a specific technology in the whole range of existing technologies. The idea is to go back to the date before a specific technology existed as such—its origin date—and to evaluate the influence of every existing technology in relation with it. Our tool is based on the role played by prior art patent citations as a historical footprint. The documents cited in the prior art search reports by patent examiners against patent applications in a particular —new—technology link the new emerging techniques to the conventional existing ones. The nature of this particular set of references, namely *who* produced the citations—the patent examiner in place of the author—and *why* they are cited—the evaluation of the novelty and non-obviousness—, is unique within the body of bibliographic references (Meyer, 2000), and explicitly points to temporal and conceptual proximity. These two factors seem fundamental to the study of history and technology.
The *Technology Roots spectrum* (*TR spectrum*) is a tool for visualizing the components at the origin of the specific technology under study, showing their relative weight as bars in a graph containing the whole range—the spectrum—of technologies. It uses the computer to exploit the network formed by prior-art citations in patent publications and the classification codes assigned to them. This tool can be used to study the history of technology and, as a technology indicator of technological origins, can also be used for defining technology metrics.

Data Collection Methodology

The data collection methodology is shown in Figure 1. First, we select the whole collection of patents published in a specific technology using classification codes. For example, if this technology is graphical user interfaces (GUI), we must use the IPC code G06F3/048, literally "Interaction techniques based on graphical user interfaces" (IPC codes and titles can be consulted at http://www.wipo.int/). In this way we get the specific "technology" collection. From this set we extract all the citations from its search reports building the "citations" collection. Then, we keep the patents filed before the specific technology has emerged, in this case 1975 (Reimer, 2005) and we obtain the "Roots" collection.

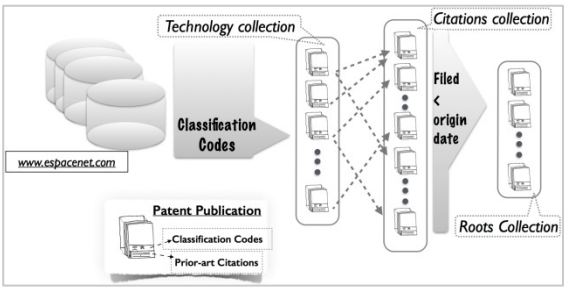


Figure 1. Data collection path

The *TR spectrum*

The set of selected patents—the "Roots" collection—is formed by patent publications disclosing technology methods, concepts, devices or systems intertwined with different aspects of the specific technology under study and filed (and therefore developed) before this technology existed —the origin's date. Analysing in turn the codes assigned to them provide us with indications of the technological foundations of the technology under study. This is why we use the expression: *Technology Roots*. Furthermore, every patent publication in the "Roots" collection is classified with a code representing a technology chosen between all possible existing technologies, this is why we use the term: *spectrum*.
The *TR spectrum* is built by aggregating the classification codes allocated to each document within the "roots" collection, and ordering this dataset in a sequence in accordance with the IPC scheme at a certain level of granularity—*section, class, sub-class, group or sub-group*—(WIPO, 2014). Changing the level of granularity we zoom out or zoom in on the techniques to have different conceptual resolutions and in consequence we can identify more technical details or we can have global views of technical fields. Figure 2 (top graph) shows the *TR spectrum* for computer graphics (CG) at the IPC *class* level. This spectrum was built using the IPC codes G06T11 (2D image generation), G06T13 (Animation), G06T15 (Image rendering), G06T17 (3D image modelling for computer graphics) and G06T19 (Manipulation of 3D models) for the "technology" collection, and the origin date was set at 1960 (Perez-Molina, 2014).
Following our methodology the "technology" collection contained 32,034 documents. Then, all

the patent publications cited in their search reports made a "citations" collection with 83,719 documents. Finally, the "roots" collection is formed by 344 patents.

A tool for studying the history of technology

The direct analysis of the main components of the spectrum provides us with an indication about the technological foundations of a specific technology. Looking, for example, at the computer graphics *TR spectrum* at IPC-*class* level (see Figure 2 top graph), it is straightforward to note that the foundations of CG are mainly in computers, electrical devices and electronics, and photography (the right-hand side of the spectrum), and to a lesser extent in medicine (left) and mechanics (left-center). The main components are G06 (computation), G01 (measuring), G09 (Education, cryptography, displays and seals), H04 (electric communications) and G03 (photography and cinematography).

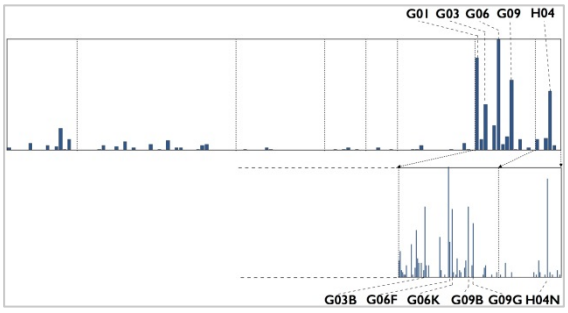


Figure 2. C.G. *TR spectrum* at IPC-*class* level (top) and partial view of the CG *TR spectrum* at IPC-*subclass* level (bottom)

At finer granularity, in other words, aggregating the dataset at the level of *sub-classes*, we have more precision in these technologies already identified. Then, it is clear from the partial view of the TR-spectrum at IPC *sub-class* level (see Figure 2 bottom graph) the importance of digital processing (G06F), television (H04N), photography (G03B), pattern recognition (G06K), educational appliances (G09B) and display control circuits (G09G). If, for instance, we are interested to know which specific technology is behind educational appliances, we zoom in on this spectral component, discovering that the most populated group is simulators (G09B9), and zooming in again we find in particular flight simulators (G09B9/08).

A tool for technology metrics

The *TR spectrum* contains information about the technological influences at the origin of a specific technology. It forms a sort of technology affiliation fingerprint of its origins, thereby it can be used as a technology identifier in technology metrics.

We have used it to get an indication of the relative distances between technologies. The different spectral bin values of the *TR spectrum* are considered as coordinates in a *technology-roots* space, thereby every particular *TR spectrum* is a point in this space. Then, applying *multi-dimensional scaling* (Wickelmaier, 2000) we have reduced the dimensionality for visualizing the relative positions of technologies. Figure 3 shows the results for four technologies—computer graphics (CG), graphical user interface (GUI), computerized tomography (CT) and Airbags—using Euclidean distance.

At present we are experimenting with other distance metrics more suitable for classification spaces.

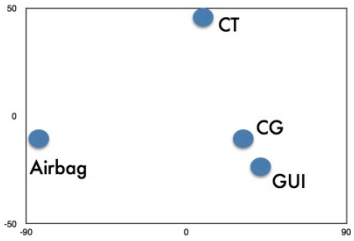


Figure 3. Relative position of CG, GUI, CT and Airbags after applying *multidimensional scaling* to its respective TR spectrums

Conclusions

We have introduced a new visualization tool—the *TR spectrum*—for identifying the technological foundations of a specific technology. We also have briefly disclosed the application of this tool for studying the history of technology and its use as a technology indicator.

References

Meyer, Martin (2000). What is special about patent citations? Differences between patent and scientific citations. *Scientometrics*, *49*(1), 93-123.

Perez-Molina, E. (2014). The Technological Roots of Computer Graphics. *IEEE Annals of the History of Computing*, *36*(3), 30-41.

Reimer, J. (2005). A History of the GUI. Retrieved, December 2014, from http://arstechnica.com/.

WIPO (2014). International Patent Classification - Guide. Retrieved, February 05, 2015, from http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf.

Wickelmaier, F. (2003, May 4). An introduction to MDS. Retrieved, May 16, 2015, from https://homepages.uni-tuebingen.de/florian.wickelmaier/pubs/Wickelmaier2003SQRU.pdf.

Modelling of Scientific Collaboration based on Graphical Analysis

Veslava Osinska[1], Grzegorz Osinski[2] and Wojciech Tomaszewski[2]

[1]*wieo@umk.pl*
Nicolaus Copernicus University, Institute of Information Science and Book Studies
ul. Bojarskiego 1, 87-100 Torun (Poland)

[2]*grzegorz.osinski@wsksim.edu.pl; wojciech.tomaszewski@wsksim.edu.pl*
Institute of Computer Science, College of Social and Media Culture
sw. Jozefa 23/35, 87-100 Torun (Poland)

Introduction

An analysis of the interrelationships between elements within dynamic structure typically involves perturbation methods based on the minimum energy. In result, the researchers use minimum distance-based algorithms and therefore the shortest path between the various components of the system. However, the history of science development shows that collaboration between the researchers in different disciplines becomes effective and fruitful when scientific explorations do not follow the "shortest possible" roads.
In current work authors present a novel approach, how to analyse and evaluate the possible collaborations ways in a small team of researchers (number of nodes is less than 100) participating in the project network KnowEscape COST Action.[1]

Data, metrics and assumption

Analysed dataset consists of 83 records characterized each member of COST network. Input data organized in 83x83 matrix, describe two years collaboration within such activities as: mobility, events organization, publishing (also for former years) and project management. The dataset was gathered using KnowEscape website (knowescape.org), ResearchGate and Mendeley services.
To describe the mutual relationships between members the graph based on Mycielski concept was constructed (Larsen, Propp & Ullman, 1995). The authors identified graphically four attractors of maximum energy. The clique represents each researcher's pair, and arbitrarily large chromatic number means any combination of disciplines. Presented visualisation (Fig. 1) was generated by using the Poincare section (PS) of the 3D space which is defined by all ties between team's members (Tamassia, 2000).
The main problem concerns identification subgroups categories with regard to scientific activity. The matrix was generated using selected nodes and links through Poincare projection (Clifford, Azuaje, & McSharry, 2006).

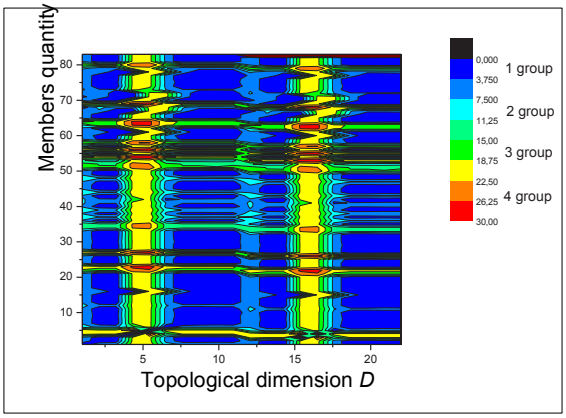


Figure 1. An iterated visualization of discrete distance routes.

Obtained iterated visualization of discrete distance routes is shown on Figure 1. As a final result we observe four clear clusters. All participants were divided on four groups by describing appropriate roles in social network: leaders, connectors, performers and outliers.
This approach was tested using algorithms adopted from medical data analysis for time series (Swierkocka-Miastkowska & Osinski, 2007, Mazur, Osinski, Swierkocka, 2009).
The authors evaluate also the dynamics of total activity by using fractal dimension (FD) of each PS image. FD is the measure of nonclassical geometry shapes and can be used as a pattern's complexity parameter (Osinska 2012).
Fractal dimension was obtained by Higuchi algorithm, so the resulting maps help to discover possible opportunities for further development of cooperation between the scientists.

Visual results

All members' activities represented by matrixes are summarized and full collaboration is weighted by appropriate real numbers. Popular application *Gephi* allows finding collaboration groups and revealing the scientists with basic roles: leader,

[1] This research is sponsored by National Science Center (NCN) under grant 2013/11/B/HS2/03048/ Information Visualization methods in digital knowledge structure and dynamics study.

subgroup leader, connector, outsider and so on. By using force directed layout (*force atlas 2*) the authors have obtained clarify configuration presented on Figure 2. As expected, the central point is occupied by the real team's leader. The closer node to central one represents the scientist who is more active in collaboration with the team's leader.

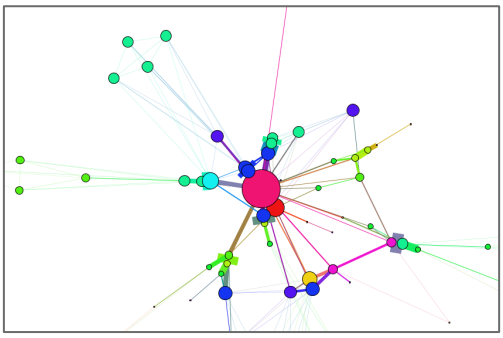
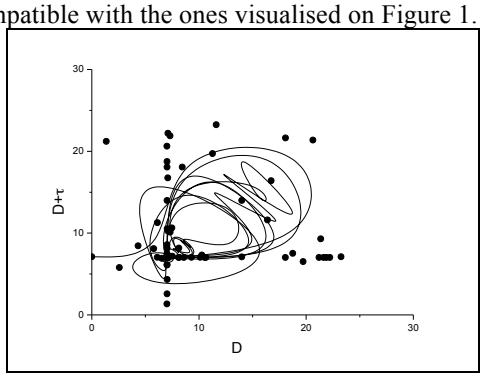


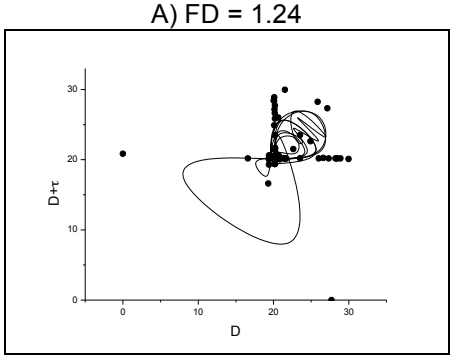
Figure 2. The graph of full activity of team's members.

Network visualisation exposes also some subgroups where intrinsic collaboration (mainly in publishing) is significant. The scientists within these groups share a common feature: geographic localisation. They work in the same country.

Simple quantitative proportional correlations between identified groups on a graph are compatible with the ones visualised on Figure 1.



A) FD = 1.24



B) FD =1.39

Figure 3. Two variations of collaboration between scientists with different social roles: A) Leader-performer; B) performer-performer.

Next step, calculation of fractal dimension, was accomplished for combinations of representatives of different groups, for example: leader-performer, subleader-leader, connector-performer and so on.

Two variations of collaboration with appropriate FD are shown on Figure 3. Fractal dimension is always lower for every pairs composed from the leader or subleader compared to the performers and connectors.

Conclusions

The authors propose new parameters for the prediction of a stable way of scientific collaboration. First is the shape of Poincare section (Return Map Poincare). For inhomogeneous academic groups where there is no self-consistency (like in this work), the level of nonlinearity can also reflect collaboration potential. It is proportional to the quantity of curves on Figure 3. The second indicator – FD shows the possibility to cooperate as well as its dynamics.

Higher fractal dimension in the case of performers can be explained by larger dynamics of predictive collaboration. This indicates the pattern is more complex. It means the pair covers significant collaboration potential.

Visualisation can help discover possible opportunities for further development of scientific cooperation. Therefore, we can observe common career landscapes of the various members and groups.

References

Clifford, G.D., Azuaje, F. & McSharry, P. (2006). *Advanced Methods and Tools for ECG Data Analysis.* Artech House Publishers.

Larsen, M., Propp, J., & Ullman, D. (1995). The Fractional Chromatic Number of Mycielski's Graphs. *Journal of Graph Theory*, *19*, 411-416.

Mazur, R., Osinski, G., Swierkocka, M. & Mikolaiczik, G. (2009). Evaluation of the dynamics of energetic changes in the brain stem respiratory centre in the course of increasing disorders of consciousness. *Activitas Nervosa Superior. 51*(5112), 69-72.

Osinska, V. (2012). Fractal Analysis of Knowledge Organization in Digital Library. In A. Katsirikou & Ch. Skiadas (Eds.) *New Trends in Qualitative and Quantitative Methods in Libraries* (pp. 17-23). World Scientific Publishing Company.

Swierkocka-Miastkowska, M. & Osinski, G. (2007). Nonlinear analysis of dynamic changes in brain spirography. Results in patients with ischemic stroke. *Clinical Neurophysiology*, *118*(12), 2822.

Tamassia, R. (2000). Graph Drawing. Ch. 21 In J.-R. Sack & J. Urrutia (Eds.) *Handbook of Computational Geometry* (pp. 937-971). Amsterdam, Netherlands: North-Holland.

Monitoring of Technological Development - Detection of Events in Technology Landscapes through Scientometric Network Analysis

Geraldine Joanny[1], Adam Agocs[2], Sotiri Fragkiskos[2], Nikolaos Kasfikis[2], Jean-Marie Le Goff [2] and Olivier Eulaerts[1]

[1] *geraldine.joanny@ec.europa.eu*
Joint Research Centre, European Commission, Brussels (Belgium)

[2] CERN, Geneva, (Switzerland)

Introduction

Monitoring technological development is an important challenge for research organisations and regulators. For decision-makers, the detection of early signals of technology maturation is key to designing proper standards and regulations. Anticipating the arrival of new technologies also allows policy-makers to develop and implement fit-for-purpose research or industrial policies. Scientometric analysis (in this case using both publications and patents) is a powerful tool to monitor technological fields and can be used to detect events in the lifecycle of a technology (Rotolo et al., 2014).

Objectives

- to analyse different cases (historical) of technological change by monitoring the evolution of patterns of collaboration between research organisations, the apparition of new keywords and/or subject categories in articles as well as changes in quantitative data such as patent or publication counts;
- to investigate whether network analysis can be used for the detection of events related to technological change;
- to identify potential indicators of technological maturation useful in the context of early warning to regulators.

Methods

Results relating to 4 technologies are presented here. Publications for each technology were retrieved from the Web of Science Core Collection database and patents from Thomson Innovation. To select the technologies, a semantic search was used in the abstract, title and author keywords of the publications.

Different network landscapes were then created using the retrieved patents and publications: sociograms showing how organisations collaborate together (through co-publishing and co-patenting); keywordgrams based on co-occurrence of author keywords in articles; and subject-category-grams based on subject categories given by Thomson Reuters. These three types of network landscapes were created and analysed for each technology.

Results

Shale Gas and horizontal drilling

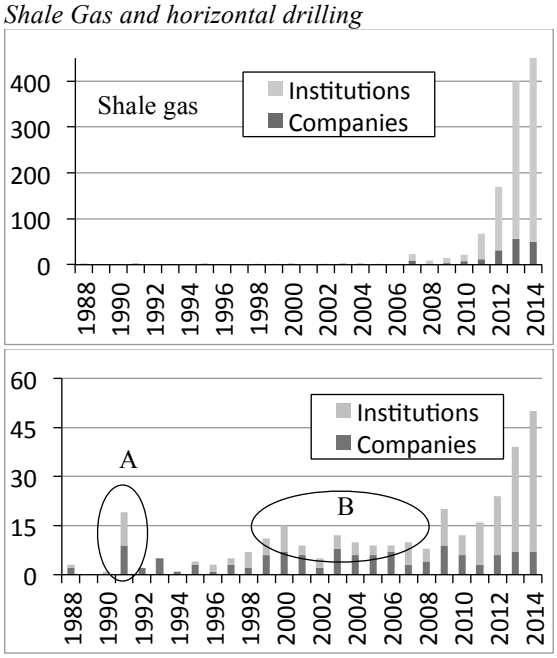


Figure 1. Number of patents and publications for horizontal drilling and shale gas from 1988.

Figure 1 shows that the number of patents and publications mentioning "shale gas" in the abstract, title or keywords started to increase noticeably in 2007 and boomed from 2011 onwards. By contrast, articles mentioning horizontal drilling, one of the key enabling technologies for "shale gas" appeared earlier (A) and rose from the year 2000 onwards (B). In addition, comparison with press content analysis shows that the rise in articles mentioning "shale gas" correlates with an increase of occurrences of press articles about shale gas (data not shown), which leads to think that this rise does not correspond to a technological trend. This shows that for the prediction of technological change the subjacent technologies - not the broad concepts - are more meaningful for the early detection of technological change.

The 2nd graph of Figure 1 shows the need to build composite indicators to avoid false positive signals. The peak of publication activity in 1991 is indeed not correlated to increased activity in other

indicators such as volume of patents or variation of number of players, for example (data not shown).

3D-printing - Detection of new uses of a technology
The number of patents and publications on fused-deposition modeling (a key enabling technology of 3D-printing) is growing steadily from 1995 to nowadays (data not shown). The subject categories of the journals in which the selected publications were published are manifold and evolve in time. As shown in Figure 2, from 1998 to 2014 a few clusters of new subject categories appear. In 1998 the articles relating to fused deposition modeling were belonging to engineering, material science and automation, which are categories describing the core of this technology. Categories describing applications of 3D-printing appear as of 2001, i. e., earlier than the entry of the first 3D printer on the market (2009).

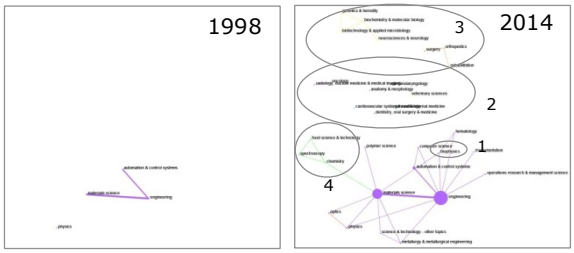


Figure 2. Subject categories for publications on fused-deposition modeling in 1998 and 2014. The circles show appearance of new non-core subject categories. 1. Biophysics (2001), 2. Radiology (2004), dentistry (2005), oncology (2006) 3. Genetics, Biochemistry (2007), Neurosciences (2008) 4. Food science and chemistry (2011).

CRT - Detecting substituting technology
The study of the author keywords for publications related to cathode ray tube (CRT) allowed to observe the emergence of the replacing technology, Liquid Crystal Display, in the CRT space. Figure 3 shows various synonyms of LCD in the keywordgram for CRT. The LCD nodes are quite big, showing their relative importance. The keyword LCD or its synonyms appear in 35 out of 649 publications or 5% of the publications.

Silicon wafer for microelectronic and for solar cell
Two application lifecycles can be observed for silicon wafers by analysing the number of related publications and patents (data not shown). These two lifecycles culminate respectively around the years 2000 and 2010. Analysing the keywordgram for the selected publications we see the keyword "silicon solar cells" appearing in 1999, and being increasingly used until 2011. Figure 4 shows its co-occurrence with other keywords in 2014. The emergence of this keyword reflects the apparition of a new use of silicon wafers for solar applications.

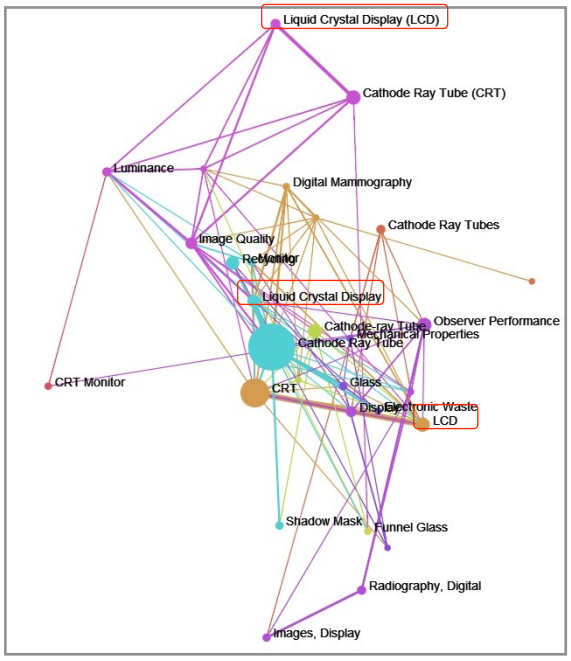


Figure 3. Author keywords view for Cathode Ray Tubes in 2014.

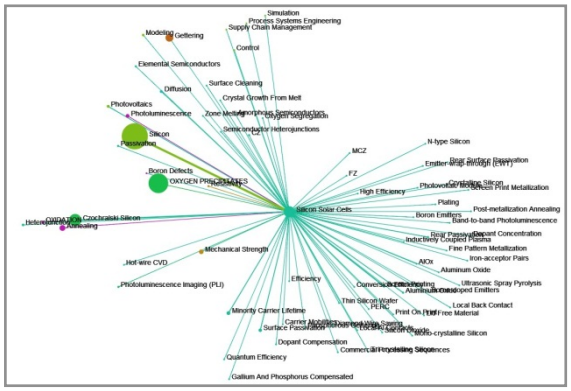


Figure 4. Centric view of keyword "Silicon Solar Cells" and its co-occurrence with other author keywords in the publications space relating to Silicon wafers.

Conclusions

Our study suggests that network analysis can be used for the detection of events relating to technological change.
We have identified several types of indicators that could be combined in order to design an early warning system to alert decision-makers of changes in technology landscapes.

References

Rotolo, D., Rafols, I., Hopkins, M. & Leydesdorff, L. (2014). Scientometric mappings as strategic intelligence for tentative governance of emerging science and technologies. *SPRU Working Paper Series,* 10, 1-40.

Analysis of R&D Trend for the Treatment of Autoimmune Diseases by Scientometric Method

Eunsoo Sohn[1], Oh-Jin Kwon[1], Eun-Hwa Sohn[2] and Kyung-Ran Noh[1]

[1] *essohn@kisti.re.kr*, [1] *dbajin@kisti.re.kr*, [1] *infor@kisti.re.kr*
Korea Institute of Science and Technology Information, Seoul 130-741 (Korea)

[2] *ehson@kangwon.ac.kr*
Kangwon National University, Gangwon-do 245-905 (Korea)

Introduction

Autoimmune diseases (AD), referred to as abnormal immune responses of body against self-antigen, are caused by the loss of immunologic self-tolerance resulting in damage to the cells, tissues and organs. The National Institute of Health (NIH) lists more than 80 autoimmune diseases that affect varied organs of the body including rheumatoid arthritis, multiple sclerosis, systemic lupus erythematosus and so on.

Significant advances of AD have been made in the understanding of clinical and pathological mechanisms involved but, to date, a few elements have been identified as being responsible for the autoimmune process. With a better understanding of the causes and treatments of AD, many potential novel therapies have recently been developed and evaluated, focusing on cellular or molecular targets.

Although there have been several research activities carried out with scientometric tools to evaluate scientific output for individual autoimmune diseases such as rheumatoid arthritis, Crohn's and Behchet's disease (Shahram et al., 2013), there was no scientometric studies on the entire autoimmune disease to date. Density-equalizing algorithms, scientometric methods and large scale data analysis were applied to evaluate quality and quantity of scientific researches in rheumatoid arthritis (Schöffel et al., 2010). Various scientometric analysis including literature-related discovery (LRD), text-mining was more broadly performed to produce knowledge discovery such as gene expression and proteomic studies. Data mining and bioinformatics approaches for autoimmune biomarker discovery studies were also attempted (Kostoff, 2014).

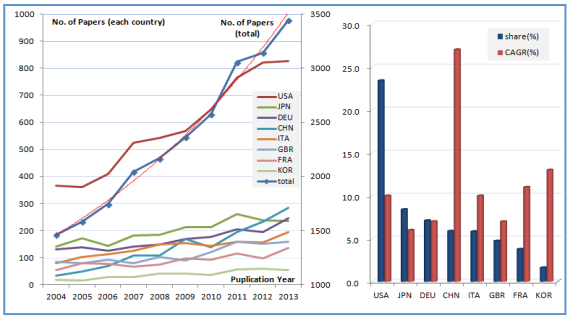
The purpose of this study is to analyze the status and trends of treatments for AD using scientometric methods, and intend to give researchers and policy-makers valuable information in the field of AD.

Data and Methods

Publications associated with the treatment of AD were retrieved from Elsevier's SCOPUS database. The query to collect data for scientometric analysis was as follows: "TS=(autoimmun*) AND TS=(therap* OR treatment*)" Total 23,587 articles published during recent 10 years (2004-2013) were collected and analyzed. Microsoft Excel, KITAS, NetMiner and VOSviewer software were combined to analyze bibliometric data. KITAS software from KISTI (Korea Institute of Science and Technology Information) was used for data extracting and cleaning. NetMiner and VOSviewer software were also used for clustering and mapping.

Results and Discussion

Figure 1 shows R&D trends over time in major countries, and the share and CAGR (compound annual growth rate) of each country based on scientific papers regarding treatments of AD. Over the last 10 years, there has been a significant growth in performance of papers with CAGR 10% in this field. Although the US quantitatively represents the largest share (23.4%), China shows the most rapid CAGR 26.6% followed by Korea (13.2%). Especially in the field of AD, Japan and Germany show a strong tendency compared with other general aspects of pharmaceuticals.



(a) (b)

Figure 1. The changes of number of papers (a) and the share and CAGR by major countries (b).

2-mode network in Figure 2 shows the co-occurrence between main countries and keywords extracted from papers, which can help identifying; which country related to; which kind of autoimmune diseases or therapeutics or treatment technologies. Circle nodes represent countries and the size of each node indicates the number of publications. The degree of relationships is

indicated by the thickness of the link and the distance between two nodes.

Keywords are divided into 2 groups, different types of AD at the bottom of Figure 2 and its technical terms at the top. In terms of the disease, high prevalence of AD including rheumatoid arthritis, multiple sclerosis, type I diabetes have shown a high correlation with US. Japan is estimated to be active in the field of autoimmune pancreatitis, autoimmune hepatitis, and Germany seems active in multiple sclerosis and type I diabetes. In particular, autoimmune thyroiditis shows a high correlation with Japan, Germany and Italy rather than US. As shown in the top of Figure 2, US is very active across all areas of the field. Advanced immunotherapies with cell-based technologies using dendritic cell, regulatory T cell (T-reg) are particularly revealed to be active in Japan and Germany as in the US.

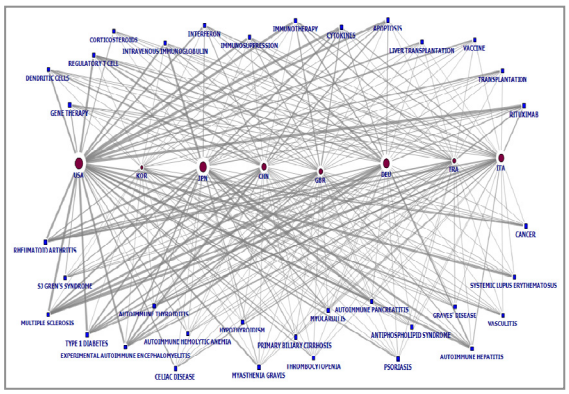


Figure 2. 2-mode network of the major countries and keywords related to autoimmune diseases.

Figure 3 provides the knowledge mapping for AD treatment drawn by co-word analysis, which shows the hot topic field or an increasing R&D productivity trend for AD treatment. To find out changes in R&D trends for treatment of AD, the dataset was divided in two time periods: 2004 to 2006 and 2011 to 2013. Several changes are found in the map of the past 3 years (2004-2006) compared with the last 3 years (2011-2013).

Figure 3 shows an experimental study using experimental autoimmune encephalomyelitis (EAE) animal model of multiple sclerosis has been disappeared in the last map (2011-2013). As time passed, clinical studies on many diseases considered to be autoimmune have been conducted with various organs and systems including endocrine, hepatobiliary, vascular systems. In addition, cell-based immune therapies with regulatory T cell (T-reg) or Th17 cells gradually have emerged in the last map (2011-2013). Immunomodulatory effects of mesenchymal stem cell (MSC) are also shown in the second figure of Figure 3. This might imply that a targeted immune therapy had been developed and successfully utilized in treating AD patients.

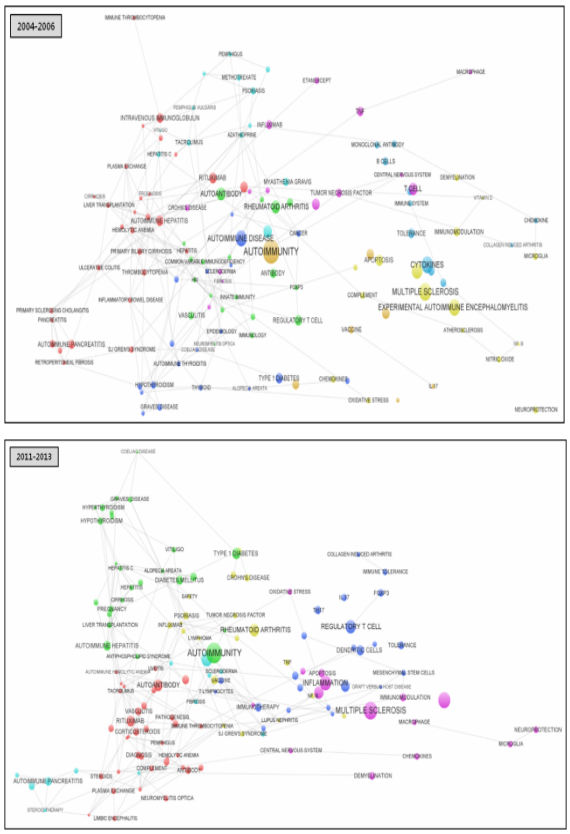


Figure 3. Co-word knowledge mapping product for the treatment of autoimmune disease.

In this study, we investigated present R&D status and trend for the treatment of AD using scientometric analysis methods. The trend in advanced R&D for the treatment of AD was identified through knowledge mapping techniques such as co-word analysis of articles and visualization technology. The results show that each country has progressive development of AD therapeutics with any other aspect. Additionally, the approach to identify the molecular and cellular mechanisms of AD underlying the immune tolerance has been increased.

References

Kostoff, R.N. (2014). Literature-related discovery: common factors for Parkinson's Disease and Crohn's Disease. *Scientometrics*, *100*, 623-657.

Schöffel, N., Mache, S., Quarcoo, D., Scutaru, C., Vitzthum, K., Groneberg, D.A., & Spallek, M. (2010). Rheumatoid arthritis: Scientific development from a critical point of view. *Rheumatoloty International*, *30*(4), 505-513.

Shahram, F., Jamshidi, A.R., Hirbod-Mobarakeh, A., Habibi, G, Mardani, A., & Ghaemi, M. (2013). Scientometric analysis and mapping of scientific articles on Behcet's disease. *International Journal of Rheumatic Diseases, 16*(2), 185-192.

Analysis of Convergence Trends in Secondary Batteries

Young-Duk Koo[1] and Dae-Hyun Jeong[2]

[1] *ydkoo@kisti.re.kr*
Korea Institute of Science & Technology Information, Gyeonggi-Inchon Branch, 145 Gwanggyo-ro, Yeongtong-gu, Suwon0si, Gyeonggi-do (Korea)

[2] *gregori79@kisti.re.krt*
Korea Institute of Science & Technology Information, Dept. of Creativity Implementation, 66, Hoegi-ro, Dongdaemun-gu, Seoul (Korea)

Introduction

Convergence refers to the creation of new technologies (or industries, markets) through the combination of two or more technologies (or industries, markets), which is promoted by technical changes, innovations, and technology diffusion, and plays a key role in changing gradual innovations to destructive innovations. Furthermore, convergence is a key factor in accelerating changes in the growth curve of technologies and the life cycle of products (Pennings & Puranam, 2001). This study was conducted to analyze convergence trends in secondary batteries and find their implications. For this purpose, useful papers and patent data for analysis were selected, collected, and processed to calculate the convergence index. This attempt is expected to provide the foundation for predicting convergence by identifying major causes that accelerate convergence. To effectively measure convergence status in this study, the diversity index suggested by Yegros Yegros et al. (2003) was used. The diversity index, which is used to measure interdisciplinary studies, considers three aspects: variety, balance, and disparity. An interdisciplinary study means the integration of different disciplines, thereby creating new academic disciplines. In this study, the convergence index was derived by the integration of different technologies into one technology.

Method of Analysis

For this purpose, the diversity index suggested by Yegros Yegros et al. (2013) was used for analysis, and IPC International Patent Classification) was used for the analysis of patents. IPC codes are assigned to individual patents and multiple codes can be specified depending on the case. In this study, IPC codes were used to analyze the convergence phenomena in secondary batteries (Stirling, 1998, Purvis et al., 2000, Stirling, 2007). The equation for each variable is given below.

$$\text{Variety} = n$$

$$\text{Balance} = -\frac{1}{\ln(n)}\sum_i p_i \ln p_i \quad (1)$$

$$\text{Disparity} = \frac{1}{n(n-1)}\sum_{ij} d_{ij} \quad (2)$$

(d_{ij} = 1-cosine coefficient)

In this equations, n means that number of IPC codes and p_i means that ratio of i IPC code.
In this study, U.S. patents about secondary batteries that had been opened or registered between January 1, 1998 and December 31, 2011 were analyzed with the IPC code for secondary batteries H010-010 using the USPTO database. In this study, we use patent data until 2011 because patent data is valid until 2011.

Table 1. Search formula for secondary batteries

Data	Search formula	Number of patents
USPTO	IPC=H01M-010*, PY=19880101~20111231	8,181

Result and Discussion

The measurement of variety through the number of IPC subclasses about patents in secondary batteries by year showed that the variety value was increasing sharply over time. In particular, the variety value greatly increased after 2009 when the number of applicants in medium- and large-sized secondary batteries increased rapidly, indicating that the variety value of secondary batteries increased with the active research related to medium- and large-sized secondary batteries. The measurement of balance by year showed that the balance value decreased between 1988 and 2000, and steadily increased again after 2003. This suggests that with the beginning of the development of the medium- to large-sized secondary batteries, research and development of various technologies have been carried out to develop the required technologies. The measurement of disparity values by year showed that the disparity value has been decreasing over time. This suggests the decreasing distance between technologies and the progress of convergence.

Figure 1. (left) Trend of variety by year; (middle) Trend of balance by year; (right) Trend of disparity by year.

In particular, the distance between technologies has become very low after 2001. As analyzed above, with the emergence of medium- to large-sized secondary batteries, convergence with other technology fields such as eco-friendly cars and solar cells has been going on.

Figures 2 and 3 show the network structure of IP codes for secondary batteries by period (1988-2000, 2001-2011). The node size indicates the number of IPCs and the length of link indicates the distance between different IPCs. The network structure of IPC codes shows that IPCs have gathered together since 2001, indicating that the relationships among different technologies have been strengthened and the distances shortened since 2001. Furthermore, IPCs related to new application fields for medium- and large-sized secondary batteries such as solar cells and wind power energy have appeared, and the distance between them and the representative IPC for secondary batteries has become closer since 2001. In other words, with the research and development of medium- and large-sized secondary batteries since 2001, the convergence in secondary batteries has become conspicuous.

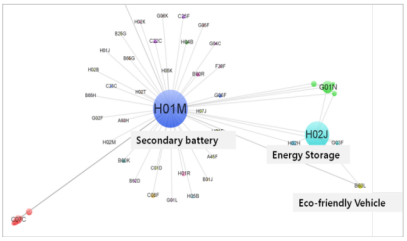


Figure 2. IPC network structure (1988-2000)

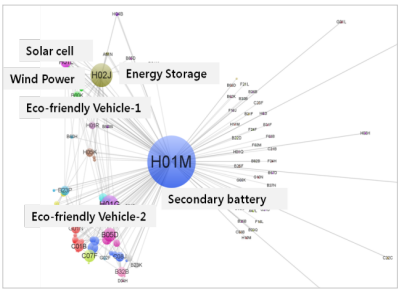


Figure 3. IPC network structure (2001-2011)

Conclusion

In this study, we analysis of convergence trend using patent data of secondary battery. As a result, it can be summarized as follows: First, as passing by year, convergence of secondary battery has increased, especially, in terms of variety and balance. This means that as increasing convergence, various field has merged and increased similarity between fields. Second, as the comparing result of IPC mapping between 1998-2000 and 2001-2011, convergence in secondary batteries is greatly increasing around the medium- and large-sized secondary batteries with the progress of convergence with eco-friendly vehicles, wind power energy, and solar energy and the decreasing distance between technologies. Predicting the convergence trends in secondary batteries has great implications to countries and companies in that they allow us to predict future industries and search for new markets and strategic partners. Furthermore, considering that existing studies used patents in a limited way due to limitations of patent analysis and limited use of time-series patent data so far, the analysis in this study was useful.

Acknowledgments

This work was supported by Construction of SMB Support System based on Industry-University-Institute Knowledge Ecosystem (K-15-L04-C01-S01).

References

Pennings, J.M. & Puranam, P. (2001). Market convergence & firm strategy: new directions for theory and research, *ECIS Conference*. *The Future of Innovation Studies*.

Purvis, A. & Hector, A. (2000). Getting the measure of biodiversity. *Nature*, *405*, 212-219.

Stirling, A. (1998). On the economics and analysis of diversity. *SPRU Electronic Working Paper.*

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, *4*(15), 707-719.

Yegros Yegros, A., D'Este Cukierman, P., & Rafols, I. (2013). Does interdisciplinary research lead to higher citation impact? *35th DRUID Celebration Conference.*

Can Scholarly Literature and Patents be Represented in a Hierarchy of Topics Structured to Contain 20 Topics per Level? Balancing Technical Feasibility with Human Usability

Michael Edwards[1], Mahadev Dovre Wudali[2], James Callahan[3], Paul Worner[4], Jeffrey Maudal[5], Patricia Brennan[6], Julia Laurin[7] and Joshua Schnell[8]

[1]*michael.edwards@thomsonreuters.com*
[2]*mahadev.wudali@thomsonreuters.com*
[3]*jim.callahan@thomsonreuters.com*
[4]*paul.worner@thomsonreuters.com*
[5]*jeff.maudal@thomsonreuters.com*
Data Center Operations, Thomson Reuters, 610 Opperman Drive, Eagan, Minnesota 55123

[6]*patricia.brennan@thomsonreuters.com*
[7]*julia.laurin@thomsonreuters.com*
[8]*joshua.schnell@thomsonreuters.com*
Intelectual Property & Science, Thomson Reuters,
1500 Spring Garden St, Philadelphia, Pennsylvania 19130

Introduction

The Intellectual Property & Science division of Thomson Reuters curates millions of records a year covering scholarly literature (Web of Science®), patents and intellectual property (Derwent World Patent Index®) and life sciences discovery (Cortellis®). These millions of records could be connected through billions of potential relationships, such as that represented by a citing relationship between literature and patents, or by different documents that pertain to similar topics. By building these relationships using machine learning techniques we hope to unite information from different data sources to enable extraction of knowledge such that the whole is greater than the sum of the parts, with minimal human effort required.

However, connecting these documents in a meaningful way is challenging from both a technological perspective as well as a usability perspective. As shown in Figure 1, studying citation patterns among approximately 250,000 articles from the Web of Science, or 1/200 of the full data set, generates a citation graph that, while rich with information, is extremely difficult to use to understand knowledge flows.

This challenge is the focus of our presentation. For this research project, we have created a graph of the topics represented in a subset of the scholarly literature and granted patents, in order to explore ways to constrain the visualization of this topic graph to emphasize usability. While many additional research areas remain, our initial findings suggest that such constraint enables users to easily explore the knowledge graph in way that maximizes understanding while minimizing user effort.

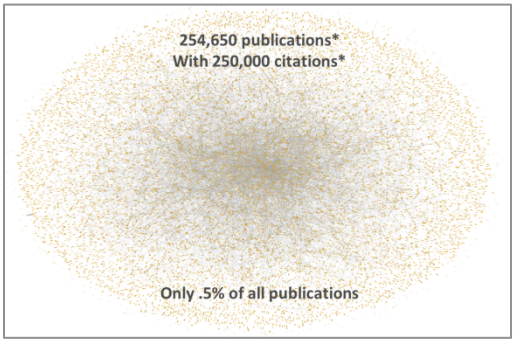


Figure 1. Ball and stick diagram of the citing relationships among a select set of publications from Web of Science®.

Generation of the Topic Graph

We chose to use topic modelling based on the latent dirichlet allocation (LDA) algorithm (Blei, Ng & Jordan, 2003) to generate connections between documents that reflect the shared knowledge among scholarly articles and granted patents. From Web of Science, we selected 27 million publications published since 1990 that had abstracts in English. Our past experience with LDA topic modelling led us to take a hierarchical approach to clustering the documents based on topics. We created a tree of over 1 million topics for the corpus, parceling out the topics into manageable chunks (20 at a glance) which were a better fit for human perception. We also created our own algorithm for applying these topics to patents, demonstrating a flexible, unsupervised technique for combining two distinct content sets. We found that the hierarchy we produced generally exhibited 4 to 5 levels of depth to the terminal nodes or documents.

Understanding the Knowledge Graph

We created the Epiphany tool to more effectively navigate the corpus of scholarly articles, using both browse and search interactions. As shown in Figure 2, the tool supports drill-down (e.g. 2.6 million articles assigned to an algorithm-focused topic; left side green), as well as search, (e.g. 8 topics strongly related to "genetic programming"; right side orange). This allows users to interact with topics and the relevant documents to understand the underlying data.

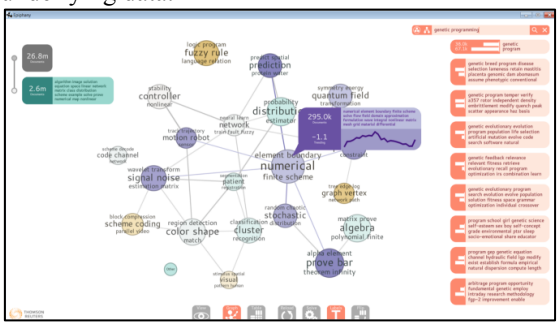


Figure 2. Screenshot of Epiphany tool showing topic clusters matching "genetic programming" search criteria.

Drilling down into the topic details is show in Figure 3. At the top in purple are statistics on the topic itself including the number of documents closely associated with the topic, the most frequent terms and the Trending metric score for the topic.

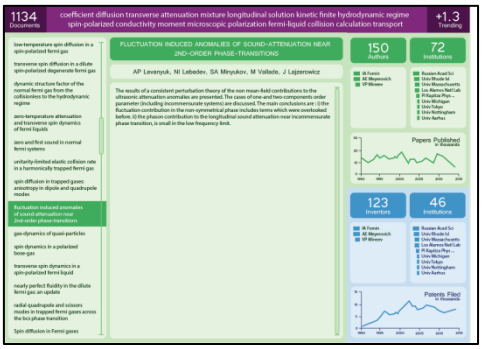


Figure 3. Screenshot of Epiphany tool Topic Details screen.

The right side of the panel contains two statistics sections, one in green for scientific papers and one in blue for patents. The header for each of the sections includes counts of the unique number of authors (or inventors) and unique number of institutions (or assignees) responsible for creation of the documents associated with the topic. Below these counts are a breakdown of the most commonly mentioned authors (inventors) and institutions (assignees). Finally, the bottom part of the statistics section is a graph of the proportion of documents assigned to this topic out of all documents published for each year.

Project Outcomes

The purpose of this research project is to test the application of scalable machine learning techniques to generate a knowledge graph that is accessible to the analyst. Now that we have developed the Epiphany tool, we have begun using it to gather feedback on this approach from a cross section of potential users. We expect to present that feedback at the ISSI2015 conference specifically to answer the question of whether a topic graph of millions of records of scholarly literature and granted patents can indeed be represented in hierarchical structure with a maximum of 20 topics at each level.

Acknowledgments

We acknowledge the support of the Intellectual Property & Science staff and the Data Center Operations staff for improvements made to this research project.

References

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, *3*, 993-1022.

A Sciento-Text Framework for Fine-grained Characterization of the Leading World Institutions in Computer Science Research

Ashraf Uddin[1], Sumit Kumar Banshal[2], Khushboo Singhal[3] and Vivek Kumar Singh[4]

[1]*mdaakib18@gmail.com,* [2]*sumitbanshal06@gmail.com,* [3]*khushbusinghal18@gmail.com,* [4]*vivek@cs.sau.ac.in*
Department of Computer Science, South Asian University, New Delhi (India)

Introduction

This paper describes our experimental framework for a text analysis based fine-grained characterization of leading world institutions in Computer Science (CS) research. Though the present paper uses CS research output data from Web of Science, it can be extended and applied to any discipline and data source. The existing well-known ranking systems, such as ARWU[1],Times Higher Education World University rankings[2], QS World University Rankings[3], SIR[4], Leiden Ranking[5] and Webometrics[6], only present an overall (or for a whole discipline) rank of institutions. These rankings may not be helpful if one is interested in knowing centers of excellence in research in a particular area (say Artificial Intelligence or Software Engineering in CS). Such fine-grained characterization could be very useful for different purposes. Prospective students looking to work in a particular specialized area may look at the fine-grained characterization and select institutions accordingly. Academicians or industry professionals looking for collaboration in a particular area can use the information for selecting potential institutions for collaboration. Similarly, funding agencies and policy making bodies in a country may identify institutions strong in different specialized areas of research. The other advantage of this kind of sciento-text characterization is that it is completely automated, verifiable and does not use any perceptual scores for ranking (such as reputation survey and perceptual scores of QS). Our system thus proposes a framework that uses scientometric data to produce a fine-grained research strength characterization of institutions and to rank them in order of their research excellence in a particular area.

Data Collection

We have demonstrated the working and suitability of our approach for CS domain. We obtained research output data for CS domain for the period 1999 to 2013 indexed in Web of Science (WoS). The data has been collected through an institution-wise search and we collected data for top 100 most productive institutions. A total of 261,154 records were obtained. This data constitutes about 34% of the total worldwide CS domain research output (784,920 records in total) for the period 1999-2013.

Sciento-Text Based Analytical Framework

Since our main objective is to produce a fine-grained characterization and consequential rankings, we had to first assign every research output to one or more particular research specialization. We identified a total of 11 major thematic areas (specializations) in CS domain research output. The 11-classes are based on perusal of data, some recent work (Gupta et al., 2011; Uddin et al., 2015) and recent research trends in the discipline. We processed each record in the data, extracted its 'title', 'author keywords' and 'abstract' fields and obtained the text contents of these fields. For classifying a record (research paper) to belong to one or more of the 11 thematic areas (specializations), a simple Naïve Bayes (NB) text classifier is used. The names of the 11 classes are embedded in table 1. For obtaining training data for the NB classifier, we used a keyword-match strategy for a part of the data. First of all, we created a term-profile for each thematic area (through a manual annotation by three independent annotators). Then, each record is checked for occurrence of any term from the term-profile of the 11 thematic classes, in its 'author keyword', 'title' and 'abstract' fields, in a sequential manner. Those records which get an exact match of keywords with one or more of the 11 thematic classes are assigned that class label. The assigned records then serve as training set for NB classifier, which is then used to classify the remaining unclassified records. In this manner, we classify each record to belong to one or more of the 11 thematic classes. After assigning thematic class to each record, we partitioned the data into 11 groups. Now, we have research output data for each of the major thematic areas (specializations) from the 100 most productive institutions of the world. This information is now used to first produce a plot of the research output landscape of the 100 most productive institutions and then to identify top ranking institutions in all the thematic areas. For ranking we use a simple average of scientometric indicator values for these

[1] http://www.shanghairanking.com/
[2] http://www.timeshighereducation.co.uk/world-university-rankings/
[3] http://www.topuniversities.com/university-rankings
[4] http://www.scimagoir.com/
[5] http://www.leidenranking.com/
[6] http://www.webometrics.info/

Table 1. Thematic Area Wise Top Ranking Institutions.

AI	CT	CHA	CN	CSA	CG	DBMS	IM	OS	SIP	SE
NTU	NTU	INRIA	INRIA	UCB	INRIA	NTU	TU	INRIA	NTU	INRIA
UCB	MIT	IBM	NTU	INRIA	SJTU	HU	INRIA	TU	UL	UCB
TU	INRIA	TU	UCB	KL	NTU	INRIA	MS	KL	UCB	HU
MS	UL	NTU	TU	NTU	UT	MIT	NUS	HKPU	NUS	UL
UGR	UM	GIT	CUHK	UL	UL	UL	HU	IBM	UIUC	MIT
CUHK	UTA	UCB	HIT	CMU	UW	NUS	NTU	UM	MS	NTU
INRIA	PSU	INTEL	UNC	TU	KL	MS	SU	UW	INRIA	UNC
HKPU	CMU	MS	UL	GIT	TU	MPG	CUHK	UCSD	TAU	UMCP
HU	UCL	PUC	SU	MIT	CUHK	CU	UL	NTU	TU	TU
UL	SU	CMU	GIT	MPG	IBM	IBM	MIT	UCB	KL	IBM

AI : Artificial Intelligence, CT: Computation Theory, CHA: Computer Hardware & Architecture, CN: Computer Networks ,CSA: Computer Software & Applications, CG: Cryptography, DBMS: Database Management System, IM: Internet & Multimedia, OS: Operating System, SIP: Signal & Image Processing, SE: Software Engineering

institutions, namely TP (Total Papers), TC (Total Citations), ACPP (Average Citations Per Paper), and HiCP (Highly Cited Papers). The absolute scores are first normalized to 0-100 range and then a simple arithmetic average is computed. One such similar ranking work (without thematic areas) is presented in a past literature (Ma et al., 2008).

Results and Conclusion

Our framework produces a detailed characterization of research output along the major research themes by the 100 most productive institutions of the world. The Figure 1 presents a plot of TP and TC values along the 11 research themes for the whole set of 100 institutions. Top ranking institutions identified in all 11 thematic research areas for the given period are listed in table 1. It can be seen that many of the institutions are almost available in each list but with different rank positions. Thus the presented results verify the importance of ranking institutions in different thematic areas rather than doing it for a broader research field. The paper thus presents an interesting framework for fine-grained characterization of leading world institutions and to identify the top ranking institutions in different thematic areas of CS domain. The work is extendable to other disciplines and data sources. The work may benefit more if we would have incorporated the number of researchers and graduate students for better insightful result but unfortunately obtaining those data for each institution is cumbersome and time consuming. See http://www.viveksingh.in/publications/issi2015/appendix.pdf for the full names of institutions.

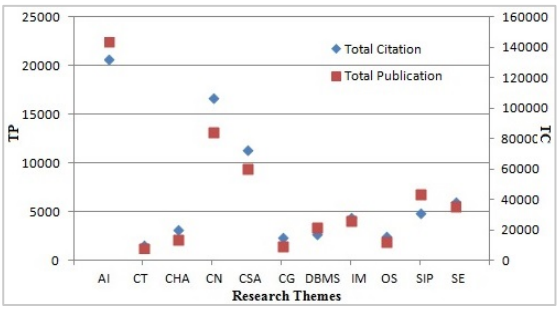


Figure 1. Thematic Area Wise Research Output and Citations.

Acknowledgments

This work is supported by research grants from Department of Science and Technology, Government of India (Grant: INT/MEXICO/P-13/2012) and University Grants Commission of India (Grant: F. No. 41-624/ 2012(SR)).

References

Gupta, B.M., Kshitij, A., & Verma, C. (2011). Mapping of Indian computer science research output, 1999–2008. *Scientometrics*, *86*(2), 261–283.

Ma, R.., Ni, C., & Qiu, J. (2008). Scientific research competitiveness of world universities in computer science. *Scientometrics*, *76*(2), 245–260.

Singh, V.K., Uddin, A., & Pinto, D. (2015). Computer Science Research: The Top 100 Institutions in India and in the World. Submitted in *Scientometrics*.

Influence of *Human Behaviour and the Principle of Least Effort* on library and information science research

Yu-Wei Chang

yuweichang2013@ntu.edu.tw
National Taiwan University, Department of Library and Information Science, No. 1, Roosevelt Rd., Taipei (Taiwan)

Introduction

The principle of least effort (PLE), a concept advanced by the American linguist George Kingsley Zipf, indicates that people complete tasks by choosing the way of least effort among various options (Zipf, 1949). To prove that the PLE is an indication of human nature, Zipf analyzed numerous empirical data collected from various human activities and used mathematical formulae to explain his findings. Zipf explained the PLE in detail in his classic 1949 entitled *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology* (HBPLE).

The PLE represents a common human behavior; it may thus be expected that the HBPLE has become visible in various fields and applied to various human activities. HBPLE was also compared with similar theories and was reconceptualized in the field of library and information science (LIS) (Austin, 2001; Gratch, 1990). The LIS publications on PLE have indicated that the concept of the PLE is connected to various topics (Bronstein, 2008; Chrzastowski, 1995, 1999; Kim, 1982; Wang, 2001).

This paper presents partial results of a research project for exploring the interdisciplinary influences of HBPLE. The focuses is this paper are on which concepts and citation functions of HBPLE were cited by authors of LIS articles that were published between 1949 and 2013. We analyzed citation frequency trends and the research topics of citing articles to identify emerging trends in the influence of HBPLE on LIS research and to determine which topics in LIS research have involved applying the concepts in HBPLE. In addition, citation context analysis was used to identify the cited concepts and the citation functions of HBPLE; thus, whether the PLE was the most frequently cited concept in HBPLE and the reasons HBPLE was cited were identified. The results may contribute to the understanding how a classic book on linguistics has influenced LIS research.

Methodology

The bibliographic records of LIS articles citing HBPLE published between 1949 and 2013 were searched and collected from the database Web of Science. The LIS journal candidates had to be included in the subject category of "Information Science and Library Science" in the 2012 Journal Citation Reports and the subject category of "Library and Information Science" in the database provided by Ulrichsweb.com. The publication language of articles had to be English and only research articles were collected. Regarding the search strategy used for collecting the citing articles, search terms were combined in two designated fields: the cited author field and publication year of the cited work.

A citing article could have two or more citation contexts referring to HBPLE. Each in-text citation was defined as an independent citation context. Of the 274 citing articles, three were excluded from the dataset because of citation errors existed between the in-text references and reference lists (two articles), or because full-text articles could not be obtained (one article). Finally, we analyzed 260 citing articles including 310 citation contexts. The records of cited concepts were analyzed and divided into several categories. The classification scheme of citation functions was developed based on a temporary classification scheme devised after reviewing previous studies and was modified during the analysis process. The main topic of each citing article was also coded.

Results

Topics of citing articles

Table 1 shows that HBPLE is more associated with bibliometrics and information retrieval research than are other research topics.

Table 1. Distribution of citing article topics.

Topics	No. of articles	Percentage
Bibliometrics	121	46.5
Information retrieval	64	24.6
Information behavior	24	9.2
Information system	12	4.6
Information service	7	2.7
Collection development	7	2.7
Information science	7	2.7
Knowledge organization	7	2.7
Management	5	1.9
Scholarly communication	3	1.2
Resource allocation	2	0.8
Information literary	1	0.4
Total	260	100.0

Cited concepts and citation functions

Table 2 shows the distribution of 17 cited concepts in 11 citation functions. The most frequently cited concept was "Zipf's law" and was mainly used for comparison with other bibliometric laws, whereas the second-most cited concept, the "PLE," was mainly used as evidence.

Among 201 citation contexts referring to the concept of "Zipf's law," 52.2% used the term "Zipf's law," 28.4% used other terms, such as "Zipfian distribution," "power law," "hypobolic distribution," and "rank-size law," and 19.4% contained a statement to describe or imply the concept of "Zipf's law." Although Zipf's law is a well-known informetrics law, not all authors have used the formal term "Zipf's law" to refer to the law emphasizing the relationship between word rank and word frequency.

Although the concept of the PLE, which is derived from Zipf's law, is the focus of HBPLE, the number of citation contexts referring to the PLE was lower than that referring to "Zipf's law." This result ran counter to our assumption that the number of citation contexts referring to the concept of the PLE would be highest. This implies that citing behavior is complicated and that various motivations for citing publications also affect the visibility of cited publications.

Table 2. Distribution of cited concepts according to citation functions.

Cited concepts	Citation functions											
	E	C	RS	H	R	D	E	F	Exp	T	M	Total
Zipf's law	29	38	30	27	21	22	17	7	4	5	1	201
Principle of least effort	15	13	8	6	11	7	1	4	8	3		76
HBPLE	2		2	2	2		1					9
Word distribution	3	1	1				1					6
Human behavior			2									2
Information cycle	2											2
Publication productivity			1	1								2
Rank				1								1
Sample size							1		1		1	3
Information nonuse			1									1
Language analysis	1											1
Lotka's law						1						1
Richer effect										1		1
R.Y. Chao	1											1
Signal information theory								1				1
Social physics								1				1
Optimization problem								1				1
Total	53	52	45	37	34	30	21	14	13	9	2	310

Note: (1)E: Evidence. (2)C: Comparison. (3)RS: Related studies. (4)H: History. (5) R: Relationship (6)D: Definitions. (7)E: Examples. (8)F: Further reading. (9)Exp: Explanations. (10)T: Terms. (11) M: Methods.

The 17 cited concepts were examined by year. Figure 1 shows large fluctuations for the two concepts of "Zipf's law" and the PLE; opposing trends appear. A "falling after rising" trend was observed in the concept of "Zipf's law" whereas a "rising after falling" trend was evident for the concept of the PLE. These opposing trends have resulted in a decreased difference in the annual percentage between the top two cited concepts. Although a close relationship exists between the PLE and Zipf's law, they exert an evidently different influence.

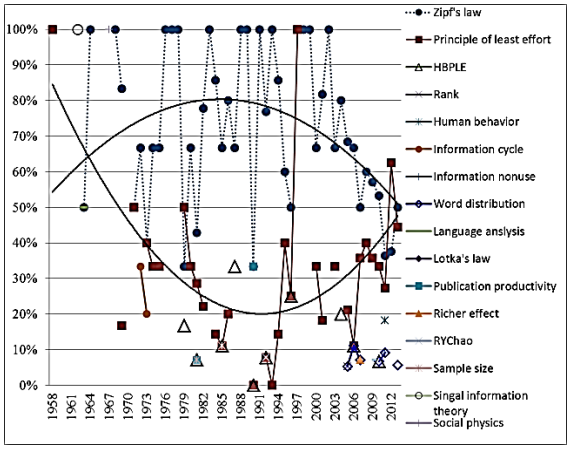


Figure 1. Changes in the percentage of cited concepts by year.

Acknowledgments

This research was supported by a grant from the Ministry of Science and Technology of Taiwan (MOST 103-2410-H-002-172).

References

Austin, B. (2001). Mooers' law: In and out of context. *Journal of the American Society for Information Science and Technology*, 52(8), 607-609.

Chrzastowski, T. E. (1995). Do workstations work too well? An investigation into library workstation popularity and the `principle of least effort.' *Journal of the American Society for Information Science*, 46(8), 638-641.

Chrzastowski, T. E. (1999). E-journal access: the online catalog (856 field), Web lists, and 'The principle of least effort.' *Library Computing*, 18(4), 317-322.

Gratch, B. G. (1990). Exploring the Principle of Least Effort and its value to research. *College and Research Libraries News,* 51(8), 727-728.

Kim, K. S., Sin, S. C. J. (2011). Selecting quality sources: Bridging the gap between the perception and use of information sources. *Journal of Information Science*, 37(2), 178-188.

White, H. (2001). Authors as citers over time. *Journal of the American Society for Information Science and Technology*, 52(2), 87-108.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.

Document type assignment accuracy in citation index data sources

Paul Donner

donner@forschungsinfo.de
Institute for Research Information and Quality Assurance (iFQ), Schützenstr. 6a, 10117 Berlin (Germany)

Introduction

The observed citation counts of publications can be divided by the average of a reference set of similar publications in order to get a relative impact measure. It is customary to define the reference set by publication date, scientific discipline and document type. Different document types (DT) have very different citation distributions, leading to very different results in calculations of indicators when separating reference sets by DT and disregarding this kind of normalization (Sirtes, 2012). Thus, when computing relative impact, the correctness of the assignment of document types to publications is crucial. The correctness of DT assignment in citation indexes has been called into question by studies of van Leeuwen et al. (2007), drawing attention to the treatment of letters and 'research letters' from medical journals as the same type in Web of Science and by Harzing (2003), illustrating how WoS is using some highly questionable assignment criteria. In this contribution DT assignments in WoS (Thomson Reuters, 2013) and Scopus (Elsevier, 2014) by their respective staff are compared to those of the publishers.

Methods and data

For this study data licenced from Thomson Reuters Web of Science and Elsevier Scopus and loaded into SQL databases was used. The databases are part of the infrastructure of the German Competence Centre for Bibliometrics project. Random samples of document identifiers were drawn from the WoS records, stratified by DT as assigned in WoS, restricted to items published in journals. Subsamples of the document types 'article', 'review' and 'letter', as well as of records not assigned to any of those three types (here called 'other') were taken. This follows the convention of distinguishing between 'citable items' and others. They were linked to the Scopus records detailing the same documents using DOIs. It follows that only documents with a DOI are used. In the resulting sample table, only the WoS and Scopus document identifiers and the DOI are saved in a row. The rows were randomized.

To each sample record, bibliographic description data comprised of article title, first author family name and initials, publication year, journal name, volume and issue were queried from the WoS data and saved along with record IDs into a separate table. Student assistants were tasked to search for the article abstract web pages online using the bibliographic information to query Google Scholar and web search. On the individual article web page of the journal, they were instructed to find the officially assigned document type, if specified, and code it as article, letter, review, other or not found. If no type was stated but it was clearly deducible from the abstract or title, this was also accepted.

A sample of 528 publications was analyzed so far, on which the following provisional results are based. For a further 90 publications, no certain DT assignment was possible. Found (true) DT and Scopus/WoS DT were tabulated and classified as true/false positive/negative. From those counts precision and recall were computed for each DT and combined precision and recall as weighted by DT occurrence frequency in the databases. The effect of false DT assignment on publication normalized citation score is measured in percent deviation.

Results

The results depicted in Fig. 1 show that in both citation indexes the accuracy of correct DT assignment is quite poor. WoS gives the correct DT in about 72%, Scopus in about 80% of cases (as weighted by shares of DT in the databases). On average WoS finds about 81% of publications of a given DT while Scopus will return about 73%. Error bars for the DT specific results are 95% posterior probability Bayesian credible intervals for the binomial proportion, using a flat beta prior with both shape parameters set to 1.

These findings necessarily have an adverse effect on the mean field/DT/year specific expected citation rates used as reference standards in obtaining normalized publication level citation scores. To give an idea of the magnitude of this effect, the normalized article citation score (3-year citation window) for publications that were assigned an incorrect DT in WoS was calculated following Waltman et al. (2011).

The differences between incorrect and correct score in percent of the correct score are plotted as a histogram in Fig. 2. Publications with zero citations are not used (N_0=34), since no difference could manifest.

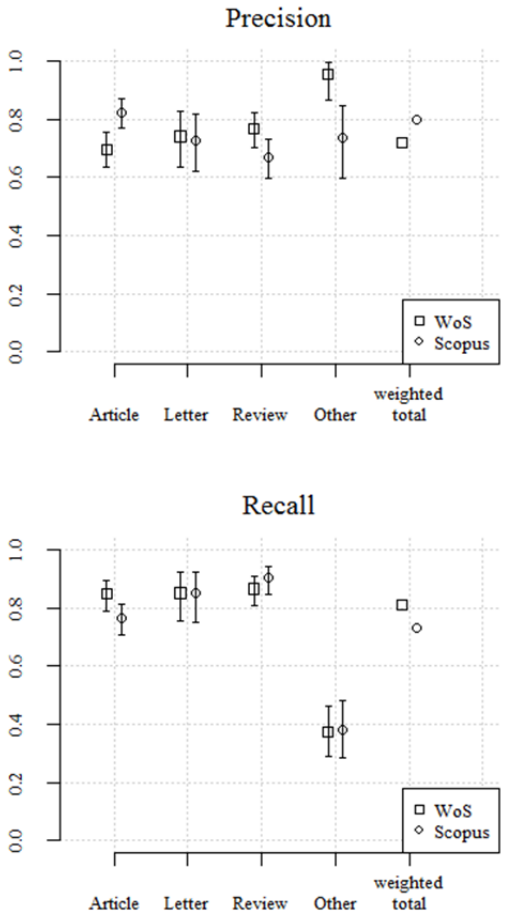


Figure 1. Precision and recall per document type in WoS and Scopus (N=528).

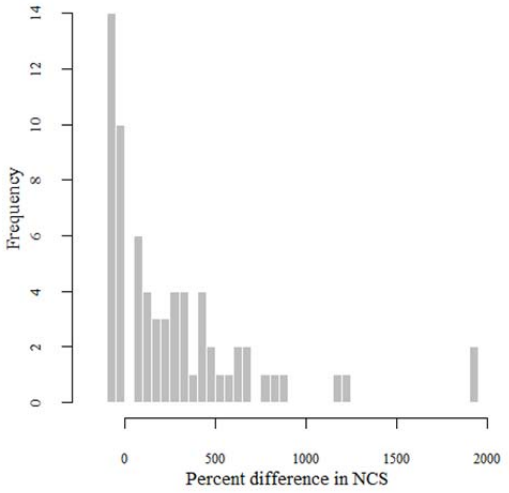


Figure 2. Percent difference in normalized citation score per document for those with wrong DT assignment in WoS (N=68).

Conclusion

Document type assignment is unreliable in both Web of Science and Scopus and will cause large errors in publications' normalized citation scores and consequently derived indicators such as field-normalized mean citation rate.

References

Elsevier B.V. (2014). Scopus Content Coverage Guide 07.14. [accessed 2015/02/06] http://www.elsevier.com/__data/assets/pdf_file/0011/242489/Content-Coverage-Guide.pdf

Harzing, A.W. (2013). Document categories in the ISI Web of Knowledge: Misunderstanding the Social Sciences? *Scientometrics*, 93(1), 23-34.

Sirtes, D. (2012) How (dis-) similar are Different Citation Normalizations and the Fractional Citation Indicator? (And How it can be Improved). É. Archambault, Y. Gingras & V. Larivière (Eds.), *Proceedings of 17th International Conference on Science and Technology Indicators (STI)*, Montréal: Science-Metrix and OST, 894-896.

Thomson Reuters (2013). Web of Science® Help. Searching the Document Type Field. [accessed 2015/02/06], http://images.webofknowledge.com/WOKRS59B4/help/WOS/hs_document_type.html

Van Leeuwen, T. N., Van Der Wurff, L. J., & De Craen, A. J. M. (2007). Classification of "research letters" in general medical journals and its consequences in bibliometric research evaluation processes. *Research Evaluation*, 16(1), 59-63.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, 5(1), 37-47.

Measuring the Impact of Arabic Scientific Publication: Challenges and Proposed Solution

Raad Alturki [1]

[1]*ralturki@ccis.imamu.edu.sa*
Department of Computer Science, Al-Imam Mohammad Ibn Saud Islamic University,
P.O.Box: 5701, Riyadh:11432 (Saudi Arabia)

Introduction

Citation Indices are very useful tools that were firstly used to help finding articles easily and then, used to provide information about research output. They can be used as indicator to measure research performance, provide information about trends in research and compare and rank the research output of countries, institutes and authors. It is well known that English is the universal language for science and technology and that have resulted in having many citation indices like Web of Science (Formerly ISI) and SCOPUS. It has been reported in the literature that such Indices overlook and hide publications in other languages (van Leeuwen et al., 2001) and that -with other reasons- have resulted in having indices for other languages like Chinese, Portuguese and Korean. Arabic publications is one of the least represented in the scientific community despite its been spoken by more than 200 million which makes it the fifth spoken language in the world (Gordon Jr., 2005). This work investigates the possibility of making a Citation Index for Arabic literature and addresses the challenges associated with that. This is supported by initial implementation of web based Arabic Citation Index (ACI).

Challenges

This section discusses challenges associated with non-English citation indices with special focus on the one dealing with Arabic literature. In order to have citation index for any language, it is very important to make it integrate with other English-based indices. Non-English citation indices should be able to read citations from other indices in order to see how any article or language is impacting the scientific community. This raises some issues of how to make cross languages referencing; if an article written in Chinese has cited other article in Korean, how the Chinese/Korean indices will identify this citation. This problem is not easy to be solved unless if there is a well established standardization for citations which allows identifying any article in any language. Such identifier should be unique across the globe and can be used in every citation. Luckily, Digital Object Identifier (DOI) can be used to serve this purpose while the adoption of using DOI in referencing is not yet being very popular as citation styles are still not considering that as part of the cited article. Having DOI as a compulsory in each citation style makes it easier for articles to be identified, then cited and discovered in citation indices across languages.

Unfortunately, there is no enough information about the scientific contribution written in Arabic. One of the most accurate information we found is the number of periodicals that have ISSN. According to a report by ISSN foundation, in 2012 there were 4489 new periodical record in Arabic which makes it the 26th most registered language in the world. The ISSN records do not represent only scientific journals but it registers any types of periodical. Also, there is a report by Thomson Reuters about the contribution of Arab countries recorded in their databases. The report shows that the number of scientific documents produced in those countries is around 13,574 in 2008 (Adams et al., 2011) where most of the written articles are in English. In fact, there are many journals written in Arabic that are not well recognized in the internet and digital libraries. We have noticed that Arabic scientific journals are still focusing on publishing printed format with no much focus on the electronic version.

In reality, there are some digital libraries that aggregate articles of major Arabic journals and provide electronic versions of such articles. However, having seen some of the main digital libraries and aggregators in Arabic, we still believe such aggregators have some issues as they provide the articles as scanned documents that cannot be indexed automatically. Also, such digital libraries do not have the full bibliographic information like title, abstract, authors, year of publishing, publisher name, volume, ISSN and list of references. Having bibliographic information is vital for building any citation index as they are the raw data to draw the relationship between article and scientific work in term of citations. If bibliographic information is not available for any reason, the PDF electronic version of the article could be used to extract the bibliographic information. Extracting such information from any electronic file can be done with some challenges if the article is saved as text rather than picture. The process becomes very

sophisticated if article is saved as picture where scanning should be done properly. Then Arabic text recognition algorithm should be used to recognize text used when current algorithms in Arabic are not reliable and accuracy rate is low.

Additional challenge in working with Arabic literature is the lack of standardization of the structure and the location of different section in articles. Any software that scan or parse the paper will make some assumptions of the location of the title, authors and abstract. Google scholar software that extract bibliographic information from files directly without having bibliographic information assumes that first line is the title which is written in large font. It has been stated in a study of Arabic journals that "instructions to authors" are generative and are not precise enough (Alkholaifi, 2001). That results in having different interpretations of instructions specially in using referencing style. Variations in formatting could happen at different places of the article, including authors' names, authors' salutation (Dr, professor), availability of abstract and list of references. List of references can be written in mixture of two languages at the same time (Arabic and English) which makes extraction harder. The extraction program should be able to work with different languages at the same time and be able to differentiate between different citing styles.

Extracted Information from article could include errors that can be stored in the index. The program should be aware of such errors and correct them before storing. Detecting errors is not an easy task as it should understand the context of the information. Names sometimes could be recognized as error or misspelled words as some names could have different variations or do not have a direct meaning especially if the name is not Arabic. After the information about any specific word is stored in the index, a query can be done to find a specific article or articles in certain subject. For this reason, search query should be able to consider all possible errors that user might have done when entering the keywords beside the stemming and lemmatization process that happens at indexing phase. In fact, there are several Arabic spelling correction techniques (Manning et al., 2006; Attia et al., 2012; Larkey et al., 2002; Rytting et al., 2011; Shaalan et al., 2012). Using such techniques will be of great important in implementing any Arabic based citation index. These techniques in Arabic are similar to other languages with few differences include the morphological analysis and context understanding of the language where Arabic language is complex in comparison to English.

The proposed system

The overall architecture of the system is shown in Figure 1 where it shows the five main components: Crawler, Parser, Matcher, Database and User Interface. This architecture is inspired by the typical design of search engines as they share similar concepts. One major difference between the two systems is that citation indices use citations as way to rank and measure the impact of an article whereas search engines normally uses the links and other metrics as a way to rank sites and documents.

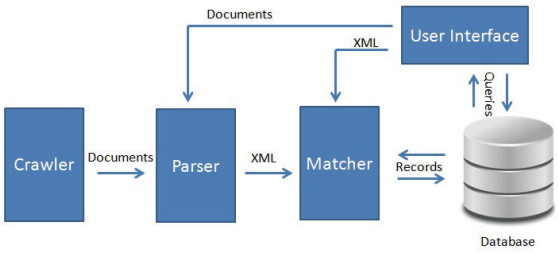


Figure 1. The proposed Architecture of ACI.

References:

Adams, J., King, C., Pendlebury, D., Hook, D. & Wilsdon, J. (2011). Global research report. Middle East, *Evidence*, Thompson Reuters.

Alkholaifi, M. (2001). Documenting citations: an analytical study of publishing policy in some journals. *Journal of King Fahd National Library* vol. 6.

Attia, M., Pecina, P., Samih, Y., Shaalan, K. F., & van Genabith, J. (2012). Improved Spelling Error Detection and Correction for Arabic. *Proc. COLING,* 103-112.

Gordon Jr, R. G. (2005). Ethnologue: Languages of the World, Dallas, Tex.: SIL International. *Online version*: http://www. ethnologue. com.

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002). Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. *Proc. 25th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 275-282.

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval.* Cambridge: Cambridge University Press.

Rytting, C. A., Zajic, D. M., Rodrigues, P., Wayland, S. C., Hettick, C., Buckwalter, T., & Blake, C. C. (2011). Spelling correction for dialectal Arabic dictionary lookup. *ACM Transactions on Asian Language Information Processing* (TALIP), *10*(1), 3.

Shaalan, K. F., Attia, M., Pecina, P., Samih, Y., & van Genabith, J. (2012). Arabic Word Generation and Modelling for Spell Checking. *Proc. LREC*, 719-725.

Van Leeuwen, T., Moed, H., Tijssen, R., Visser, M., & Van Raan, A. (2001). Language biases in the coverage of the Science Citation Index and its consequencesfor international comparisons of national research performance. *Scientometrics*, *51*(1), 335-346.