STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Open Science – a loosely coupled discourse? Comparing Open Science and Open Innovation from a bibliometric point of view

Clemens Blümel[*], Florian Beng[**]

[*]*bluemel@dzhw.eu*
DZHW – German Centre for Higher Education Research & Science Studies (Berlin Subsidiary), Schützenstraße 6a, Berlin, 10117 (Germany)

[**]*beng@dzhw.eu*
DZHW – German Centre for Higher Education Research & Science Studies (Berlin Subsidiary), Schützenstraße 6a, Berlin, 10117 (Germany)

Introduction

Open Science is currently high on the agenda of science policy officials, funding agencies and scholarly associations. It is understood as an umbrella term that covers different initiatives aiming at establishing new forms of how research is produced (Open Data), reviewed (Open Peer Review), communicated (Innovative Dissemination or Open Access), and evaluated (Open Metrics or Alternative Metrics). While we still lack a clear definition of Open Science, many of the propagators and observers agree in that the opening up of the science system should account for the whole process of scientific knowledge production, reaching from discovery, data sampling, review to the final publication (Nielsen 2012). Open Science is therefore associated with fundamental changes in the scientific publishing and communication system. The concept of Open Science itself, however, is only rarely understood. Several articles argue (Dickel & Franzen, 2016; Fecher & Friesike, 2014; Nielsen, 2012) that there are multiple ways of how the term is understood and how openness should be implemented as a consequence. Up to now, there are only few studies which attempt to explore the structure of the academic discourse in Open Science. It is still unclear, how the concept of open science is integrating scholarly debates and as to whether it has established structures similar to other scientific fields or debates. This is rather different to another concept proposing openness in the social, scientific and economic world, Open Innovation, which seems to spur very decisive discussions (Huizingh, 2011).

As has been suggested elsewhere (Crane, 1972; Gläser & Laudel, 2001), bibliometric studies can provide insights into the intellectual or social organization of a research topic or knowledge area. Studying citation patterns, for instance, can provide information about the common knowledge base or the common ways of authoring or authorizing knowledge. In a similar way, the analysis of keyword networks can enlighten the cognitive structure of a given field (Callon et al. 1983). Up to now, however, there is only scarce bibliometric information about the scholarly discourse of Open Science. Against this backdrop, we have attempted to study the scholarly discourse of Open Science by using bibliometric and network analysis tools. Our goal was to determine how the Open Science discourse has evolved over time, how it is structured and how the different associated terms relate to one another. In order to evaluate as to whether Open Science (OS) has evolved into an established scientific topic, we have attempted to compare OS with another concept which propagates openness in innovation

and research, that is, Open Innovation. Open Innovation has emerged in the field of management studies, but has soon been taken up not only in the academic, but also in the science policy realm.

In bibliometrics, a rich body of work has emerged which aims to inform the study of scientific fields, knowledge areas or disciplines (Gläser & Laudel, 2001; Hicks, 2004). There have been various suggestions for exploring the institutionalization of research topics into mature fields. One way for evaluating as to whether a certain topic has evolved into a mature or proper field, is to analyse the number and distribution of journals in which research is published. For the topic of Open Science, however, such attempts of studying formation appear to be rather inappropriate, since Open Science is a truly interdisciplinary topic with many different sites for publication. Coming from network analysis, one alternative way to explore as to whether a topic has been established as a discourse, is to analyse the cohesion of the networks of its seminal terms. Cohesion means the extent to which a networks remains connected when the nodes with the highest degree of connectedness is removed.

Methods and materials

For our case, we rely on metadata of the bibliographic database Web of Science which covers different scholarly indices. We have analysed the scholarly debate on Open Science in two major work packages which will be described in further detail below. The first consists of the construction of the corpus and the analysis of its major bibliometric attributes, e.g. publication dynamics, citation patterns, and dominant places of publication. The second package consists of an analysis of the scholarly discourse based on keyword networks which we derived from the initial corpus.

First, we employed a traditional bibliometric analysis of the literature based on the Science Citation Index Expanded (SCIE), the Social Science Citation Index (SSCI), the Conference Proceedings Citation Index (CPCI) and the Arts & Humanities Citation Index (AHCI) in Web of Science, searching for the terms "open science" and "open innovation" in the titles and keywords, discovering a seed of publications for the two different corpora.[1] In a second step, we have integrated further keywords and introduced them in the search query for the respective corpora. We have found that open science co-occurs with other keywords and keyword families, such as "open access", "open scholar*", "open research*", "collaboration*", "open data"), leading to a set of 321 publications, covering many different aspects and debates of open science. After analysing the corpus of documents, we found that the corpus only to a limited extent reflected the debate on Open Data in science, which has led us to set up a second search for publications using the keyword "open data". Yet, the first results of this query showed that many publications referred to the Open Data debate in the e-government discourse. Therefore, publications using Open Data as a keyword have been integrated only if they also used "scien*", „academi*", or „publish*" as a keyword, leading to additional 25 publications in the corpus. This has led to a total of 346 publications in the databases of Web of Science for Open Science (including the newly found keywords) and 1601 for Open Innovation. In a third step, we have identified a corpus of citing publications for each set. Publications which cite more than one document (above citation median in the seed of corpus), but which are not themselves included in the initial seed, have been integrated in the different corpora. This approach resulted in identification of 2704 items for Open Innovation and 425 publications for Open Science. Finally, we have checked for duplicates in all the sets and came up with a total of 3144 publications for both corpora in the databases of Web of Science. Metadata for publication year, citation counts, keywords and titles have been extracted in order to account for the evolution of the topic. We identified a list

---

[1] Abstracts have not been accounted for, because this was supposed to lead to too many false positives in the sample.

of highly cited documents, taking into account differences between documents which are highly cited within the corpus of publications and those which are highly cited, but mainly by documents from outside the corpus.

As a second analysis, we have constructed keyword networks based on above mentioned corpus of publications. We extracted all keywords with their respective publication ID and processed them, using the statistical package R. We then transformed the keywords and stemmed them using the wordStem package, a suffix stripper algorithm developed by Martin Porter. We finally processed them to generate a weighted matrix. This list has then been processed with the network analysis and visualization package Gephi. The Force Atlas 2 algorithm has been used to layout the keyword network and modularity maximization to assign them to clusters. We also calculated other network parameters (e.g. centrality measures for nodes and weight distributions for edges), to filter out less important keywords and links.

Results

Evolution of the topic

The analysis clearly shows that Open Science is a dynamic topic within the scholarly community. Figures 1 displays the number of publications per year between 2001 and 2016. These figures have increased rising from 1 in 2001 to more than 100 in 2016. The growth of publications since 2013 has been particularly dynamic: while 32 scientific articles have been published in 2013, this number has increased to 102 in 2016, which means that the yearly output has tripled within this short period. With some uncertainty, it is likely that trend will continue in the years to come. It could be argued that this increase reflects the rising political attention the topic receives. Even more dynamic was the development for specific sub-keywords such as open data. Their yearly output has even grown faster within the last years.

Figure 1: Publication Output for Open Science and Open Innovation
(raw data: Web of Science, processing: DZHW)
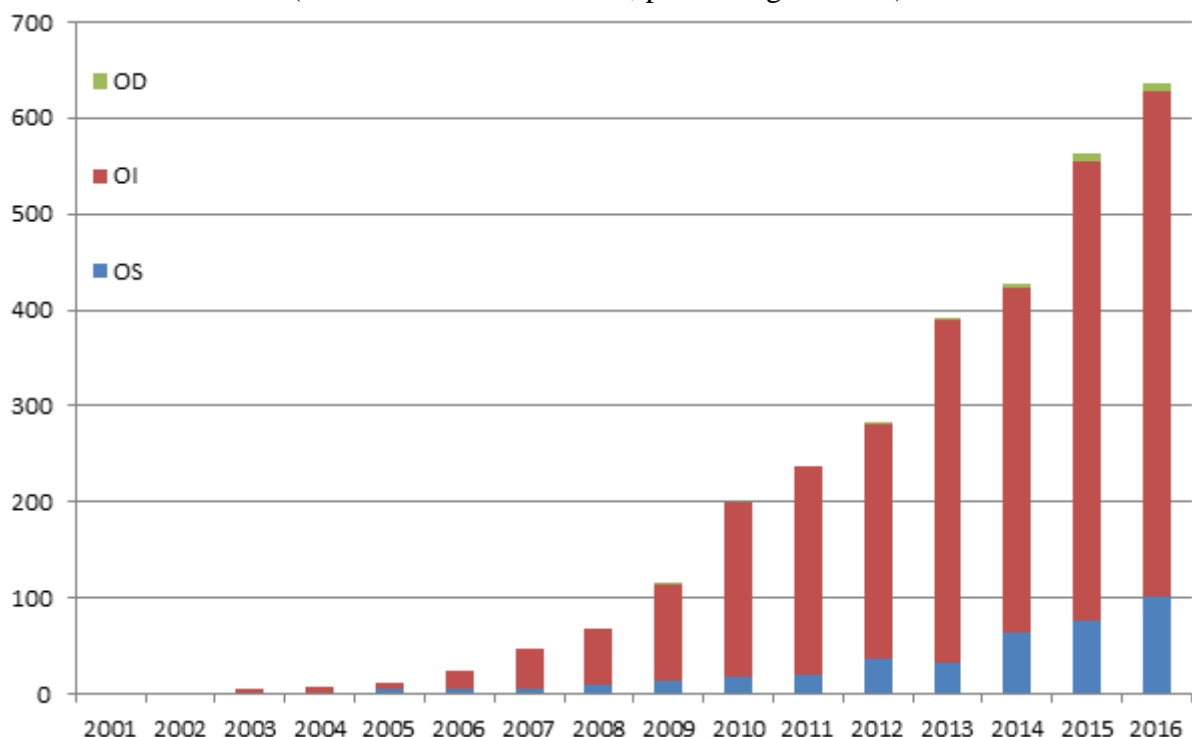
As figure 1 shows, Open Innovation has been much more subject to scholarly discussion. Publishing activity has been much stronger in Open Innovation than in Open Science.

Publications have grown enormously since 2003 which was the year when the concept has been coined by Henry Chesbrough. We can observe a strong increase between 2009 and 2010 which reflects the increasing visibility the concept received when the first review articles appeared. Since then, the growth has been steady and the yearly output of publications has reached 526 in 2016. It is likely that this growth will continue in the near future.

This relative strength in publishing activity is also reflected in higher citation counts for Open Innovation. The most important articles which can be understood as conceptual scrutinizations are among the most highly cited within the respective corpus of Open Innovation. One of these articles is Henry Chesbroughs 'Era of Open Innovation' which was published in 2003. As mentioned above, other reviews, state of the art articles or articles regarding the future development of the field exist and are also among the most highly cited (Dahlander & Gann, 2010; Huizingh, 2011). The existence of such articles also signifies the rising interest and stabilization of Open Innovation as an academic topic (Bastide, Courtial, & Callon, 1989). Furthermore, we could identify in the Open Innovation corpus of highly cited works other established and highly cited papers related to innovation research. For example, this accounts for the paper published by Cohen and Levinthal in 1990 (Cohen & Levinthal, 1990). It can be argued that these citation patterns suggest that Open Innovation shares a common knowledge base with innovation research or that at least there are strong ties to this field.

By contrast, citation figures reveal that the topic of Open Science is not gaining similar attraction among the scholarly communities. There are only very few articles with more than 100 citations. We cannot identify highly cited conceptual papers which would point at some shared knowledge base in conceptual development such as that those of Chesbrough in Open Innovation. Also, we do not find specific documents which aim at ordering or defining the field, such as review or state-of-the-art-articles. However, we find several papers with particular focus on biomedicine. Partly, this can be explained by the strong representation of these fields in the scientific databases. Many topics and issues related to Open Science can be associated with scholarly topics in clinical or laboratory medicine. Yet, there seems to be specific connection to a very recent debate regarding the quality of biomedical publications (Ioannidis John P. A., Oliver, Sander et al., 2014). There is a substantial amount of authors in biomedicine who clearly relate the transition from closed to open science with an increase in transparency, and, as a consequence, in quality. The insistence of the quality issue is also indicated by the dominance of the replication topic in biomedicine (Levy 1990). Despite these publications, however, there seems to be neither indication of a specific disciplinary or scholarly focus nor a specific topical focus of the highly cited papers. Based on this analysis, the distribution of topics seems to be rather diffuse. Taken together, this suggests that Open Innovation does not only seem to gain more attention within the scholarly communities, but that it also appears to be more anchored to specific scientific fields than this could be said for the discourse of Open Science.

Analysis of the keyword networks
In order to find out more about how the scholarly discourse in Open Science is organized, we conducted a network analysis of keywords. Such kind of analysis also allows for establishing claims about the conceptual development of the field. The analysis is concentrated on those keywords which have been provided by the authors themselves (author keywords). We have therefore not included those keywords provided by the databases as those might bias the representation of the respective keywords. As mentioned above, the analysis is based on the layout and modularization algorithms provided by Gephi. This algorithm clusters a network in various subgraphs and plots the nodes and edges based on a measure of attraction. Keywords are placed within the same clusters based on the frequency of their co-occurrence. Results are

presented in the figures 2 and 3. Nodes represent keywords, while publications where they co-occur are displayed as edges. Figure 2 also shows the relationships between the major keywords (keywords clusters) and other subnetworks and their relations to each other, allowing for claims about the general cohesion and integration of the keyword network.

Figure 2: Open Science and Open Innovation Keyword Network (with search terms),
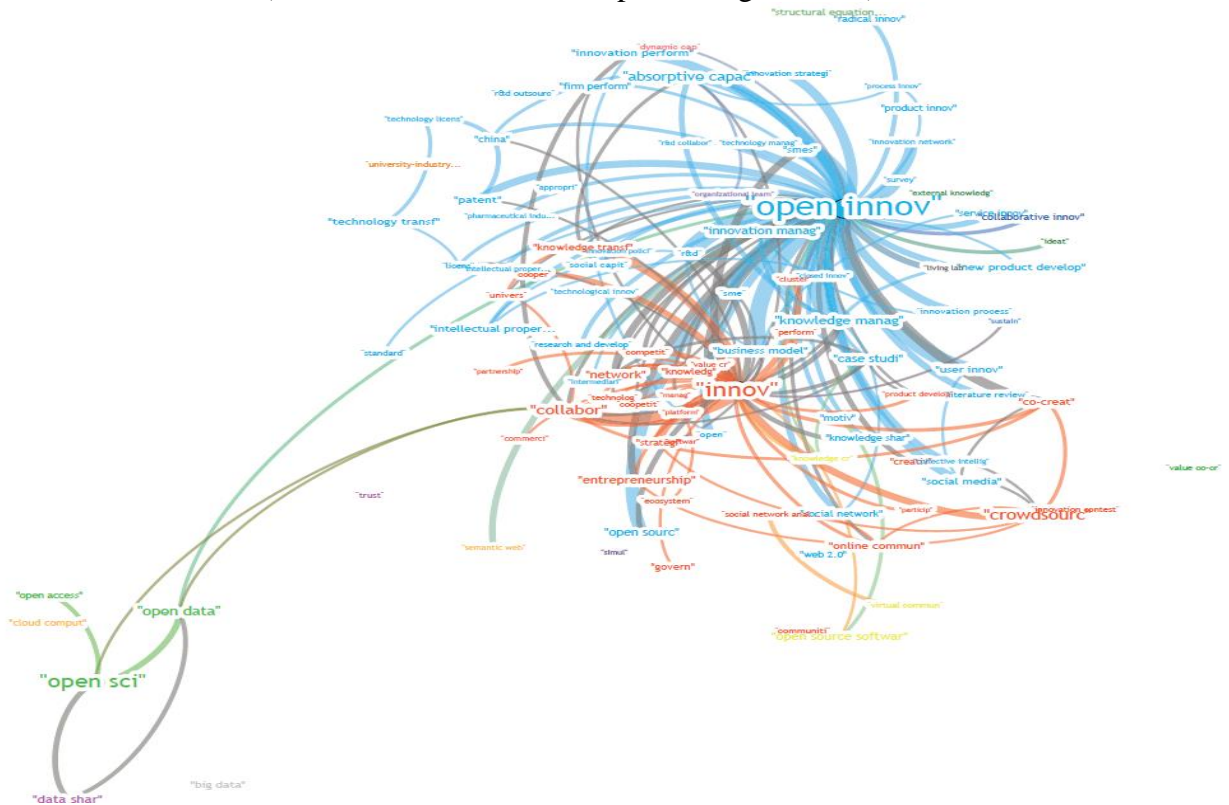(raw data: Web of Science, processing: DZHW)

Figure 2 shows the results for the combined corpus of publications related to Open Innovation and Open Science. The size of the different corpora is also reflected in the size of the different keyword networks of Open Science and Open Innovation. In general, the keyword graph signifies that Open Science and Open Innovation are more or less clearly distinct academic discourses. Many of the important keywords in Open Innovation, such as knowledge management are not related to Open Science. On the other hand, we also find almost no relations between keywords of "open science", such as "open data" or "data sharing", and keywords within Open Innovation. This appears to be counterintuitive, since in science policy, Open Science and Open Innovation are referred as similar and closely related concepts as both of them highlight the value of Openness in knowledge production (Crozier 2015). We also find differences in the intensity of interrelation within the different keyword networks of Open Science and Open Innovation.

Figure 2 also demonstrates that Open Innovation is not only larger in size than the Open Science corpus, but also more strongly and cohesively connected internally. There are a number of densely populated keyword networks such as "user motivation", "collaboration", "transfer", "patents", which have developed strong subclusters within the Open Innovation network. Moreover, we can observe that these clusters are not only connected internally, but also related to each other, in particular through the keyword cluster "innovation". This shows that Open Innovation seems to be clearly nested within the larger structures of innovation research. The strong internal cohesion of Open Innovation becomes even more obvious, if we

have a closer look at the Open Innovation keyword with a lower threshold (Figure 3). We find that Open Innovation keyword network as three-partite structure:

First, there is a strong concentration of interconnected keywords referring to „innovation", which organizes a number of specialized discourses. Second, we find a specific concentration of keywords related to "absorptive capacity", which is a highly established concept within innovation research, And, third, we find a cohesively connected subnetwork of keywords organized around the concept of „Crowdsourcing", which seems to be closely related to other keywords such as "user innovation", "co-creation" or "social media". This network reflects the debate about the digitization of economic activities and the innovation dynamics of digital platforms. These three core networks are particularly connected with each other which can be shown that these components still connect to each other when the central keyword "open innovation" is deleted.

Figure 3: Keyword Network Graph for Open Innovation (without searchterms),
(raw data: Web of Science, processing: DZHW)
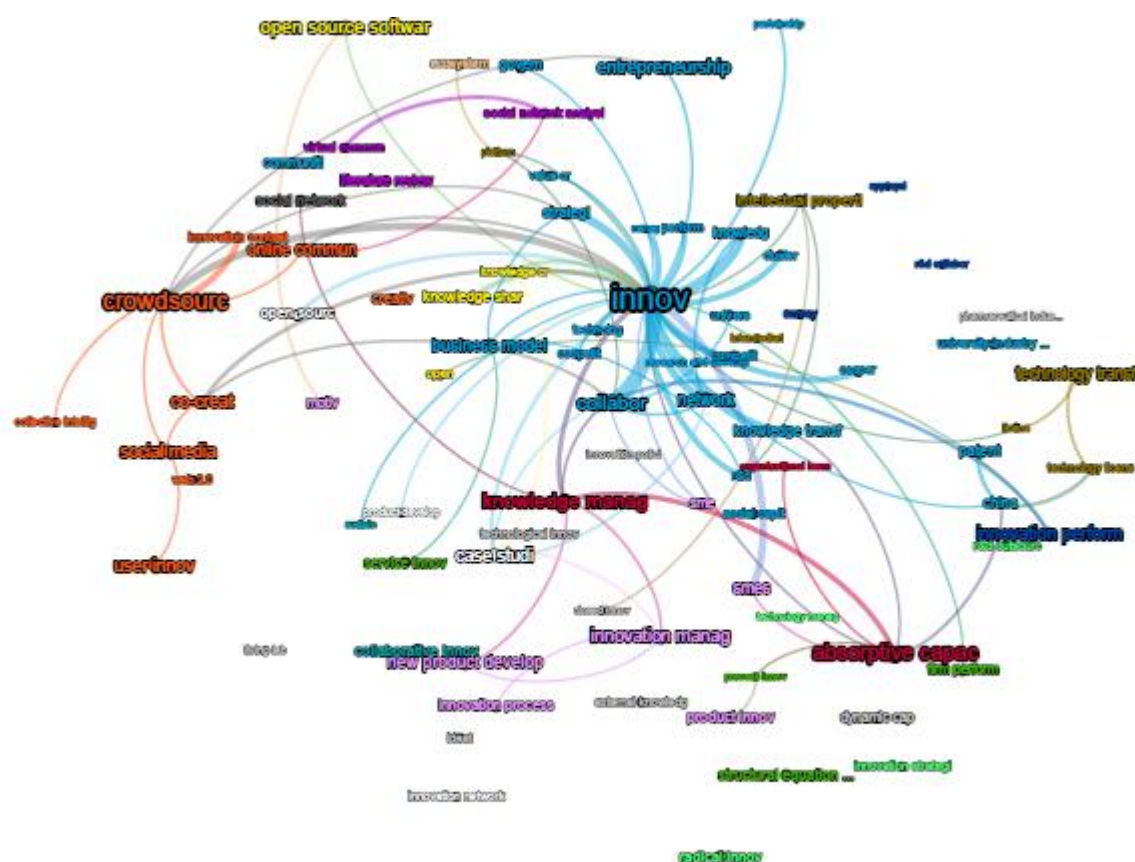
The structure of keywords is rather different in the realm of Open Science. While there are some keywords clusters, such as "open access" or "data-mangement", these have not yet generated connections between each other. Figure 4 shows that there are some major keyword clusters in Open Science, in particular "open access", "data sharing", and "open data", representing a substantial amount of papers dealing with those topics. Other keywords with minor dominance are "knowledge transfer" and "university". It is obvious, however, that these generate only scarcely integrated clusters of keywords which points at a rather low development of the discourse. Moreover, these major clusters are not related to each other. For example, we do not find many edges (that is, publications) connecting open access with open data. The keyword clusters again indicate major influence of the biomedical community

on the discourse. In particular this accounts for the keyword cluster around data sharing which is highly connected with keywords from the biomedical field such as "clinical trials", "neuroimag(ing)" or "biobank". This suggests on how the recent dynamic of the debate on data sharing is related to specific fields and issues in this field.

Due to a threshold applied in Figure 2, some keywords with only few occurrences in the corpus are not displayed, which somewhat biases the representation of Open Science. Therefore, we also provide a visualization of the keyword network within Open Science with lower threshold rates for the occurrences of keywords (Figure 4).

Figure 4: Keyword Network Graph for Open Science (without searchterms),
(raw data: Web of Science, processing: DZHW)
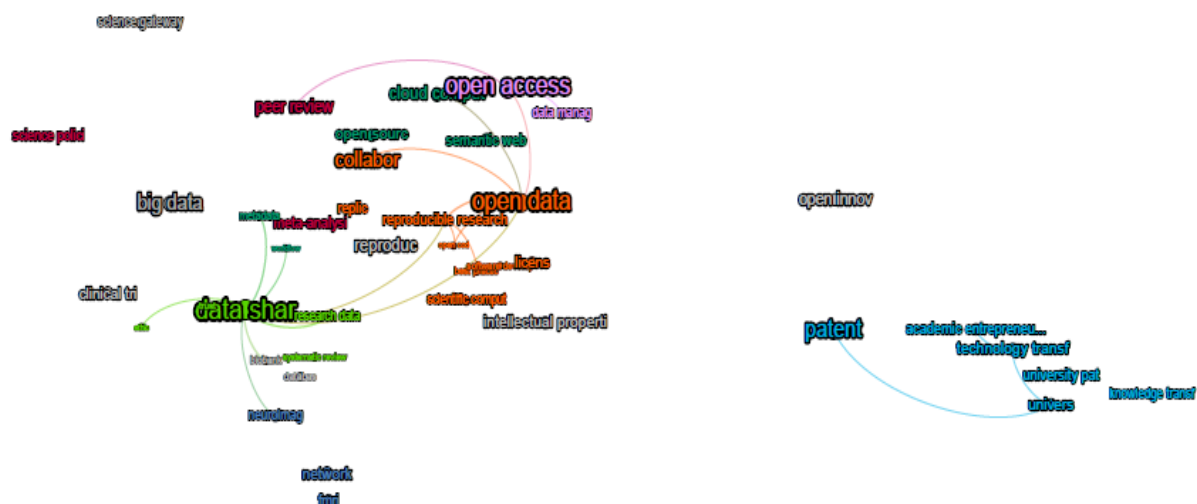
What seems to be particular surprising is that the various keyword clusters are not connected to each other, that is, the keyword clusters of "open access", "open data", and "data sharing" are only scarcely connected with each other and even to a lower extent connected to other important keyword clusters such as "knowledge transfer" or "university-industry linkages". Again this seems to be a rather counterintuitive finding since many Open Science propagators claim that openness is needed for the whole process of knowledge production. This shows the low level of cohesion within the Open Science network which becomes even more obvious when the main keyword "open science" is deleted from the network visualization. Now the distinction between the above mentioned keyword subnetworks becomes clearly visible.

Discussion and conclusion

This contribution aimed at providing insights into the evolution and the structure of the scholarly discourse of Open Science. By comparing Open Science with another academic concept that propagates Openness in Science and Innovation, that is, Open Innovation, we attempted to explore as to whether there are differences in the way these different concepts develop and are nested within the scholarly discourse. The analysis shows that Open Science and Open Innovation are both rising topics, with a strong growth for Open Science since 2013. Yet, the figures also show that the corpus of Open Innovation publications is not only much bigger, but also seems to be more anchored in a scholarly discourse, because they seem to share a common knowledge base with other academic fields, such as innovation research. By applying network analysis of the keywords chosen by authors, it could be also shown that the Open Science discourse is still rather sparsely connected internally when compared with Open Innovation. Furthermore, we could show that Open Science has not developed strong ties to other scholarly debates. There are, however, only few ties that relate the different

realms in Open Science, such as Open Access, Open Data, and Data Sharing. This appears counterintuitive given the claim of Open Science to account for the whole process of scientific knowledge production. These findings have implications, not only for scholars, but also for policymakers, because they contradict the claim of policy makers which often argue that Open Science is already an established and well developed set of practices and concepts (EC 2016; Crozier 2015).The analysis may suffer from the missing coverage of grey literature such as policy papers, position papers or scientific reports. Yet, a preliminary analysis of policy reports and documents of the EC financed publication platform Zenodo[2] reveals that the representation of Open Science in this literature is very different from those established in the aforementioned scientific databases in the respect that these documents' keywords on Open Science are particularly frequent and often relate Open Science keywords with those in the Open Innovation discourse, such as "collaboration", "technology", "intellectual property" or "software". Moreover, we found in this database that Open Science is closely related to other semantics and labels of policy relevance of which some have similar meanings, like "science 2.0", and some relate to broader problems between science and society, such as "responsible research and innovation". These preliminary findings seem to suggest that there are major differences in the keyword networks of Open Science in the (often policy related) grey literature, compared to the keyword networks in the scholarly discourse. Given the heterogeneity and co-occurrence or prehistory of Open Science with similar labels and semantic predecessors, such as science 2.0., open scholarship or responsible research and innovation, we contend that more research is needed in order to understand the heterogeneous discourse of Open Science.

References

Bastide, F., Courtial, J. P., & Callon, M. (1989). The use of review articles in the analysis of a research area. Scientometrics, 15, 535–562.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From Translations to Problematic Networks: An Introduction to Co-Word Analysis. Social Science Information, 22(2), 191–235.

Cohen, W., & Levinthal, D. (1990). A new perspective on learning and innovation. Administrative Science Quarterly, 35, 128–152.

Crane, D. (1972). Invisible colleges: diffusion of knowledge in scientific communities. Chicago: Chicago University Press.

Crozier, Thomas (2015): Science Ecosystem 2.0: how will change occur? European Commission. Brussels (EUR 27391 EN). DOI: 10.2777/67279.

Dahlander, L., & Gann, D. M. (2010). How open is innovation? Research Policy, 39(6), 699–709. https://doi.org/10.1016/j.respol.2010.01.013

Dickel, S., & Franzen, M. (2016). The "Problem of Extension" revisited: new modes of digital participation in science. JCOM, 15(1), A06_en.

European Commission (2016): Open Innovation, Open Science, Open to the World - A Vision for Europe. European Commission - DG Research and Innovation. Brussels. Online verfügbar unter DOI: 10.2777/061652.

Fecher, B., & Friesike, S. (2014). Open Science: One Term, Five Schools of Thought. In S. Bartling & S. Friesike (Eds.), Opening Science (pp. 17–47). Cham: Springer International Publishing.

---

[2] The Zenodo platform also provides keywords (tags) of publications.

Gläser, J., & Laudel, G. (2001). Integration of scientometric indicators into sociological studies: methodical and methodological problems. Scientometrics, 52, 411–434.

Hicks, D. (2004). The four literatures of Social Science. In H. Moed, W. Glänzel, & U. Schmoch (Eds.), Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems (pp. 473–496). Dordrecht: Kluwer Academic Publishers.

Huizingh, E. K. (2011). Open innovation: State of the art and future perspectives. Technovation, 31(1), 2–9. https://doi.org/10.1016/j.technovation.2010.10.002

Ioannidis John P. A., Oliver, Sander, Greenland, S., Hlatky Mark A., Maceleod, M. R., Moher, D., Schulz Kenneth F., & Tibshirani, R. (2014). Research: increasing value, reducing waste 2: Increasing value and reducing waste in research design, conduct, and analysis. Lancet, 383, 166–175.

Levy, David M.; Feigenbaum, Susan (1990): Testing the replication hypothesis. In: Economics Letters 34 (1), 49-53. DOI: 10.1016/0165-1765(90)90180-9.

Nielsen, M. (2012). Reinventing discovery: the new era of networked science, Princeton: Princeton University Press. Princeton NJ: Princeton University Press.