Scoping paper
Knowledge Exchange study on metrics for datasets

The Knowledge Exchange (KE) Working Group on Research Data is seeking a professional consultant or a team of consultants to produce a landscape study on metrics for datasets from both a cultural and a technical point of view. Any bidding party should have expertise with respect to the publication, citation, use and reuse of research data and should be aware of current discussions on the assessment of research performance.

Introduction

The "metric of science" is a field which emerged after World War II. So-called scientometrics or quantitative studies of science were a response to the growth of the science system and its impact on technological innovation eventually leading to the raise of knowledge-based economies. Public spending on the sciences, or academia as a whole, needed to be monitored. Indicator systems as the Science and Technology Indicators of the NSF or comparable European systems account for human capital and R&D investment as the input side to science.

The output of scientific production is usually measured in terms of publications, citations, and patents. It was the innovation of citation indexing and the related establishment of bibliographic databases (as Web of Science, or later SCOPUS) which allowed indicator-based evaluation of scientific productivity to be developed as a craft of its own. These methods are partly still controversial, as there is no consensus on the truly scientific production of these metrics or on the appropriate use of them in research evaluation and assessment. In addition, there are areas of scientific practice which are not or not easily measurable.

Among them is the (re)use of data. Large parts of science rely on "Big Data" creation, sharing and curation – examples include high energy physics, bioinformatics, medical-clinical information, but also data in the domain of cultural heritage.[1] Similar to instruments and tools, data are among the research technologies applied by different fields. But research technologies can play a hidden role, sometimes as part of the ubiquitous infrastructure, and their precise contribution to research output cannot easy to be measured.

The digital age allows for even more easy access to data and information than ever before. At the same time, huge efforts have been made to create research infrastructures to produce, process, exchange and store data. How can the impact of those efforts be made visible? Can resources for research and their use and re-use be mapped? How can organisations operating as information providers assess their contribution to basic and applied research? To what extent are the pre-

[1] See e.g. the report „One Culture. Computationally Intensive Research in the Humanities and Social Sciences" by Christa Williford and Charles Henry, June 2012: www.clir.org/pubs/reports/pub151/pub151.pdf

conditions for research – such as data access – measurable: directly, implicitly or maybe not measurable at all? How could individual researchers benefit from attempts at measuring the uses of their data? And how could funders, research performing organizations and universities take the contribution of researchers to the data infrastructure into account?

Those questions are currently discussed from different perspectives and in isolation. Computer science, and in particular the Linked Open Data movement, creates new data models and representations which *in principle* allow a data web (semantic web), however their implementation is just at the beginning. In fact, a lot of data are still isolated from each other and are not recorded and documented in a systematic way across scientific disciplines. Research infrastructures support the data life cycle, but often they are designed around specific communities, and not always also equipped with features that monitor their use. Science and technology studies and science policy studies look into the data (re)use behaviour of different scientific communities, but often don not connect to quantitative studies of science. As a result, issues such as data formats, data citations, open access to data, data re-use, data management plans as prerequisite for funding, and issues of research assessment are mainly discussed in very different fora and not connected to each other. Moreover, there is a tendency to separate the technological aspects (the backbone for a possible data metrics) from the actual research practices, issues of quality assurance, and social and political behaviour (science dynamics and data cultures), not to mention the necessary context of a critical-historical account of metrics in past, present and future.

Motivation for the study

Increasingly, the necessity to provide open access to data is recognized as a foundation for good science and scholarship. Researchers are increasingly encouraged to share their research data or are requested to share data from projects for which they apply for funding. Data sharing though has a different status and function in different scientific communities, which have thus moved at different speeds. Among the most significant barriers and enablers for data sharing are researchers' motivations, linked closely to mechanisms for professional recognition and reward. Giving credit to researchers who share their data by recognising reuse of that data as a scholarly achievement is therefore viewed as an important incentive. An appropriate reward system will likely require metrics – but these must be robustly founded and fair. Examples of things to which metrics may be applied include data publications, data citations, and usage statistics. Obstacles to the integration of data-related metrics to academic systems of professional recognition and reward may be considered technical and cultural. For example, any metrics can only be implemented on top of clear documentation schemes for research data and their use. Just as importantly, the attitudes of researchers and other stakeholders need to be understood when considering mechanisms to reward researchers contributing data to the open research data infrastructure. A number of recent documents call for the development of such mechanisms, and the commissioned study is supposed to contribute to discussions in this field by making qualified statements on the possibilities of using metrics for data sets in the context of research evaluation and assessment considering that:

- The Knowledge Exchange (KE) vision is "To make a layer of scholarly and scientific content openly available on the Internet."[2]

[2] http://www.knowledge-exchange.info/Default.aspx?ID=68

- The **Surfboard for Riding the Wave (2011)** report for the Knowledge Exchange includes the recommendation: "Advocate making published data sets and citation metrics count in research assessment exercises in research assessment exercises in the four countries including, as a preliminary step, the registration of datasets in the annual reports of research institutes and universities"[3] (p. 32).
- The Communication from the European Commission **Towards better access to scientific information: Boosting the benefits of public investments in research[4] (2012)** states that: "Many researchers and innovative enterprises are reluctant to share what they perceive to be 'their' data and are concerned that others will unfairly benefit from their efforts. Researchers, moreover, may not want to invest time in the practicalities of depositing their data. Systematic reward and recognition mechanisms for data sharing, such as citation mechanisms and measurements of the data citation impact, are not yet in place" (p. 7).
- The European Commission **recommendations on access to and preservation of scientific information (2012)** call for "[I]nstitutions responsible for managing public research funding and academic institutions that are publicly funded [to] assist in implementing national policy by putting into place mechanisms enabling and rewarding the sharing of research data" (p. 6-7).[5]
- The ALLEA declaration **Open Science for the 21st century (2012)** states that: "Scientists and their organisations should apply open sharing principles to the ***data*** that underpins such publications, including 'negative' results; measures should be put in place for quality assurance and preservation of such data for re-use" (p. 5).[6]
- The report **Science as an open enterprise (2012)** by the Royal Society includes the recommendation to "develop more sophisticated systems of attributing credit for researchers' development and dissemination of data resources – and crucially the use of such resources by others" (p. 72), June 2012.[7] DFG Strategy Paper: **Taking Digital Transformation to the Next Level**. The Contribution of the DFG to an Innovative Information Infrastructure for Research (2012), July 2012 (p. 15).[8]

Aims of the study

The Knowledge Exchange Working Group on Research Data commissions a study that will inform the research performing community, the information infrastructure providers, funding agencies and policy makers about the status quo in the area of science metrics, specifically with regard to the area of research data. This study will include an overview of the domain considering existing solutions, a critical assessment of possibilities for their use and suggestions for further actions.
In particular, this study will

1. provide an overview of the status quo in the area of science metrics, specifically with regard to the area of research data, taking into account the status quo of research data citation;

[3] http://www.knowledge-exchange.info/Default.aspx?ID=469
[4] http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf
[5] http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf
[6] http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/allea-declaration-1.pdf
[7] http://royalsociety.org/policy/projects/science-public-enterprise/report/
[8] http://www.dfg.de/download/pdf/foerderung/programme/lis/strategy_paper_digital_transformation.pdf

2. critically assess the current solutions and developments in this area considering the risks and benefits involved, possibly including observations on the differences between quantitative and qualitative metrics, and the appropriateness of their use for different purposes;
3. consider technical aspects along with cultural implications taking behavioural components of crucial stakeholders into account;
4. sketch possible uses of current solutions in the context of research evaluation and assessment;
5. consider possible developments and unintended effects;
6. distil requirements, suggestions and possible actions for the development of the field over the next five years (2013-2018) for the different stakeholders involved.

Scope of the study

The study is to provide an overview of the various existing methods in different scientific communities that result in or may be adapted for reliable metrics to measure the impact of available, shared datasets. This should include developing an understanding and mapping of disciplines according to their "data richness" and possibly with regard to the types of data produced and shared. There have been various innovations in the identification of digital objects (e.g. applying DOIs to research data) and ways of identifying authors or creators are being explored and promoted (ORCID). Such initiatives may have a bearing on the mechanics of how metrics can be applied to the citation of research data, but other technical components should be surveyed (e.g. ARKs, Handles, URIs). A variety of data archiving platforms have been developed (DataVerse, CKAN) and document repository software has been adapted to the needs of data (e.g. DSpace, ePrints, Fedora). Initiatives like Dryad provide a repository for a broad research community to deposit data substantiating scholarly articles. The role of social media in research practices is expected to have an influence for data practices and their eventual metrics. Figshare[9] provides a platform for data sharing with some typical social media metrics (views, likes, downloads). The Impact Story[10] experiment and the alt-metric[11] product (like Figshare, from Digital Science) reflect this trend and should be surveyed and analysed; the effect of such (also commercial) developments for metrics around data sharing should be considered. A reflection on possible (unintended) effects of a future "data metrics" should be part of assessments and recommendations that result from this study. The study should therefore not only just take into account the bibliometric and technical aspects, but should also approach the matter from the perspective of evolving trends in scientific communication from a cultural perspective. The attitudes of various stakeholders should be analysed. Such stakeholders are researchers, research performing organisations, research administrations, data archives and data sharing publishers as well as research funders. Of particular importance will be the close involvement of researchers who do already share data as well as of researches that are reluctant to share data. Quality assurance in the form of peer review of datasets is viewed as a related topic, but placed out of scope due to the added complexity this brings in an already complicated field.

[9] http://figshare.com/
[10] http://impactstory.it/
[11] http://www.altmetric.com/

Suggested Literature

- Mooney & Newton. (2012). The Anatomy of a Data Citation: Discovery, Reuse and Credit. *Journal of Librarianship and Scholarly Communication* 1(1):eP1035.
- Lane. (2010). Let's make science metrics more scientific. *Nature, 464*, p. 488-489.
- NSF (National Science Foundation), 2011. Changing the Conduct of Science in the Information Age. 8 November 2010. s.l.: NSF. Retrieved August 21 2012, from the NSF website: http://www.nsf.gov/pubs/2011/oise11003/oise11003.pdf
- Wouters & Costas. (2012). *Users, Narcissism and Control*. Retrieved May 23, 2012, from the SURF website: http://www.surf.nl/en/publicaties/Pages/Users_narcissism_control.aspx
- Van der Graaf & Waaijers. (2010). *Over kwaliteit van onderzoeksdata.* Retrieved May 23, 2012, from the SURF website: http://www.surf.nl/nl/publicaties/Pages/Verkennendonderzoek.aspx
- Relevant publications from the competence centre for bibliometrics with regard to indicator assessment: http://www.bibliometrie.info/en/publications/publications.html
- German Council on Science and Humanities: Recommendations on the Assessment and Management of Research Performance. http://www.wissenschaftsrat.de/download/archiv/1656-11_engl.pdf

Other relevant developments and sources to consider as input

- MESUR project, http://mesur.informatics.indiana.edu/
- SURE2 project, http://www.surf.nl/en/projecten/Pages/SURE2.aspx
- The ANDS data citation forum: http://community.ands.org.au/viewforum.php?f=181&sid=b438cf206e0efabdc55bffb2d268fa07
- Data Citation Index - Web of Knowledge: http://wokinfo.com/products_tools/multidisciplinary/dci/and http://wokinfo.com/media/pdf/DCI_selection_essay.pdf
- Reilly, S., Schallier, W., Schrimpf, S., Smit, E., Wilkinson, M. (2011). Opportunities for Data Exchange Report on Integration of Data and Publications. http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf
- Outcomes from the conference on Scientometrics 2012 http://www.uni-regensburg.de/library/scientometrics/index.html and other relevant events.

The author(s) of the study are responsible for researching further relevant material.

Output

The outcomes of the study will be presented in a written report targeting a broad audience ranging from individual researchers to research organizations and research institutions (e.g. universities), and from infrastructure providers (such as data centers and libraries) to research and research infrastructure funders. The study will be presented at a workshop on 11-12 April 2013 in Berlin, Germany. The study will be made available under the Creative Commons Attribution 3.0 License. The Knowledge Exchange will be responsible for format and the design of the publication.

Services required

Knowledge Exchange is seeking a (team of) consultant(s) to conduct and report on the study as laid out above. Candidates are invited to briefly outline their expertise on the topic; propose the methods that they think would be most appropriate to achieve the aims and objectives set out above; and provide a brief concept for the realization of the study, including a proposal for the title and structure of the study. We encourage innovative electronic documents that include examples or evidence in multimedia formats, if appropriate. Any bidding party should provide Knowledge Exchange with the budget needed to realise the text. Any bidding party should be willing to discuss preliminary results and open questions in telephone or video conferences with the Knowledge Exchange Working Group on Research Data and should be available to participate in the workshop in Berlin on 11 and 12 April 2013 in order to present the findings of the study. Any bidding party should provide a budget needed to realise the text along with the bid. Expenses related to the workshop in Berlin (i.e. travelling and accommodation) will be reimbursed separately by the Knowledge Exchange.

Timeline

The outcomes of this study will be input for a workshop titled “Making Data Count” planned for 11 and 12 April 2013 in Berlin, Germany. A final draft is thus to be delivered by March 2013.

Milestone	Date
Deadline for proposals	Wednesday 7 November 2012, 17:00 (pm) CET
Award of contract	Monday 19 November 2012
First draft	Monday 11 February 2013
Comments on first draft	Monday 25 February 2013
Final draft	Monday 18 March 2013
Presentation	Thursday or Friday, 11 and 12 April 2013

Contact

Please send your bid to the following contacts. For any inquiries relating to the bid, please contact the Knowledge Exchange Coordinator

Keith Russell
Knowledge Exchange Coordinator
Danish Agency for Culture
H.C. Andersens Boulevard 2, 2nd floor, DK-1553 Copenhagen V Denmark
Tel +31 (0)30 234 66 99, Fax +31 (0)30 233 29 60 |
kru@knowledge-exchange.info

or the Knowledge Exchange Research Data Working Group

Dr. Angela Holzer
Programme Officer
German Research Foundation - Scientific Library Services and Information Systems-
D-53170 Bonn
Tel. +49 (228) 885-2358, Fax +49 (228) 885-2777
angela.holzer@dfg.de