OPEN ACCESS TO DATA AND THE 'BERLIN DECLARATION'

Jens Klump*, Joachim Wächter and the STD-DOI Consortium

GeoForschungsZentrum Potsdam, Data Center, Telegrafenberg, 14473 Potsdam, Germany
Email: jens.klump@gfz-potsdam.de

ABSTRACT

The 'Berlin Declaration' was published in 2003 as a guideline to policy makers to promote the Internet as a functional instrument for a global scientific knowledge base. Since knowledge is derived from data, the principles of the 'Berlin Declaration' should similarly apply to data. Today, access to scientific data is hampered by structural deficits in the publication process. Data publication requires incentives for authors to publish data, availability of published data in long-term repositories, and an adequate licence model to protect the intellectual property rights of the authors, yet allowing further use of the data by the scientific community.

Keywords: *open access, access to data, science policy, intellectual property rights, creative commons licence*

1 INTRODUCTION

On 22 October 2003, a group of leading research institutions and research funding institutions published the 'Berlin Declaration on Access to Knowledge in the Sciences and Humanities' (Berlin Declaration, 2003) in order to "[…] promote the Internet as a functional instrument for a global scientific knowledge base and human reflection and to specify measures which research policy makers, research institutions, funding agencies, libraries, archives and museums need to consider." The Berlin Declaration has since been signed by 47 scientific bodies worldwide and is supplemented by the 'Communiqué on Science, Technology and Innovation for the 21st Century' issued at the OECD ministerial meeting, 29-30 January 2004.

2 DATA PUBLICATION TODAY

Since scientific knowledge is ultimately derived from data, the 'Berlin Declaration' and the OECD Communiqué should apply to data as well (Arzberger et al., 2004). However, the scientific discourse today is hampered by structural problems in the publication process (Figure 1). The size of the data sets used in a scientific publication often prohibits their publication as data tables and, as a result, data used as the basis of a publication are rarely published anymore. The lack of access to scientific data is an obstacle to interdisciplinary and international research. It causes unnecessary duplication of research efforts and the verification of results becomes difficult, if not impossible (Dittert et al., 2001). In addition, cases of scientific misconduct in recent years have highlighted the importance of making scientific data accessible. As a consequence, the German Science Foundation adopted access to data as part of their policy in their 'Recommendations for Good Scientific Practice'. To make this policy effective, scientists themselves need to be convinced that preparing their data for online publication is a worthwhile effort. It would be an incentive to the author, if publishing of data became a citeable publication, which would add to his reputation and ranking among his peers.

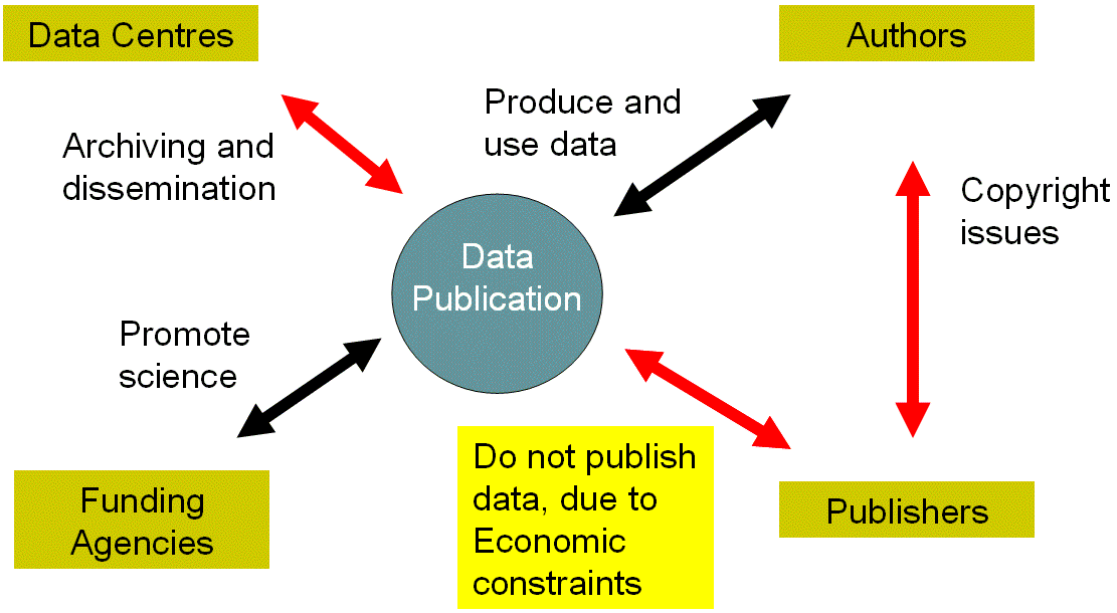

Figure 1: Stakeholders in data publication. Economic constraints in many cases prohibit data publication in 'traditional' scientific journals.

For data to be citeable it is necessary that they can be referred to in a persistent way. Simply making data available though the 'web' is not enough. The location of internet resources, and thus their URL, may easily change, which in most cases means to the user that the data are lost. Therefore, a prerequisite for data access via the internet is the use of persistent identifiers, such as DOI or URN, to be able to always locate the desired dataset (Paskin, 2004).

Another prerequisite for data publication through the internet is that the data are stored in data repositories that guarantee long-term availability. This condition is met by modern data centres, some of which are part of the ICSU system of World Data Centers, which make data accessible through their web portals (Lautenschlager, 2004).

3 THE PROJECT "PUBLICATION AND CITATION OF SCIENTIFIC PRIMARY DATA"

The German CODATA group initiated a project on publication and citation of scientific data (http://www.std-doi.de) which was funded by the German Science Foundation DFG for the period 2003-2005. This project uses persistent identifiers (both DOI and URN) to identify electronic datasets. The identifier is then resolved to the valid location (URL) where the this dataset can be found, thus meeting one of the prerequisites for citeability of online scientific data. In addition, the data publications are included into the catalogue of the German National Library of Science and Technology (TIB).

In this project TIB acts as a registration agency for persistent identifiers. For every data publication it requests a set of metadata to be incorporated into the library catalogue. The data sources are the participating data centres WDC-MARE (Bremen/Bremerhaven), WDC Climate (Hamburg) and GFZ Potsdam. The data centres act as registration agents for scientific and technical data DOIs. These data centres are also responsible for quality control in their data domains, at the same time they also act as long-term archives. The project participants thus encompass all functions necessary for the publication of scientific data.

4 INTELLECTUAL PROPERTY RIGHTS

The Open Access Initiative defines the following criteria for open access:

- Irrevocable free access, worldwide,
- The licence to copy, use, distribute, transmit and display the work publicly,
- The licence to make and distribute derivative works if proper attribution of authorship is given,
- Availability through at least one online repository with long-term archiving capability.

The criteria of accessibility and long-term persistence are met by modern data centres with online access and by the use of persistent identifiers for digital data objects. In addition, publication of scientific data also requires that the intellectual property rights of the data author are guarded by an adequate licence model that allows open access to the data within the boundaries of 'fair use', including the right to produce derivative works.

'Fair Use' is an issue in the 'Berlin Declaration' and was discussed at the 2. Berlin Declaration Conference in May 2004 at CERN, Geneva. Here, Schlögl and Velden (2004) recommended the Creative Commons Licence System (http://www.creativecommons.org) in their Roadmap Proposal as an appropriate licence system for publications in the sciences and humanities. The Creative Commons Licence System is a toolbox to assemble a licence tailored to the requirements of the author. Of the available options the most appropriate for scientific and technical data are:

- By attribution (proper attribution of authorship must be given)
- Non-commercial (the work may not be used commercially)
- Share alike (derivative works must be published under the same licence type)

Publications in science are traditionally credited by attribution, commonly called "citation". It would be considered bad scientific practice if someone produced a derivative work without proper attribution of authorship. In this sense, data publication should be treated analogous to a 'traditional' publication.

A "Science Commons" licence was launched in December 2002 to supplement the Creative Commons Licence. It is still work in progress. Matters are complicated by questions arising from industry involvement, especially in biomedical research and software development. A Science Commons Licence needs to be compatible with existing licence systems such as BSD, GNU, etc., and with patent law. The matter is further complicated by open questions on the issue of liability in case of incorrect data. The Science Commons proposal is currently hosted at Stanford Law School and is backed by 31 associates, among them Rice University, Harvard Law School, MIT, O'Reilly Publishers, and the Public Library of Science.

5 CONCLUSIONS

Applying the 'Berlin Declaration' to data requires a publication system for data beyond 'traditional' media. The criteria of accessibility, persistent identification and long-term availability need to be met to comply with the declaration. The project 'Publication and citation of scientific primary data' showed prototypically how these criteria can be met and by the end of 2005 a publication system for scientific data will be available to the scientific community through the project participants.

This publication system needs to be supplemented by an adequate licence model that allows scientists to use the published data, create new works derived from these, and in turn publish their new data, while respecting the intellectual property rights of the original author and the principles of 'fair use'. The options available in the Creative Commons Licence System suit many fields of scientific research. A separate Science Commons Licence System is desirable and necessary, especially in applied research, but it is still work in progress due to open questions regarding liability issues and compatibility to existing licensing systems.

6 REFERENCES

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. & Wouters, P. (2004) Promoting Access to Public Research Data for Scientific, Economic, and Social Development. *Data Science Journal 3***,** 135-152.

Berlin Declaration on Access to Knowledge in the Sciences and Humanities (2003), Retrieved December 22, 2004 from the World Wide Web: http://www.zim.mpg.de/openaccess-berlin/.

Dittert, N., Diepenbroek, M. & Grobe, H. (2001) Scientific data must be made available to all. *Nature 414***,** 393.

Lautenschlager, M. (2004) WDC Network for Earth System Research, *19th International CODATA Conference*, Berlin, Germany.

Paskin, N. (2004) Digital Object Identifiers for scientific data sets. *19th International CODATA Conference*, Berlin, Germany.

Schlögl, R. & Velden, T. (2004) Berlin 2 Open Access - Roadmap Proposal. *Berlin 2 Open Access: Steps Toward Implementation of the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, CERN, Geneva, Switzerland.