German Data Forum
(RatSWD)

www.germandataforum.de

RatSWD
Working Paper Series

Working Paper No. 149

DataCite - A global registration agency for research data

Jan Brase

July 2010

Federal Ministry
of Education
and Research

Working Paper Series of the German Data Forum (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

The RatSWD Working Paper Series is edited by:

Chair of the RatSWD (2007/2008 Heike Solga; since 2009 Gert G. Wagner)

Managing Director of the RatSWD (Denis Huschka)

Contact: German Data Forum (RatSWD) | Mohrenstraße 58 | 10117 Berlin | office@ratswd.de

DataCite - A global registration agency for research data

Jan Brase
DataCite
German National Library of Science and Technology (TIB)
Hannover, Germany
jan.brase@tib.uni-hannover.de

Abstract— **Since 2005, the German National Library of Science and Technology (TIB) has offered a successful Digital Object Identifier (DOI) registration service for persistent identification of research data.**

In 2009, TIB, the British Library, the Library of the ETH Zurich, the French Institute for Scientific and Technical Information (INIST), the Technical Information Center of Denmark, Canada Institute for Scientific and Technical Information (CISTI) the Australian National Data Service (ANDS) and the Dutch TU Delft Library all signed a Memorandum of Understanding to improve access to research data on the internet.

The goal of this cooperation is to establish a not-for-profit agency called DataCite that enables organisations to register research datasets and assign persistent identifiers to them, so that research datasets can be handled as independent, citable, unique scientific objects.

Keywords- Persistent Identifier; Research Data

I. Background

Knowledge, as published through scientific literature, often is the last step in a process originating from research data. These data are analysed, synthesised, interpreted, and the outcome of this process is generally published in its result as a scientific article.

Only a very small proportion of the original data are published in conventional scientific journals. Existing policies on data archiving notwithstanding, in today's practice data are primarily stored in private files, not in secure institutional repositories, and effectively are lost [1]. This lack of access to scientific data is an obstacle to international research. It causes unnecessary duplication of research efforts, and the verification of results becomes difficult, if not impossible [2]. Large amounts of research funds are spent every year to re-create already existing data [3].

Progress in sharing of scientific data has been made at a fast pace. Infrastructures such as grid exist for storage. Methodologies have been established by data curation specialists to build high quality collections of datasets. These include standards for metadata (provenance, copyright, author of a dataset), registration, cataloguing, archiving and preservation. A large number of disciplines benefit from these methodologies and high quality datasets.

A. Issues

When published, datasets often do not follow the same process as articles. While articles are duly incorporated in digital libraries and can be referenced – in a persistent manner – in other articles, datasets are not published, or published only on the researcher's web site and, if referenced at all, only referenced by the corresponding URL. Such publication model raises a number of issues (see Figure 1):

- Poor preservation properties (e.g. if the researcher moves to another institution, the link may become invalid);
- Poor quality of the documentation;
- Limited impact and academic recognition (dataset cannot be searched or found except from article reference or web search);
- Lack of data quality assessment.

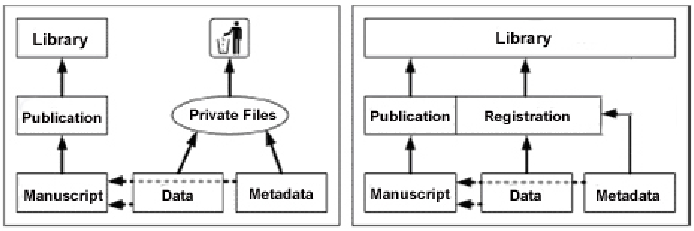

Figure 1. The traditional publication method for datasets on the left, a possible new structure on the right

II. Dataset registration

Dataset identification is a key element for allowing citation and long term integration of datasets into text as well as supporting a variety of data management activities. Also, to foster a culture of data integration, scientists need to be convinced that preparing their data for online publication is a worthwhile effort. It would be an incentive to the author if a data publication had the rank of a citeable publication, adding to his reputation and ranking among his peers. To achieve the rank of a publication, a data publication needs to meet the two main criteria, persistence and quality. Whereas the latter is a very difficult concept that should be made part of the workflow of data integration in the data producers, data persistency is a rather simple problem.

Simply making data available on the 'web' is not sufficient. The location of internet resources, and thus their URL, may easily change, which in most cases means to the user that data are lost [4]. This happens, for instance, if the data are deposited by a researcher in his personal page and the researcher moves from one institution to another. Additionally, this method of data publication makes very little impact since the way by which the dataset may be discovered by another researcher is either:

- Through a web search: Although scientific publications can easily be found through a web search, using the title as a stabile metadata element, the lack of well-defined titles and other metadata makes web-search for datasets difficult. The probability of a page containing the dataset to be found will mainly depend on the quality of the description that surrounds it on the page.
- Through the information in an article: Sometimes the information in an article enables readers to actually identify the location of a dataset, or at least provide contact information of the researcher who collected the data.

Both methods of accessing the dataset have clear limitations in terms of the potential impact of a dataset. It is not surprising that researchers naturally tend to focus their efforts on article publication instead of dataset publication.

For encouraging dataset publication, both the identification of dataset and the awareness of researcher of the availability of this dataset have to be dramatically improved.

A. The Digital Object Identifier (DOI)

DOI: The Digital Object Identifier DOI was introduced in 1998 with the funding of the International DOI Foundation (IDF). It is a registered trademark and DOI names can only be assigned by official DOI registration agencies that are a member of IDF. There are a total of currently 8 Registration agencies worldwide. The DOI system is technically based on the non-commercial Handle system of the Corporation for National Research Initiatives (CNRI). Since 2006, there is an ISO working group (ISO WG 26324) involved in the standardisation of the DOI system.

Registration agencies are responsible for assigning identifiers. They each have their own commercial or non-commercial business model for supporting the associated costs. The DOI system itself is maintained and advanced by the IDF, itself controlled by its registration agency members. Using the Handle system, there is a central free worldwide resolving mechanism for DOI names. DOI names from any registration agency can be by default resolved worldwide in every handle server; DOI therefore are self-sufficient and their resolution does not depend on a single resolution server. A standard metadata kernel is defined for every DOI name. Assigning DOI names involves the payment of a license fee by the Registration agency but their resolution is free.

DOI has emerged as the most widely used standard for digital resources in the publication world. It is currently used by all major scientific publishers and societies (Elsevier, IEEE, ACM, Springer, Wolters Kluwer International Health & Science, New England Journal of Medicine, etc.). The registration for the publishing sector is centrally run by the independent DOI Registration agency CrossRef, which assigns DOI names for 2609 members in the publishing sector. It is also used by the European Commission through its publication agency the Office of Publications of the European Community (OPOCE).

B. Citability through DOI names

While the interoperable and long-term preservation of linkage in scientific publication has been largely achieved through DOI over the last 5 years, dataset publication has not reached a similar maturity level. As mentioned in the last sections, the issue of access to datasets has grown more and more important in the different European research areas, none of these approaches however has yet established a workflow or a functional infrastructure for data registration.

A promising approach to establish dataset citation using DOI names has been started by the Organisation for Economic Co-operation and Development (OECD) for their own datasets. All statistical datasets published by the OECD in their annual factbook can be cited using DOI names [5].

In the academic sector, an established approach within Germany that is actively used by scientists is the Data Registration agency for scientific data at the German National Library of Science and Technology (TIB). TIB is the German National Library for all areas of engineering as well as architecture, chemistry, information technology, mathematics and physics, its holdings comprise around 7.3 million volumes of books, microforms and CD-ROMs, as well as around 18,000 subscriptions to general periodicals and specialist journals. TIB ranks as one of the world's largest specialist libraries, and one of the most efficient document suppliers in its subject areas.

In cooperation with several World Data Centers, over 650,000 datasets have been registered with DOI names as persistent identifiers by TIB. A selection of more than 1,500 datasets that are a part of scientific publications are furthermore directly accessible through the online catalogue of TIB and the German Common Library Network (GBV) [6].

As a major advantage the usage of the DOI system for registration permits the scientists and the publishers to use the same syntax and technical infrastructure for the referencing of datasets that are already established for the referencing of articles. For example:

The dataset:

Lambert, F. et al; (2008): Dust record from the EPICA Dome C ice core, Antarctica, covering 0 to 800 kyr BP, doi:10.1594/PANGAEA.695995

is used and cited in the article:

Lambert, F. et al; (2008): Dust-climate couplings over the past 800,000 years from the EPICA Dome C ice core, Nature, 452, 616-619, doi:10.1038/nature06763

The citation of the dataset and of the underlying article follows the same standards and is therefore easy to adapt by scientists [7].

C. *The Model of Data Registration at TIB*

Since 2005, TIB has been an official DOI Registration Agency with a focus on the registration of research data. The role of TIB is that of the actual DOI registration and the storage of the relevant metadata of the dataset. The research data themselves are not stored at TIB. The registration always takes place in cooperation with data centers or other trustworthy institutions that are responsible for quality assurance, storage and accessibility of the research data and the creation of metadata. Figure 2 illustrates this structure in more detail.

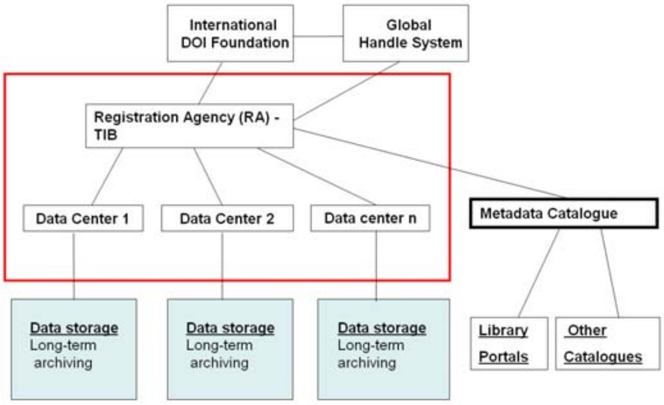
International DOI Foundation
Global Handle System
Registration Agency (RA) - TIB
Data Center 1
Data Center 2
Data center n
Metadata Catalogue
Data storage Long-term archiving
Data storage Long-term archiving
Data storage Long-term archiving
Library Portals
Other Catalogues

Figure 2. The overall structure of TIB's DOI Registration Agency

III. DATACITE

Access to research data is nowadays defined as part of the national responsibilities. During the last years most national science organisations have addressed the need to increase the awareness of and the accessibility to research data.

Science itself nevertheless is international, scientists are involved in global unions and projects, they share their scientific information with colleagues all over the world, they use national information providers as well as foreign ones.

When facing the challenge of increasing access to research data, a possible approach should be a global cooperation for data access with national representatives.

- a global cooperation, because scientist work globally, scientific data are created and accessed globally.

- with national representatives, because most scientists are embedded in their national funding structures and research organisations .

TIB, the British Library, the Library of the ETH Zurich, the French Institute for Scientific and Technical Information (INIST), the Technical Information Center of Denmark, Canada Institute for Scientific and Technical Information (CISTI) and the Dutch TU Delft Library all signed a Memorandum of Understanding to this effect during the meeting of the International Council for Scientific and Technical Information (ICSTI) in Paris on 2 March 2009.

The goal of this cooperation is to establish a not-for-profit agency that enables organisations to register research datasets and assign persistent identifiers to them, so that research datasets can be handled as independent, citable, unique scientific objects.

The key point of this approach is the establishment of a Global DOI Registration agency for scientific content called DataCite that will offer to all researchers dataset registration and cataloguing services. DataCite shall be carried by non-commercial information institutions and libraries instead of publishers. This approach will allow easy access to the DOI system for non-commercial information institutes and libraries worldwide.

The objective of establishing an independent global DOI RA is to pool together resources of various interested local agencies. The benefits will be the following:

- Reduced infrastructure cost
- Better integration of the national infrastructures
- Reference implementation of the service in a distributed fashion
- Advanced distributed search capabilities for improving researchers' awareness of available datasets

Practical DataCite can be implemented by widening the DOI model of TIB to a model of local agencies. This approach follows the example of the publishing industry in which the (often competing) publishers together use the central infrastructure of CrossRef to assign their DOI names.

Following TIB's model, data curation, maintenance and storage are not in the responsibility of the joint agency. Through its local partners it will furthermore offer services to existing national and international repositories and initiatives and therefore closing the gap between data infrastructure and information providers.

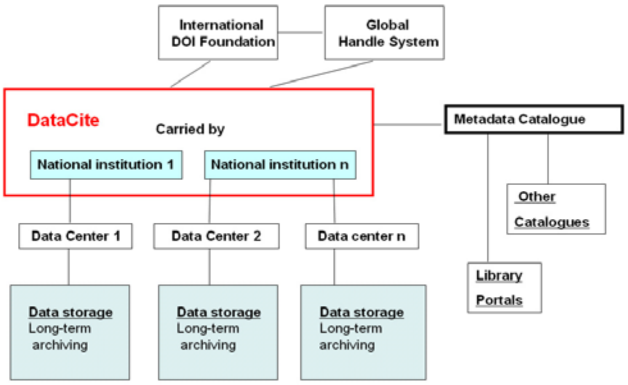

Figure 3. *The overall structure of DataCite*

A. *Structure*

The structure of DataCite will be the following:

One central office will be located at TIB as the central address and responsible body for the International DOI Foundation (IDF), with a managing agent and technical staff. Each DataCite member will host its own office of the RA, allowing him to directly contact any data center in his domain. The partners are allowed to build up their own technical infrastructure for DOI registration or use the central infrastructure at TIB. The members elect an advisory board. Affiliated members, e.g. members who are interested in establishing standards and exchanging expertise in accessing research data, but who are not in interested in assigning DOI names, are welcome and will advise the exectutive board.

There will be one central metadata repository containing the descriptions of all registered data sets, with standardised interfaces to the partners own repositories and applications.

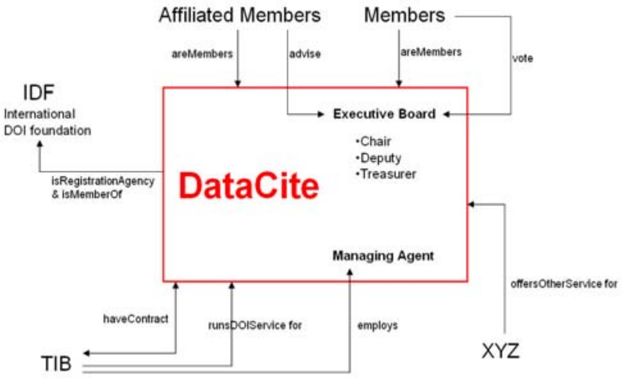

Figure 4. *The structure of DataCite in more detail*

The metadata and workflow definitions will be standardised through all partners. Every partner will have the right to develop its own business models for re-financing the registration costs.

DataCite will officially start at January 1st, 2010 and will always remain open for other institutions to join under the same rules and obligations.

For more details on the foundation of DataCite, we refer to [8].

B. *Partners*

The fist institutions to establish DataCite together with TIB are (in alphabetical order):

- **Australian National Data Service (ANDS):** ANDS is funded by the Australian Commonwealth Government's Department of Innovation, Industry, Science and Research (DIISR). It aims to influence national policy in the area of data management in the Australian research community, inform best practice for the curation of data and transform the disparate collections of research data around Australia into a cohesive collection of research resources

- **British Library (BL)**: The British Library (BL) is the national library of the United Kingdom. It is one of the world's largest research libraries, holding over 150 million items in all known languages and formats; As a legal deposit library, the BL receives copies of all books produced in the United Kingdom and the Republic of Ireland, including all foreign books distributed in the UK.

- **Canada Institute for Scientific and Technical Information (CISTI);** National Research Council Canada Institute for Scientific and Technical Information (NRC-CISTI) is Canada's national science library and leading publisher of scientific information. It provides Canada's research and innovation community with tools and services for accelerated discovery, innovation and commercialization.

- **ETH Zurich Library**, Switzerland: The ETH-Bibliothek is the largest library in Switzerland and the main library of the Swiss Federal Institute of Technology. In addition, it functions as the Swiss center for information on science and technology. The Library holds more than 6.9 million items, including maps, old prints, audiovisual materials, journals, databases and much more.

- **Institute for Scientific and Technical Information (INIST-CNRS), France**: INIST is a unit of the French National Center for Scientific Research (CNRS) under the administrative authority of the French Ministry in charge of scientific research. Its mission is to facilitate access to findings of all fields of worldwide scientific research. INIST-CNRS relies on one of the most important collections of scientific documents in Europe to provide a whole range of information services and Information portals providing access to electronic resources and dedicated to specific scientific communities.

- **National Technical Information Center Denmark**: The Technical Information Center of Denmark is DTU’s center for scientific information provision, information management and information competences as well as the Danish national technical information center. The Technical Information Center of Denmark acts as a modern university library and as a center for management of the university’s own research information. The information of the center is primarily disseminated and handled in a digital form and secondarily on the basis of printed collections. The public premises of the center are first and foremost designed to support the information searching and learning of the student.

- **TU Delft Library, The Netherlands**: TU Delft Library is the biggest technical-scientific library in the Netherlands. Its task is to safeguard the provision of technical-scientific information in the Netherlands. It focuses as much as possible on digital service in the field of technical science information. The TU Delft Library is the hub of knowledge for technical and scientific information in the Netherlands. It supports research and education within TU Delft and at the national level. The 3TU.Datacentre is an initiative of the libraries of TU Delft, TU Eindhoven and the University of Twente under the auspices of the 3TU.Federation. The 3TU.Datacentre will provide storage of and continuing access to technical-science study data.

ACKNOWLEDGMENT

The foundation of DataCite is the work of many. Special acknowledgement goes to:

Adrian Burton from ANDS, Adam Farquhar and Elisabeth Newbold from BL, Angela Gastl, Arlette Piguet and Wolfram Neubauer from ETH Zurich, Herbert Gruttemeier from INIST, Maria Heijne and Jeroen Rombouts from TU Delft Library, Alfred Heller and Mogens Sandfaer from Technical Information Center of Denmark, Francois Dube from CISTI and Irina Sens and Uwe Rosemann from TIB.

REFERENCES

[1] Lawrence, S et al (2001) Persistence of Web References in Scientific Research. IEEE Computer 34 (2), 26-31. http://www.fravia.com/library/persistence-computer01.pdf

[2] Dittert, N., Diepenbroek, M. & Grobe, H. (2001) Scientific data must be made available to all. Nature 414 (6862), 393. doi:10.1038/35106716.

[3] 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.

[4] Koehler, W. (2004) *A longitudinal study of Web pages continued: a report after six years*. Information Research 9 (2). http://informationr.net/ir/9-2/paper174.html

[5] Green, T (2009), *We Need Publishing Standards for Datasets and Data Tables*, OECD Publishing White Paper, OECD Publishing. doi: 10.1787/603233448430

[6] Brase, J. (2004) *Using Digital Library Techniques - Registration of Scientific Primary Data*. Lecture Notes in Computer Science 3232, 488-494.

[7] Altman M., King G., *A Proposed Standard for the Scholarly Citation of Quantitative Data*, D-lib Magazine, March/April 2007, Vol 13 No.3/4

[8] Brase, J., Farquhar, A., Gastl, A., Gruttemeier, H., Heijne, M.., Heller, A.et al.(2009). *An approach for a joint global registration agency for research data* Information Services & Use" 29 (2009) 13–27 ISSN 0167-5265, doi: 10.3233/ISU-2009-0595