PRINTER-FRIENDLY FORMAT Return to Article

D-Lib Magazine

January/February 2015
Volume 21, Number 1/2

The Tenth Anniversary of Assigning DOI Names to Scientific Data and a Five Year History of DataCite

Jan Brase and Irina Sens
German National Library of Science and Technology, Germany
{jan.brase, irina.sens}@tib.uni-hannover.de

Michael Lautenschlager
German Climate Computing Centre, Germany
lautenschlager@dkrz.de

DOI: 10.1045/january2015-brase

Abstract

As part of a project initiated by the German Research Foundation (DFG), the German National Library of Science and Technology (TIB) assigned its first DOI names to scientific data in summer 2004. The goal was to use persistent identifiers as part of a broader effort to make scientific datasets citable research outputs. The effort begun by TIB led to the creation and funding of DataCite on 1 December 2009. During the past five years DataCite has grown into a global consortium that has assigned over four million DOI names to scientific datasets and other research artefacts. It is a successful cooperative effort led by scientists, librarians and researchers. This article highlights its development and gives an overview of DataCite's recent work.

Keywords: Research Data Citation, DOI Names, Persistent Identifier, DataCite

1 The Concept

In summer 2004 the German National Library of Science and Technology (TIB) assigned its first DOI names to scientific data in order to make scientific datasets citable research outputs. This took place as part of a project initiated by the German Research Foundation (DFG). Five years later the work of TIB led to the funding of DataCite as a global consortium focussed on the assignment of citation references and DOI names to research data and other scientific outputs. DataCite has become a global consortium with 30 members from all over the world that has assigned over four million DOI names to scientific datasets and other research artefacts. DataCite is an example of a successful cooperative effort that evolved early ideas based on scientific workflow into an international organisation.

The DataCite development process started around 2000. An analysis of the first concept of integration of research data in scientific workflows was characterised in [LAU2002] together with a concept to improve data availability in literature by citing scientific primary data. Most of the analysis is still valid today because the envisaged cultural changes in scientific workflows have been slow in coming.

In principle, scientists are prepared to provide data, but the necessary extra work that they must perform in processing, context documentation and quality assurance for that data is often neither appreciated nor acknowledged. The classic mode of distributing scientific results is their publication in professional journals. These articles are recorded in the "citation index". The index is used for a performance evaluation of scientists. Scientific data publications have been taken into account infrequently until now.

Project data are widely spread among research institutes and are collected and governed by scientists. Due to

inadequate performance of the extra work involved, project data are often poorly documented, barely accessible and not maintainable over long time periods. Large amounts of data are unused as they are only known and accessible to a small group of scientists.

Discussion about verification of scientific results resulted in changes to the existing rules of good scientific practice in scientific research organizations like DFG (Deutsche Forschungsgemeinschaft), HGF (Helmholtz Gemeinschaft Deutscher Forschungszentren) and MPG (Max-Planck-Gesellschaft). The rules also include guidelines for data access. Primary data of a publication nowadays have to be stored and made accessible for at least ten years to allow a verification of results.

One has to differentiate between ways to enforce good scientific practice resulting from official regulations on one hand, and personal motivation of the individual scientists on the other. Until now enforced regulations in the field of science did not prove to be helpful when attempting to fill data systems continuously. Preferable, then, is encouraging personal motivation to publish primary data and thus make them accessible for a long period. Publications of primary data should be citable as publications, so that the data set may be cited together with the author when being used further. By this, scientific primary data is not exclusively understood to be part of a scientific publication, but may have its own identity.

In existing scientific journals, there is no room for repeating data work, i.e. systematical use of existing methods to complete a data basis and by this make it usable for later scientific applications. Repeating data work is not original scientific effort, but is necessary to support science and should be acknowledged as well.

The general idea of treating scientific data as independently published digital objects arose around the end of the last millennium. The developments in publishing, the Internet in general, and the use of the Digital Object Identifier (DOI®) to link article citations, motivated the discussion on adapting the methods of managing electronic publications to research data, or more generally, to digital entities in common [MUN1998].

During the 17th CODATA (Committee on Data for Science and Technology) Conference in October 2000 in Baveno, Italy, two members of the official German CODATA delegation that still existed at that time, M. Lautenschlager from the German Climate Computing Centre (DKRZ) and J. Wächter from the German Research Centre for Geosciences (GFZ), discussed possibilities for citing research data in scientific literature. The general idea behind the concept of the adaptation of publication of electronic media to research data is to give credit to data authors and to improve verification of scientific results. The usage of persistent identifiers like DOI names allows transparent data access independently from their actual storage location. The resulting discussion paper [LAU2000] was debated with the German Research Foundation (DFG) and at the end of 2000 the DFG agreed to fund a national working group to develop an implementation strategy for publication and citation of research data.

Early in 2001 this working group investigating the "Possibility of Citing Scientific Primary Data" was established with a composition of 50% librarians and 50% scientists.[1] In March 2002 the group presented a report, "Conception of citing scientific primary data" [LAU2002] to the DFG. Central points of this conception were to assure data quality and to store primary data for the long-term together with their citation reference.

The guiding principle was to follow the methodology for the publication of scientific literature as closely as possible, which means that scientific data are irrevocable after publication and that they have a suitable granularity for integration of data citation references into reference lists of scientific articles. An additional requirement was that the research data publication concept has to be applicable to multiple, ideally all, scientific disciplines.

The analogy between research data publication and publication in scientific literature led to the selection of the DOI as persistent identifier and cooperation with the International DOI Foundation (IDF) for the definition of a suitable DOI metadata profile for research data. Another reason was that most scientists were already familiar with the use of DOI names through their article publication. URN as an alternative were used for theses in libraries but were not well-known in the scientific world. The organisational structure followed the IDF structure shown in Figure 1, but with libraries as DOI registration agencies (here the German National Library of Science and Technology (TIB)) and the scientific long-term data archives as DOI registrants (here the German ICSU World Data Centers).

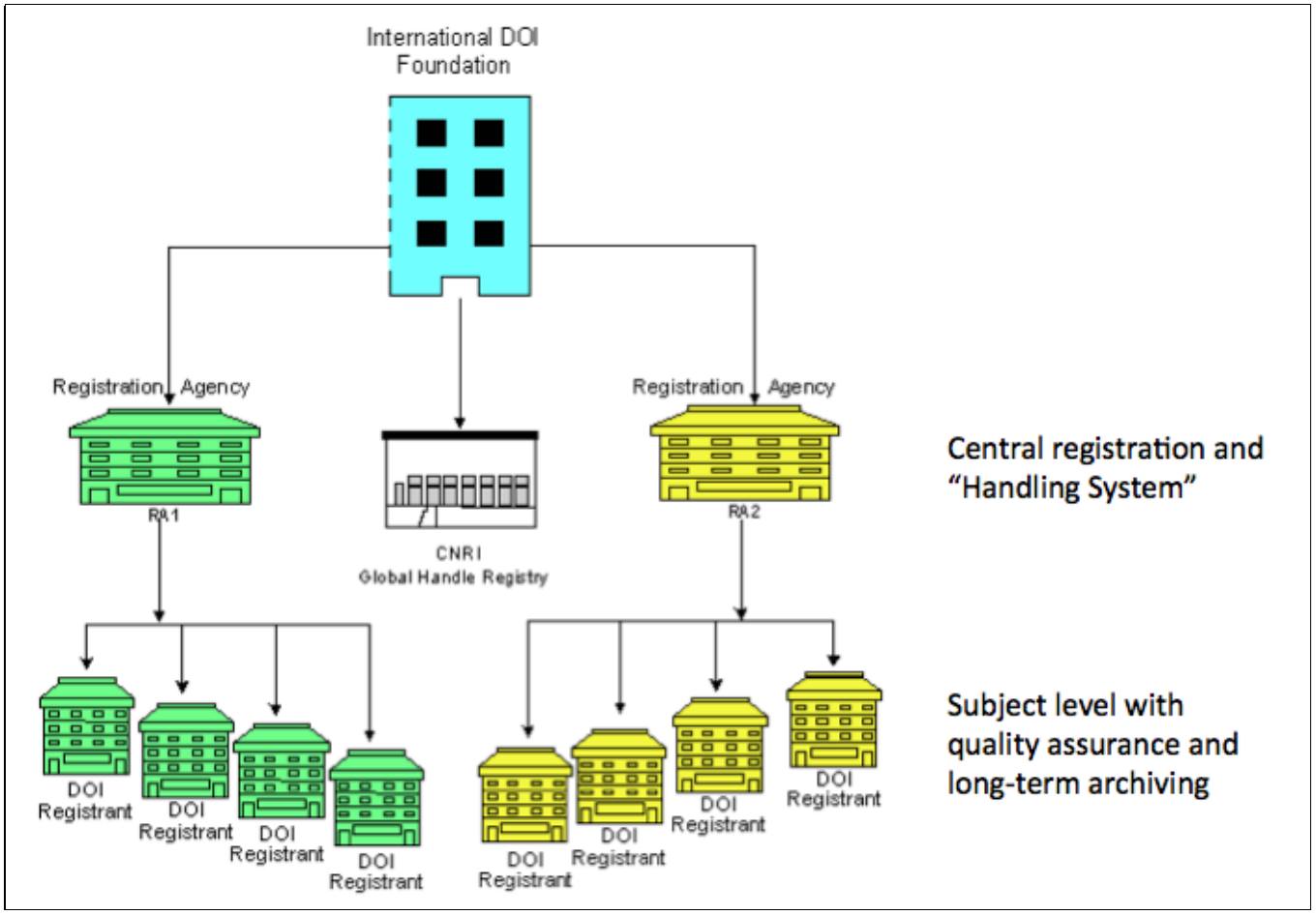


Figure 1: Organisation of the International DOI Foundation (IDF) (Source: DOI Handbook, 2002)

These concepts were addressed in the subsequent implementation projects, and the adapted IDF organisational structure (Figure 1) has finally been transferred to DataCite and is still reflected in the DataCite web pages along with basic principles from the conception phase.

2 The Project

As a result of the working group's presentation and the discussion of the concept of citing scientific primary data, the DFG funded the implementation of the concept in two phases; a pilot implementation (2003 — 2005) and development and improvement as a service (2006 — 2009). DFG set a number of constraints. Applicability of the scientific data publication should be interdisciplinary and international. One specific recommendation had been made with respect to the identifier system: URN (Uniform Resource Names) as a purely non-commercial system had to be used in parallel to the suggested DOI system (Digital Object Identifier).

The pilot implementation began in 2003, initially focused on geoscience, and was funded for two years. Partners in this project were three German data archives (German Climate Computing Centre, German Research Centre for Geosciences and the Helmholtz Centre for Polar and Marine Research) and the German National Library of Science and Technology (TIB) that would act as a registration agency for research data. The goal of the project was to ensure that quality controlled scientific data are classified as irrevocable and are registered, together with DOI name and citation reference, in a CNRI handle service and in library catalogues, in order to:

- allow for transparent data access;
- foster verification of scientific results;
- allow for data citation in scientific literature; and
- give credit to data authors.

The group defined a standard set of metadata elements for DOI registration that are based on the ISO 690-2 and the

Dublin Core metadata to describe the required bibliographic metadata including the citation reference. The standard metadata set was registered as STD-DOI (Scientific and Technical Data DOI) metadata profile with the IDF. In parallel TIB joined IDF as a full member to act as an official DOI Registration Agency. In the definition of the metadata profile for publication of digital research data, emphasis was given to cross-disciplinary applicability.

In 2004 TIB became a DOI Registration Agency and the pilot for STD-DOI research data publication started operation. The implementation was introduced in [BRA2004, BRA2005]. The first registered DOI, 10.1594/WDCC /EH4_OPYC_SRES_A2, was for a climate model experiment. STD-DOI registrations from the project partners appeared later that year in the system. By the end of 2004 the number of research data entities which finished the STD-DOI data publication process, and therefore gained a DOI registration and were registered in TIB's library catalogue, increased to 30 in total from different geoscience research fields. These 30 data sets were the first research data sets to ever appear as individual objects in a library catalogue worldwide.

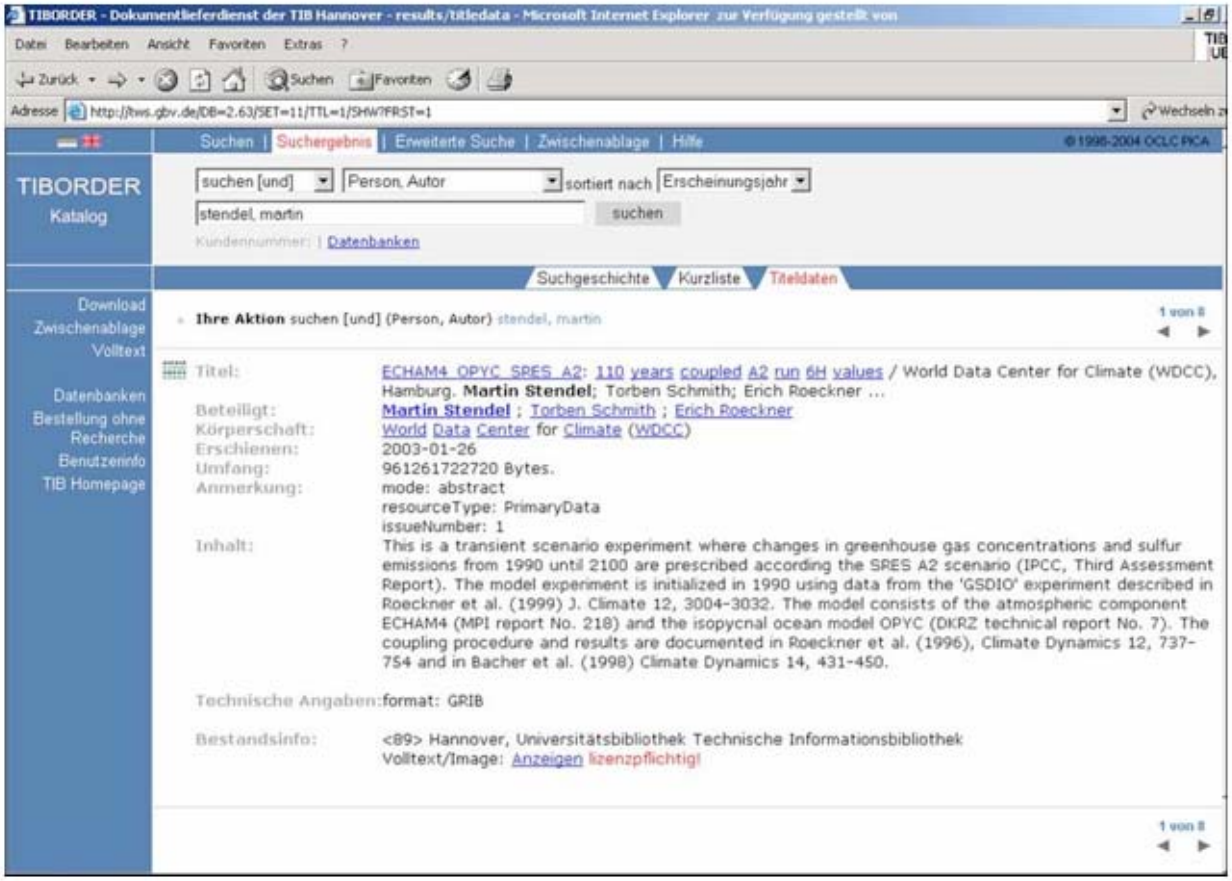


Figure 2: The first registered dataset as a query result at TIB in 2004.
The related DOI (10.1594/WDCC/EH4_OPYC_SRES_A2) is still resolvable and referenced data can be accessed from the WDCC.

In 2005 the TIB Advisory Board recommended integrating the operation of a non-commercial registration agency for scientific-technical primary data into TIB's standard services. This was a major step in the direction of sustainability. The service was quite successful in Germany and started to grow, enough that in 2006 the DFG funded a follow-up project for two and a half years. The focus of that project was to:

- introduce research data publication as operational service in Germany;
- expand from geosciences to more scientific disciplines;
- integrate international partners,; and
- define sustainability and a business model.

A data publication process for independent research data entities had been established following the analogy of the publication process in scientific journals as closely as possible. The role of the publisher has been assigned to the responsible data archive as DOI publisher or publication agent. The publication requirements are defined in the

STD-DOI data publication process and include irrevocable, fully documented data entities for unrestricted (scientific) use, a citation reference and final review of data and metadata by the data author. A peer review process for scientific data could not be established during the course of the project because of its complexity. Quality assurance is clearly a discipline dependent topic and discussion of objective criteria for data quality is an open issue in most of the scientific disciplines. At this point the analogy between publication of scientific journal articles and the publication of research data entities is broken. For the data publication process the quality flag "approved by data author" has been chosen. The STD-DOI data publication finishes with the author's assurance: Yes, these data version is final, will never been changed and is open for public use. Figure 3 shows the main steps in the publication process for independent scientific data entities.

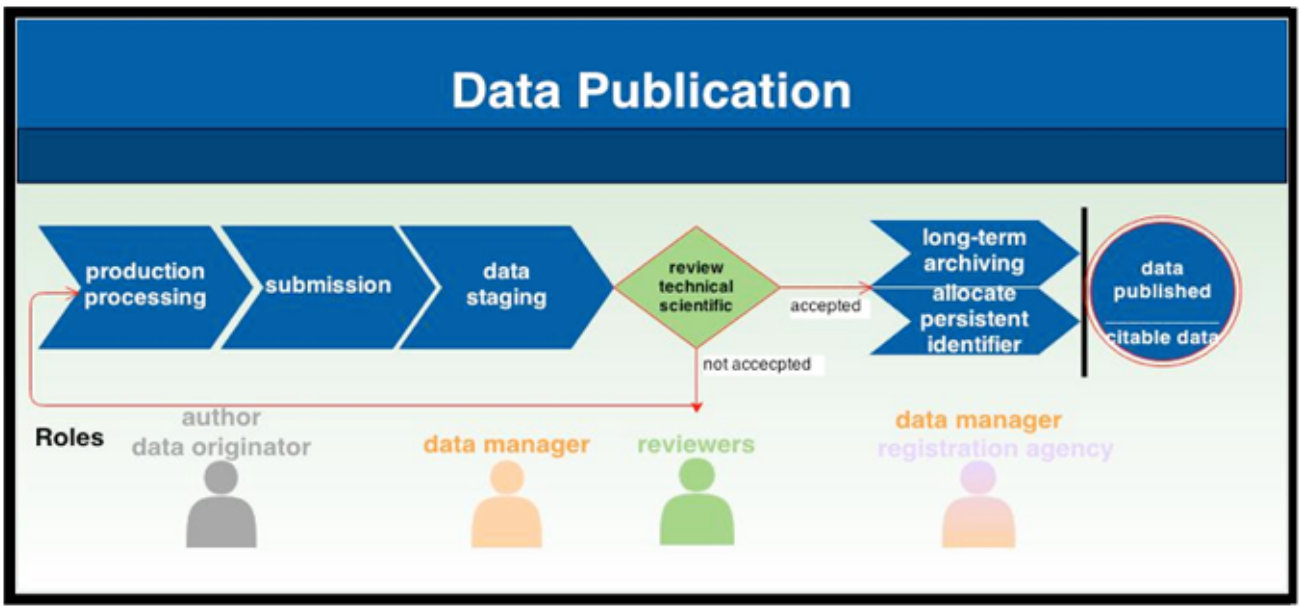


Figure 3: Data publication process for independent scientific data entities.

During the end of this project phase an international organisational structure came under consideration (Figure 4). The growing international interest required a more general structure for a scientific data publication infrastructure. The German organisational model, with TIB as publications agency and with long-term data archives which were under contract with TIB and operated as publication agents, needed to be transferable to other countries.

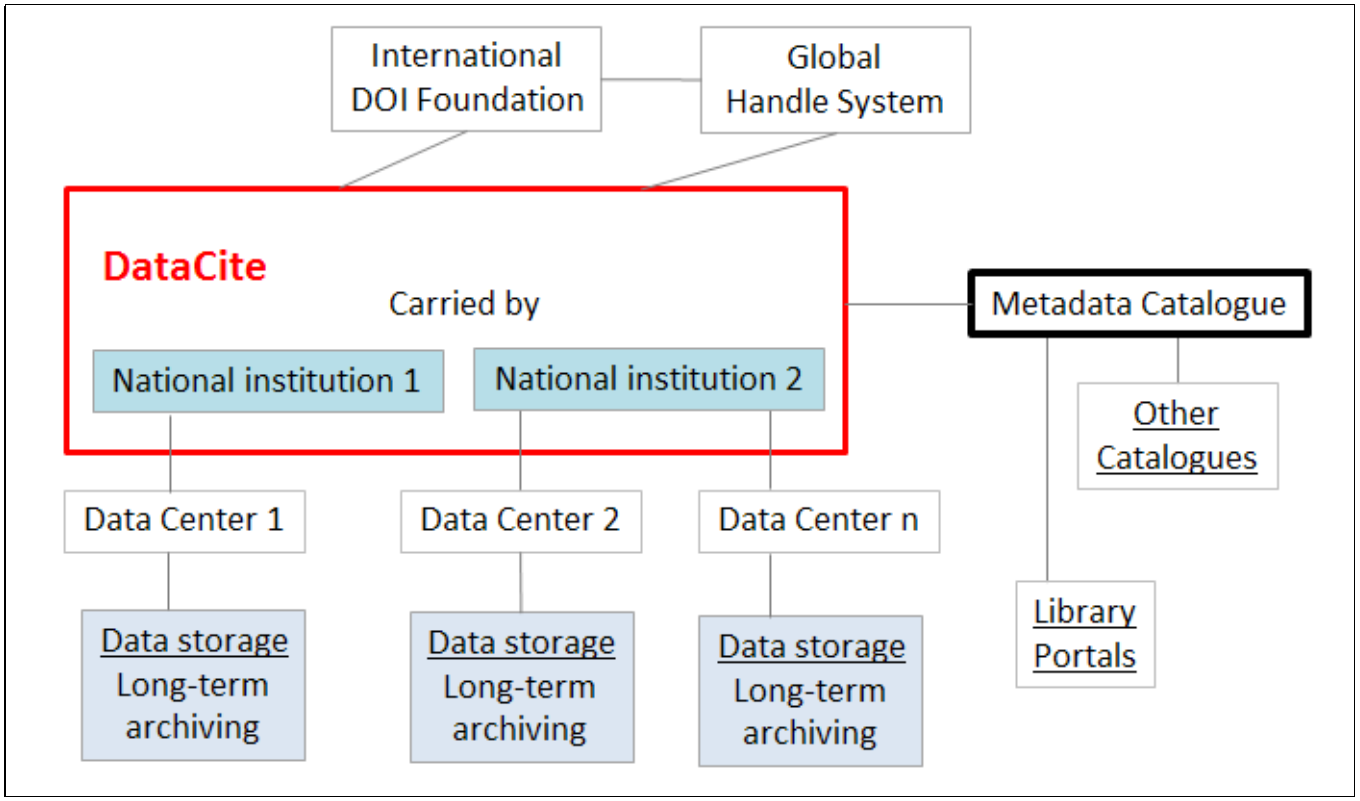


Figure 4: International organisational structure of publication agencies and publication agents which finally led to DataCite.

The service gained 17 new partners from different disciplines, including data centres from outside of Germany. The international interest finally led to the founding of DataCite [FAR2009] and the transition of STD-DOI profiles to the DataCite data model and infrastructure. In 2012 the STD-DOI handle server at TIB shut down and more than 1 million scientific data DOIs were transferred to the DataCite handle servers and integrated in the DataCite repository.

3 Gathering of Partners

Between 2007 and 2008 the work of the project was presented at various events in Germany and internationally. Further data centers outside of Germany expressed interest in assigning DOI names to their data sets, and technical libraries from other countries became interested in offering DOI registration as part of their service. Beginning with the partners of the consortium TechLib (Tu Delft Library, Library of the ETH Zurich in Switzerland and Technical Information Center of Denmark (DTIC)) TIB started to discuss the possibility of a joint collaboration of Technical Libraries in Europe to act as a DOI Registration Agency for scientific data. But it was not until the British Library and the French Information Center INIST expressed interest in late 2008 to join such an effort that the idea began to really take shape.

In February 2009 at a workshop on data citation jointly organised by the International Council of Scientific and Technical Information (ICSTI) and CODATA in Paris, the six libraries signed a Memorandum of Understanding to "establish a not-for-profit agency that enables organisation to register research datasets and assign persistent identifiers to them." [DLB2009] During the next months contacts were established with the California Digital Library (CDL), Purdue University, Canada Institute for Scientific Information (CISTI) and the Australian National Data Service (ANDS) which expressed interest in contributing to the new agency. The scope was then no longer a European one but rather a global one.

4 DataCite

DataCite was funded by seven members on December 1st 2009 at the British Library in London.

Figure 5: The first board of DataCite on December 1st, 2009: Alfred Heller (DTIC), Pam Bjornson (CISTI), Uwe Rosemann (TIB), President Adam Farquhar (BL), Patricia Cruse (CDL) and Jan Brase (TIB) as the managing agent.

From a legal perspective, DataCite is a German non-for-profit association (Verein), with the DataCite office at TIB in Hannover, Germany. During the next years, DataCite constantly grew in members, customers and DOI registrations. An annual event was introduced as an important gathering of members and stakeholders to propagate data citation. Events have taken place in Hannover (2010), Berkeley (2011), Copenhagen (2012), Washington DC (2013) and Nancy, France (2014).

In 2015, the consortium consists of 30 members in 16 countries. This emphasizes the need for local representatives to work together with a globally organized framework such as DataCite. Four DataCite members are located in Germany: the German National Library of Science and Technology (TIB), the Leibniz Information Centre for Life Sciences (ZB MED), the Leibniz Information Centre for Economics (ZBW) and the Leibniz Institute for the Social Sciences (GESIS). These institutes are working mostly with discipline-specific datacenters within the scope of their scientific area. Other European members include: The Library of the ETH Zürich in Switzerland, the Library of TU Delft in the Netherlands, the L'Institut de l'Information Scientifique et Technique (INIST) in France, the Technical Information Center of Denmark, the British Library, the Swedish National Data Service (SND), the Conferenza dei Rettori delle Università Italiane (CRUI) in Italy, the Library and Information Centre of the Hungarian Academy of Sciences (MTA KIK), and the University of Tartu in Estonia. The representatives of North America are the California Digital Library, the Office of Scientific and Technical Information (OSTI), the Purdue University and the Canada Institute for Scientific and Technical Information (CISTI). DataCite members from Australia and Asia are the Australian National Data Service (ANDS), the National Research Council of Thailand (NRCT), and the Japan Link Center (JaLC) respectively. A DataCite member from Africa is the South African Environmental Observation Network (SAEON) and last but not least, the European Organisation for Nuclear Research (CERN) is an international member of DataCite.

Affiliate members who support DataCite but do not actively assign DOI names are: Beijing Genomics Institute (BGI); Korea Institute of Science and Technology Information (KISTI); Digital Curation Center (DCC); Harvard University Library; Gesellschaft für wissenschaftliche Datenverarbeitung mbH (GWDG): Open Researcher and Contributor ID (ORCID); Institute of Electrical and Electronics Engineers (IEEE); and Inter-university Consortium for Political and Social Research (ICPSR).

5 Services

The number of DOI names registered by DataCite, in cooperation with 350 data centers from all over the world, has increased to over 4 million. In the time between 2012 and 2014, the technical infrastructure for the registration of DOI names, along with several additional services, was established and expanded. The core element of the service infrastructure is the DataCite Metadata Store (MDS). In addition to the MDS DataCite provides a search tool (DataCite Search). Furthermore, DataCite offers a detailed statistic portal where stats and numbers of registered and resolved DOI names are displayed.

In co-operation with CrossRef, a content negotiation service was established. The content negotiation service enables the user to persistently resolve all DOI names directly to their metadata in XML or RDF format. Another implemented tool from this co-operation is the Citation Formatter which provides the citation of DataCite and CrossRef DOI names in various formatting styles.

In 2012 Thomson Reuters started to build up their Data Citation Index. It provides information about research data from various research data repositories. In August 2014 Thomson Reuters and DataCite announced an official collaboration to ensure that all high quality research data from repositories worldwide that work with DataCite will be harvested by the Data Citation Index.

Notes

[1] The group consisted of Dr. Michael Lautenschlager as the Speaker; Dr. Joachim Wächter; Carola Kauhs (Max Planck Institute for Meteorology); Dr. Manfred Reinke (Alfred Wegener Institute); Prof. Dr. Gerhard Schneider (University of Freibug); Dr. Irina Sens (German National Library of Science and Technology); Dr. Uwe Ulbrich (University of Cologne).

References

[BRA2004] J. Brase (2004): *Using digital library techniques — Registration of scientific primary data* in "Research and advanced technology for digital libraries" Springer LNCS 3232, ISBN 3-540-23013-0.

[BRA2005] J. Brase, U. Schindler and M. Diepenbroeck (2005): *Webservice infrastructure for the registration of scientific primary data* in "Research and advanced technology for digital libraries" Springer LNCS 3652, ISBN: 3-540-28767-1

[DLB2009] European Initiative to Facilitate Access to Research Data, *D-Lib Magazine*, Volume 15, No. 5/6 ISSN 1082-9873, http://doi.org/10.1045/may2009-inbrief

[FAR2009] A. Farquhar, J. Brase, A. Gastl, H. Gruttemeier, M. Heijne, A. Heller, A. Piguet, J. Rombouts, M. Sandfaer, I. Sens (2009): *Approach for a joint global registration agency for research data*. Inf. Serv. Use 29 (1):13—27. http://doi.org/10.3233/ISU-2009-0595

[MUN1998] M. Mundt (1998): *Der DOI (digital object identifier) ein verlagsorientiertes Indexierungswerkzeug auch anwendbar auf Datensätze?* Term Paper at Potsdam University 1998. http://doi.org/10.2312/GFZ.misc.370184

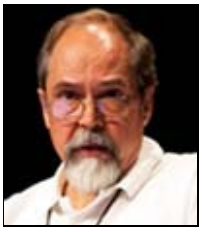
[LAU2000] M. Lautenschlager and J. Wächter (2000): *Publikation und Zertifizierung von wissenschaftlichen Daten*. Tischvorlage CODATA Landesausschusssitzung 29.11.2000.

[LAU2002] M. Lautenschlager, C. Kauhs, M. Reinke, G. Schneider, I. Sens, U. Ulbrich and J. Wächter (2002): *Conception of Citing Scientific Primary Data*. Final Report of CODATA Working Group "Possibility of Citing Scientific Primary Data".

[LAU2003] M. Lautenschlager and I. Sens (2003): Konzept zur Zitierfähigkeit wissenschaftlicher Primärdaten. Information — Wissenschaft & Praxis, No. 54, 463—466.

About the Authors

Jan Brase has a degree in Mathematics, and a PhD in Computer Science. His research background is metadata, ontologies and digital libraries. Since 2005, he has been head of the DOI Registration Agency for research data at the German National Library of Science and Technology (TIB). He was also Managing Agent of DataCite from its founding until December 2014. DataCite was founded in December 2009 and has set itself the goal of making online access to research data for scientists easier by promoting the acceptance of research data as individual, citable scientific objects.



Michael Lautenschlager is leading the Data Management Department at DKRZ (German Climate Computing Centre) and is also director of the ICSU World Data Center for Climate (WDCC) at DKRZ. Dr. Lautenschlager started his career at MPI-M (Max-Planck-Institute for Meteorology) in Earth system modeling. Since 1991 he is active in scientific data management at DKRZ and MPI-M. The use of persistent identifiers has always been a focus of his work. It was WDCC-DKRZ that minted the first DOI for data in 2004.

Irina Sens is currently the Deputy Director of the German National Library of Science and Technology in Hannover and has served in this capacity since 1999, and is also head of the collection development and metadata department. She was a Project Coordinator of several Digital Library projects. She is member of the DataCite Board since 2011. Dr. Sens studied chemistry and mathematics at the University of Marburg, Germany from 1984 — 1990 where she also received her Ph.D. in chemistry in 1993. After completion of her doctoral studies she served a librarian trainee at the State and University Library from 1993 — 1995.

PRINTER-FRIENDLY FORMAT Return to Article