Approach for a joint global registration agency for research data

Brase, Jan; Farquhar, Adam; Gastl, Angela; Gruttenmeier, Herbert; Heijne, Maria; Heller, Alfred; Piquet, Arlette; Rombouts, Jeroen; Sandfær, Mogens; Sens, Irena

Published in:
Information Services & Use

Link to article, DOI:
10.3233/ISU-2009-0595

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

Information Services & Use 29 (2009) 13–27
DOI 10.3233/ISU-2009-0595
IOS Press

Approach for a joint global registration agency for research data

Jan Brase [a,*] Adam Farquhar [b], Angela Gastl [c], Herbert Gruttemeier [d], Maria Heijne [e], Alfred Heller [f], Arlette Piguet [c], Jeroen Rombouts [e], Mogens Sandfaer [f] and Irina Sens [a]

[a] *German National Library of Science and Technology (TIB), Hannover, Germany*
[b] *The British Library, London, UK*
[c] *ETH Library Zurich, Zurich, Switzerland*
[d] *Institute for Scientific and Technical Information (INIST)-CNRS, Vandoeuvre les Nancy, France*
[e] *TU Delft Library, Delft, The Netherlands*
[f] *Technical Information Center of Denmark, Lyngby, Denmark*

Abstract. The scientific and information communities have largely mastered the presentation of, and linkages between, text-based electronic information by assigning persistent identifiers to give scientific literature unique identities and accessibility. Knowledge, as published through scientific literature, is often the last step in a process originating from scientific research data. Today scientists are using simulation, observational, and experimentation techniques that yield massive quantities of research data.

These data are analyzed, synthesized, interpreted, and the outcome of this process is generally published as a scientific article. Access to the original data as the foundation of knowledge has become an important issue throughout the world and different projects have started to find solutions.

Global collaboration and scientific advances could be accelerated through broader access to scientific research data. In other words, data access could be revolutionized through the same technologies used to make textual literature accessible.

The most obvious opportunity to broaden visibility of and access to research data is to integrate its access into the medium where it is most often cited: electronic textual information. Besides this opportunity, it is important, irrespective of where they are cited, for research data to have an internet identity.

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 this white paper we discuss the possibilities to open this registration to a global consortium of information institutes and libraries.

Definition

In this white paper the definition of the term "data" or "research data" follows the definition given by the U.S. National Institutes of Health (NIH) Grants Policy Statement as "recorded information, regardless of the form or medium on which it may be recorded, and includes writings, films, sound recordings, pictorial reproductions, drawings, designs, or other graphic representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing or computer programs (software), statistical records, and other research data".

*Corresponding author: Jan Brase, German National Library of Science and Technology (TIB), Welfengarten 1b, 30167 Hannover, Germany. Tel.: +49 511 762 19869; E-mail: jan.brase@tib.uni-hannover.de.

1. Background

Knowledge, as published through scientific literature, often is the last step in a process originating from research data. These data are analyzed, synthesized, 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 [11]. 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 [5]. Large amounts of research funds are spent every year to re-create already existing data [2].

Data have always been at the heart of scientific progress. They are the raw material out of which research can be carried out and what many publications are based upon. Data integration with text is therefore an important aspect of scientific collaboration. It allows verification of scientific results and joint research activities on various aspects of the same problem. Data integration is instrumental for the successful realization of multidisciplinary research, academia-industry collaboration and the development of new products in large scale engineering projects (e.g. in the aerospace, ship building or automotive industries).

Recognizing the need for data sharing, several scientific communities have organized data collection, archiving and access to serve their community needs. For instance, earth and environmental studies data are collected and shared on a world-wide level through the World Data Center System (http://www.ngdc.noaa.gov/wdc/). Data publication is an essential component of every large scientific instrument project (e.g., the CERN Large Hadron Collider). In fact, the development of grid technology can be linked to infrastructure requirements that were raised by the volume of information that high energy physics experiments generates and by the need to share this information among physicists across the globe. Similar examples can be found in geophysics, chemistry, astronomy, biology, etc.

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, cataloging, archiving and preservation. A large number of disciplines benefit from these methodologies and high quality datasets. Figure 1 illustrates how formal dataset publication effectively transforms data into information and ultimately knowledge.

1.1. Issues

Unfortunately, a large body of data used for research is not published following established best practices.

Problem 1. A large volume of research data is not shared at all. Since academic recognition is mainly achieved through publication, sharing datasets is a time consuming task not adequately compensated. In addition, other considerations such as the researcher liability in releasing datasets, unclear dataset ownership, or the unavailability of a repository to the researcher are factors that hinder data sharing best practices.

Problem 2. 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,

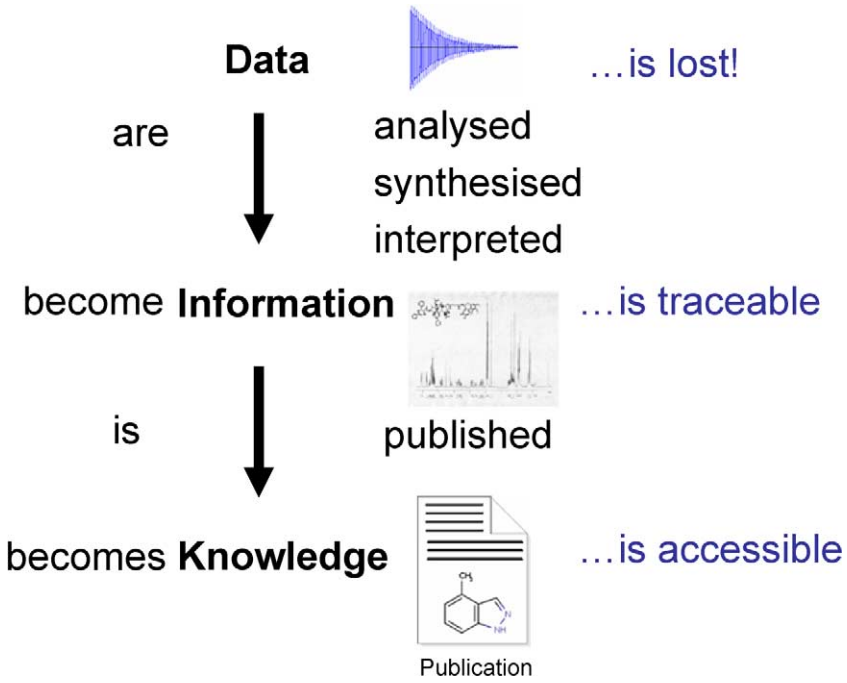

Fig. 1. From data to knowledge through publication.

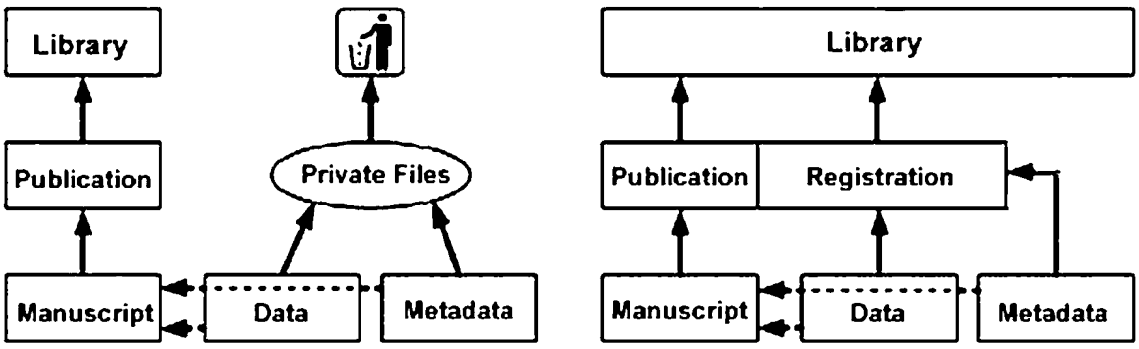

Fig. 2. The traditional publication method for datasets on the left, a possible new structure on the right.

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 Fig. 2):

(i) poor preservation properties (e.g. if the researcher moves to another institution, the link may become invalid);
(ii) poor quality of the documentation;
(iii) limited impact and academic recognition (dataset cannot be searched or found except from article reference or web search);
(iv) lack of data quality assessment.

2. Data citation techniques and integration with text

Integration of data into texts and unlocks citation services provides incentives for a researcher to publish datasets. Moreover, the publication of datasets and the inclusion of the dataset into library catalogues improves its potential impact since more researchers will become aware of its availability.

Currently, a large number of datasets are either directly referenced through their location (e.g., a URL) or through community-specific registries. URLs are subject of preservation issues since location of resources may change over time. Community specific registries introduce interoperability issues (due to heterogeneity in resolution services). Moreover, resolution of such identifier is difficult, since often one has to first identify the specific resolution engine that was used for issuing the identity.

In a number of scientific communities, there is no established data repository or dataset quality assessment protocol. In such cases, datasets are published by researchers using ad-hoc approaches. For instance, a group may make a dataset available on a web-page and communicate about it through an article, providing a URL to the dataset. Such a publication model raises preservation and quality concerns. Documentation of the dataset, if provided, would typically be based on an ad-hoc document format. License and copyright for using the dataset would often be unclear and it would not be possible to leverage metadata harvesting protocols for improving the visibility of the dataset. Publishing a dataset using standards compliant methodology (e.g., with Dublin-core metadata) is time consuming. In addition, these protocols are often not known by researchers, as dataset curation is not part of their specialties and tasks.

For academic researchers, dataset publication is not rewarded by an academic recognition proportionate to the effort. For a dataset to "count" as a publication, it would need to follow similar publication process as an article: be properly documented, be reviewed for quality, be searchable in catalogues, and be citable in articles. Moreover, in a way equivalent to citation count for articles, dataset usage needs to be measured to provide an indication of impact on the scientific community and a driver of academic recognition.

Research is a global endeavour and dataset identification and cross-referencing shall be accomplished at a global level. Raising the awareness of researchers of available datasets is also important for providing the best research possible. Similarly, publicizing the availability of dataset resources worldwide is essential, to achieve a full valorization.

2.1. Global awareness

Access to research has become an important issue throughout the world, identified by different organizations and individuals.

In its 2007 report "Cyberinfrastructure Vision for 21st Century Discovery" [14] the National Science foundation (NSF) remarks:

> Science and engineering research and education have become increasingly data-intensive as a result of the proliferation of digital technologies, instrumentation, and pervasive networks through which data are collected, generated, shared and analyzed.
> Worldwide, scientists and engineers are producing, accessing, analyzing, integrating and storing terabytes of digital data daily through experimentation, observation and simulation. Moreover, the dynamic integration of data generated through observation and simulation is enabling the development of new scientific methods that adapt intelligently to evolving conditions to reveal new understanding.
> The enormous growth in the availability and utility of scientific data is increasing scholarly research

productivity, accelerating the transformation of research outcomes into products and services, and enhancing the effectiveness of learning across the spectrum of human endeavour.

In 2007 the Organisation for Economic Co-operation and Development (OECD) has published their "OECD Principles and Guidelines for Access to Research Data from Public Funding" [15]. It identifies the important aspects from the perspective of the public funders:

> The rapid development in computing technology and the Internet have opened up new applications for the basic sources of research – the base material of research data – which has given a major impetus to scientific work in recent years [. . .].
>
> Besides, access to research data increases the returns from public investment in this area; reinforces open scientific inquiry; encourages diversity of studies and opinion; promotes new areas of work and enables the exploration of topics not envisioned by the initial investigators.

Further initiatives or reports addressing the issue of research data are for example:

- The report of *The Interagency Working Group on Digital Data* (IWGDD) in the US [13].
- The report *Shared Responsibilities in Sharing Research Data: Policies and Partnerships* by the European Science Foundation (ESF) and the German Research Foundation (DFG) [7].
- The *Strategic Coordinating Committee on Data and Information* (SCCID) established by the International Council for Science (ICSU).
- The *Digital Curation Center* (DCC) in the UK (http://www.dcc.ac.uk/about).
- The European *Alliance for Permanent Access* (http://www.alliancepermanentaccess.eu/index.php?id=1).
- The *UK Research Data Service* (UKRDS) (http://www.ukrds.ac.uk/).
- *Australian National Data Service* (ANDS) [3].
- *Research Data Canada*, by the National Consultation on Access to Scientific Research Data (NCASRD), Canada (http://data-donnees.gc.ca/eng/ncasrd/index.html).

2.2. State-of-the-art: Data infrastructures in Europe

Europe has a large number of infrastructures that cater for dataset publication needs at various levels.

Horizontal infrastructures are providing generic ICT services for dataset publication, storage or processing. The pan European backbone network GEANT and the Enabling Grids for E-sciencE (EGEE) project are representatives of such horizontal infrastructures, providing respectively connectivity and grid services. They form the bottom layer of commodity services (data storage, data transport, computation, etc.) that may be used for any sort of research, from physics to biology through social science.

In contrast, vertical infrastructures provide community specific solutions for achieving data sharing in a particular discipline. These solutions typically cater for all the steps in the dataset publication workflow, allowing online submission of datasets, their registration, their publication as well as advanced search and exploration using graphical user interfaces. Examples include:

- Art and humanities repositories: Archeology Data Service (ADS) TextGrid; Netherlands Historical Data Archive (DANS).
- High energy physics: CERN, DESY.
- Biology: BioGRID (interaction dataset).

Ideally, vertical infrastructures for data curation would be built on top of lower level infrastructures. This is indeed the case. For instance, any digital library or repository today leverages the ubiquitous

connectivity offered by the Internet. However, higher level functionalities are largely based on in-house developments that are not easily interoperable between communities. This is easily explained considering that:

1. *Legacy:* At the time of setting up community infrastructure, no high level commodity for dataset curation was available. This has forced communities to implement their own solutions to similar problems. Due to national initiatives, it is also common to find several distinct vertical solutions addressing a single scientific community.
2. *Heterogeneity:* Research datasets are very heterogeneous in form, complexity, size and nature. Radically different requirements from discipline to discipline make a one-size-fit-all approach to vertical solutions doomed to failure. Integration and homogeneity are desirable but should not be achieved at the expense of truly functional solutions closely aligned with the needs of the community that use it. Therefore, generic high level solutions would require extensive customization to address community-specific requirements.

3. 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 [10]. 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 stable 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.

3.1. Identifier schemes

Identification of electronic resources through persistent identifiers such as Digital Object Identifier (DOI) names or Uniform Resource Names (URN) is a well-known solution to the long term preservation of references. This approach is already widely used in long term preservation and the traditional publication world. For data access via the Internet, references provided by means of identifiers provide the location of the desired dataset in a way that is reliable and available over a long time [16].

A persistent identifier clearly identifies units of intellectual property in a digital environment and serves for administration of these units irrespectively of form and granulation. It allows the citation of the digital resource (in our case dataset) and more importantly, identifiers allow also cross-linkage of digital resources, for instance, datasets to reference articles or to source datasets from which they have been derived. Finally, since the provision of the dataset identifier is achieved through a registration mechanism, it gives specialized actors of data curation the possibility keep track of the resource, index it in large catalogues and thereby dramatically improve the potential impact of a dataset publication.

All these aspects have been identified by the scientific community as valuable and crucial for a better usage of scientific datasets [9].

A persistent identifier scheme always addresses two issues: The definition of the structure and syntax of the identifier itself; and the provision of a technical infrastructure for resolving. Today there are many different persistent identifier schemes used worldwide. The most common are URN, ARK, PURL and DOI.

URN: The formal description of the Uniform Resource Name (URN) was presented in 1994, its syntax was fully specified in 1997 as a standard from the Internet Engineering Task Force (IETF). There is, however, no central institution organizing the URN; there is no central resolution infrastructure. The URN is more a general concept with isolated implementations. In 1999 the Conference of Directors of National Libraries (CDNL) introduced the National Bibliography Number (NBN) as part of the URN system. The major national libraries in Europe assign URNs starting with urn:nbn and offer a mutual resolving infrastructure.

There is however no central resolving infrastructure. When resolving a URN, it is always crucial to know where to locate the appropriate resolving mechanisms. Furthermore, there is no standard definition of metadata schemes. There are no licence costs involved for assigning URNs. Each URN registration agency however has to establish an assigning and a resolving infrastructure.

PURL: The Persistent Uniform Resource Locators (PURL) were introduced 1996 by the Online Computer Library Center, Inc. (OCLC). PURLs are based on the http redirect mechanism. They offer a minimalistic technical approach in including a resolver address in the URL of a resource with a central resolver at the OCLC.

ARK: The Archival Resource Key (ARK) was introduced in 1995 by the California Digital Library (CDL). Like PURLs they are embedded in the htpp protocol and managed by the CDL as central organization with central resolver.

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 standardization 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).

Technically all of these persistent identifier systems could be used to register scientific datasets. The advantage of the DOI system lies in the possibility to establish citable datasets that can be handled as unique, independent scientific objects and are accepted as reference items by the STM publishers. The DOI system is well established and already part of the consciousness of the scientific community.

3.2. 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 [6].

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 600,000 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) [4].

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 [1].

Persistent identifiers are different from and complementary to local identifiers used in repositories. Local identifiers are useful for domain-specific applications or for local database management reasons. They can be used to reference the resource externally, but their validity is limited in time since such reference assumes the digital resource will remain in its current repository and that the repository structure will not evolve. Both assumptions are systematically proven wrong in the long run. By contrast, persistent identifiers are associated with the resource and remain identical regardless of the resource location; they are the preferred means for identifying the resource outside of the scope of the local system. Very often, a resource would have both a persistent and a local or domain-specific identifier. A common practise consists in building the persistent identifier from the local one at the time of registration. For instance, a DOI could look like: DOI:10.1594/***some domain specific ID***.

3.3. 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 3 illustrates this structure in more detail.

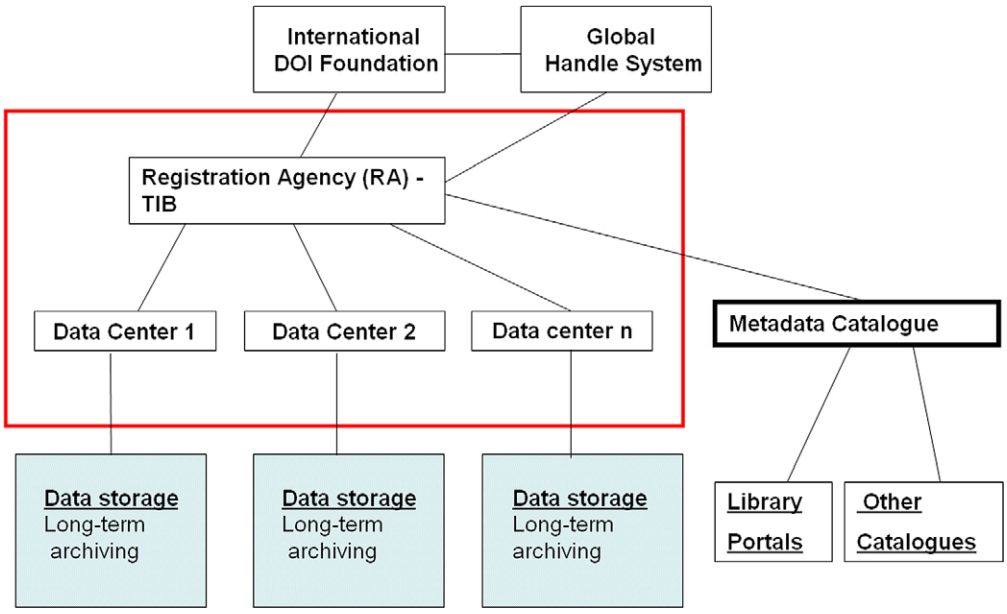

Fig. 3. The overall structure of TIB's DOI Registration Agency.

Like for every persistent identifier, costs for infrastructure, personnel and license are involved for assignment of DOI names. TIB has three ways of re-financing its costs for the DOI license and infrastructure:

- TIB has customer-relations with data centres that receive DOI names for the content.
- Costs for the registration of content that is of national interest are covered by the base funding of TIB as German National Library for all areas of engineering as well as architecture, chemistry, information technology, mathematics and physics.
- Registration of content that is a result of community funded research can be registered in cooperation with the funding agency by including the costs in the funding.

3.4. Dataset access through library catalogues

Library catalogues are classical sources for information [8]. The assignment of persistent identifiers allows further awareness of available datasets, when research data become directly accessible through library catalogues. When querying for a certain topic, users will not only receive all relevant publications as result, but also datasets collected by the corresponding researchers. Through dataset publication, researchers who collected data will gain further scientific reputation. This represents a further motivation for researchers to prepare collected data for online publication. Figure 4 shows the dataset mentioned above as result of a query to the online catalogue GetInfo of TIB.

For a registered dataset the TIB stores all relevant bibliographic metadata about the dataset. This metadata is consistent with ISO 690-2 for the citing of electronic resources and is automatically mapped to the libraries catalogue format. There is however the need for more and better metadata schemes when dealing with scientific data. At present GetInfo is the only major library catalogue in Europe to include scientific datasets.

4. Joint DOI registration agency for scientific content

Access to research data is nowadays defined as part of the national responsibilities. As shown, during the last years most national science organizations 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 organizations.

The key point of this approach is the establishment of a Global DOI Registration agency for scientific content that will offer to all researchers dataset registration and cataloging services. This joint agency 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.

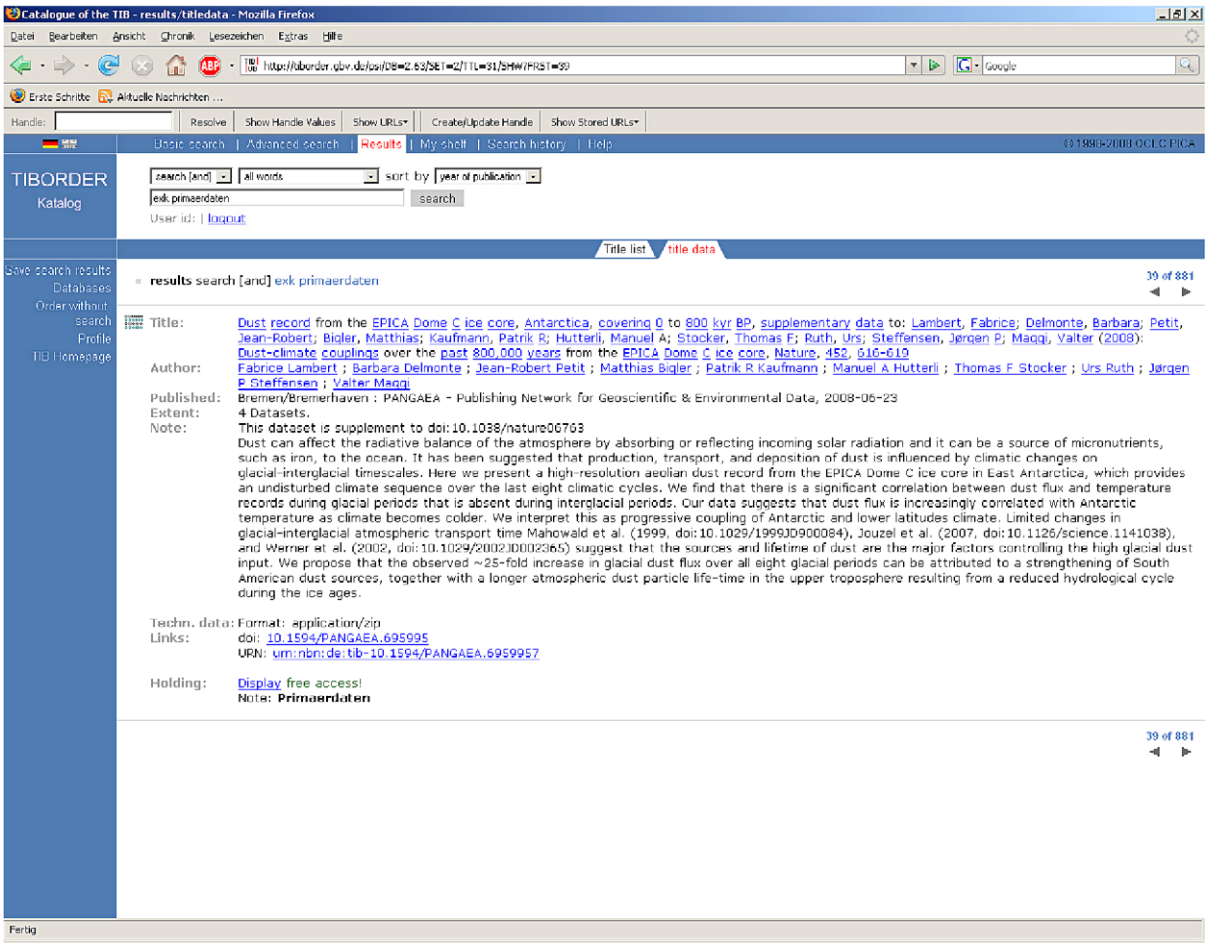

Fig. 4. A scientific dataset as result of a query to TIB catalogue – GetInfo.

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 this new DOI RA 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.

4.1. Roadmap

In a first phase the model of TIB will be opened to local agencies. These are libraries or information institutions with a national mission that includes the challenge of access to datasets. These local agencies will be direct partners of TIB and may use its infrastructure and license for DOI registration. On national level these local agencies will appear as directly responsible for the DOI registration (see Fig. 5).

In the second and final phase a new RA will be funded. This new RA will take the place of the TIB RA in the International DOI Foundation (IDF). It will be open for any information institute or library to join. The independent global DOI RA shall inherit TIB registration license and offer the existing services to other local institutions (see Fig. 6).

The structure of this DOI RA 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 consortium partner 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. If partners use their own handle server for registration these handle server will legally be operated by the joint RA. There will be one central metadata repository containing the descriptions of all registered data sets, with standardized interfaces to the partners own repositories and applications.

The metadata and workflow definitions will be standardized through all partners.

Every partner including TIB will cover the personnel costs at their offices. The costs for the DOI licences and registered DOI names will be shared by all partners, weighted by the amounts of DOI names registered by each partner.

Every partner will have the right to develop its own business models for re-financing the registration costs.

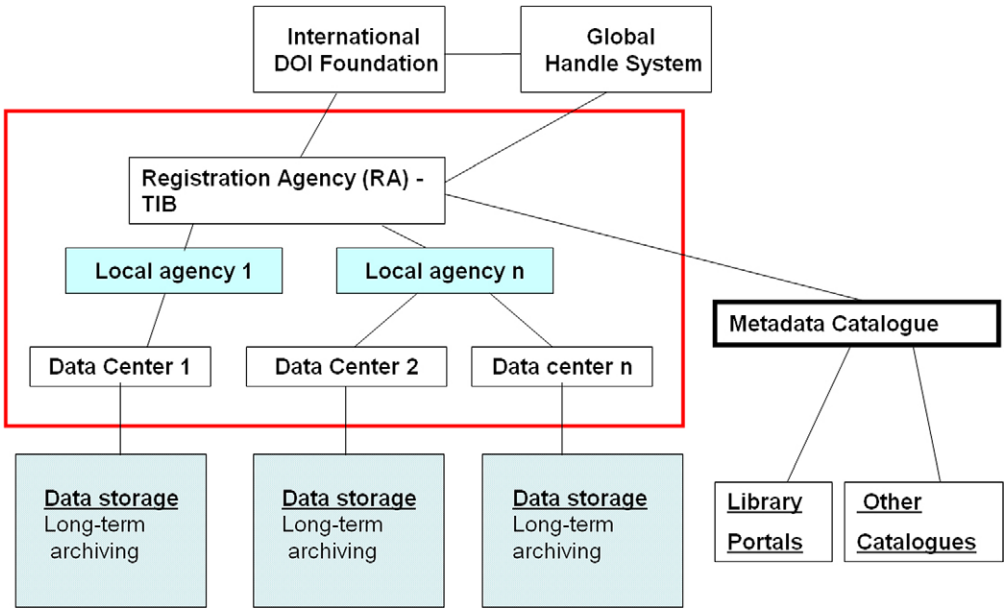

Fig. 5. The first phase of cooperation. National agencies as direct partner of TIB and responsible for their local data centers.

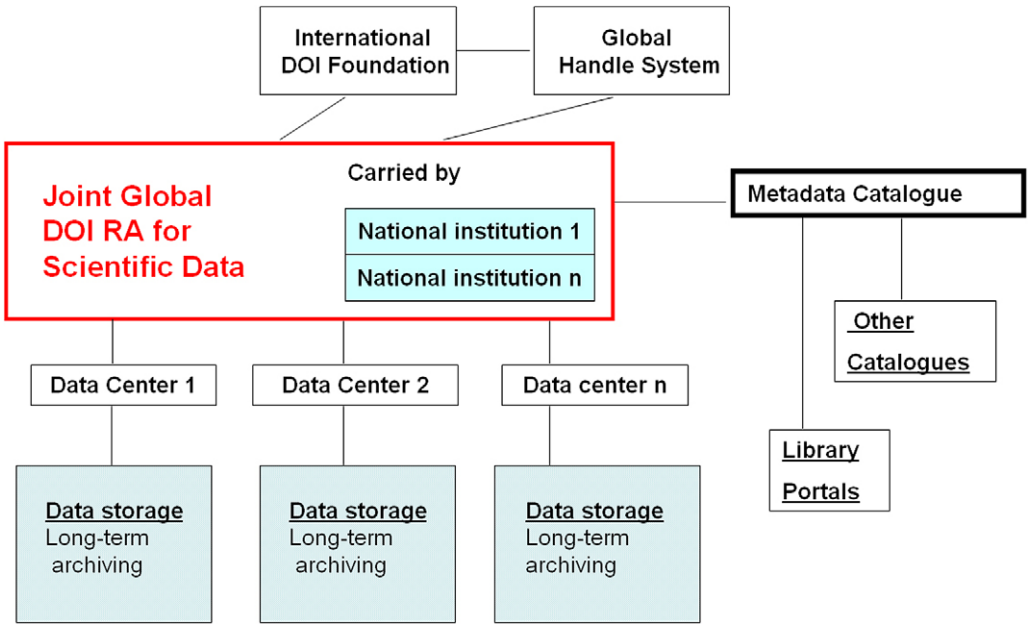

Fig. 6. In the final phase a new independent DOI RA take the place of the TIB RA.

The consortium will always remain open for other institutions to join under the same rules and obligations.

4.2. *Partners*

Institutions that have already expressed their interest to establish this agency are (in alphabetical order):

- *British Library (BL), UK*: 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.
- *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 higher education 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.

4.3. Memorandum

On 2 March 2009 the partners signed the following Memorandum of Understanding during the meeting of the International Council for Scientific and Technical Information (ICSTI) to establish a partnership to improve access to research data on the internet.

MEMORANDUM OF UNDERSTANDING

Recognizing the importance of research datasets as the foundation of knowledge and sharing a common commitment to promote and establish persistent access to such datasets, we, the signed parties, hereby express our interest to work together to promote global access to research data.

Our long term vision is to support researchers by providing methods for them to locate, identify, and cite research datasets with confidence.

In order to achieve this long term vision, we will establish a not-for-profit agency that enables organizations to register research datasets and assign persistent identifiers to them. The agency will take global leadership for promoting the use of persistent identifiers for datasets, to satisfy needs of scientists. It will, through its members, establish and promote common methods, best practices, and guidance. The organizations will independently work with data centres and other holders of research data sets in their own domains.

As a first step, this agency will build on the approach developed by the German National Library of Science and Technology (TIB) and promote the use of Digital Object Identifiers (DOI) for datasets.

Signed this day of March 2nd, Paris, France

Uwe Rosemann, *Director, German National Library of Science and Technology, Germany*
Wolfram Neubauer, *Director, ETH Library Zürich, Switzerland*
Herbert Gruttemeier, *Head of International Relations, Institute for Scientific and Technical Information, France*
Adam Farquhar, *Head of Digital Library Technology, The British Library, UK*
Mogens Sandfaer, *Director, Technical Information Center of Denmark*
Maria Heijne, *Director, TU Delft Library, The Netherlands*

References

[1] M. Altman and G. King, A proposed standard for the scholarly citation of quantitative data, *D-lib Magazine* **13**(3/4) (2007); doi: 10.1045/march2007-altman.

[2] P. Arzberger, P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Moorman, P. Uhlir and P. Wouters, Promoting access to public research data for scientific, economic, and social development, *Data Science Journal* **3** (2004), 135–152.

[3] The ANDS Technical Working Group, Towards the Australian Data Commons, A proposal for an Australian National Data Service, October 2007.

[4] J. Brase, Using digital library techniques – registration of scientific primary data, *Lecture Notes in Computer Science* **3232** (2004), 488–494.

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

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

[7] Shared responsibilities in sharing research data: Policies and partnerships, Report of an ESF–DFG Workshop, 21 September 2007.

[8] S. Inger and T. Gardner, How readers navigate to scholarly content, 2008, http://www.sic.ox14.com/howreadersnavigatetoscholarlycontent.pdf.

[9] J. Klump et al., Data publication in the Open Access initiative, *Data Science Journal* **5** (2006), 79–83. ISSN: 1683-1470, doi: 10.2481/dsj.5.79.

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

[11] Lawrence, S. et al., Persistence of Web references in scientific research, *IEEE Computer* **34**(2) (2001), 26–31; available at: http://www.fravia.com/library/persistence-computer01.pdf.

[12] Harnessing the power of digital data for science and society, Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council, Washington, DC, January 2009.

[13] D. Butler, Agencies join forces to share data, *Nature* **446** (2007), 354; doi:10.1038/446354b.

[14] National Science Foundation (NSF), *Cyberinfrastructure Vision for 21st Century Discovery*, Cyberinfrastructure Council (CIC), NSF, Arlington, VA, 2007.

[15] Organisation for Economic Co-operation and Development, *OECD Principles and Guidelines for Access to Research Data from Public Funding*, OECD Publishing, 2007.

[16] N. Paskin, Digital object identifiers for scientific data sets, in: *19th International CODATA Conference*, Berlin, Germany, 2004.