Birte Christensen Dalsgaard

Data Reuse and how Metadata can Stimulate Reuse

A workshop held on Monday 5th of December 2011 at the 7th IDCC conference in Bristol: http://www.libereurope.eu/news/e-science-workshop-on-data-re-use-how-can-metadata-stimulate-re-use

1. Introduction

The LIBER e-Science group organizes four workshops addressing different views on the role that libraries may play in the areas of e-science, research data and digital preservation. In a technical workshop that was held last December at the IDCC in Bristol the workshop participants were confronted with these three large elephants in the room. To prepare and stimulate the discussion among the workshop members, four introductory presentations were given. David Giaretta, director of the Alliance for Permanent Access (APA) gave a keynote presentation in which he outlined the vision of the high level expert group on what a scientific e-infrastructure should look like in 2030 [2].

Rob Grim from Tilburg University and the Open Data Foundation (ODaF) then gave a talk on e-Science, research data and the role of libraries [12]. The next speaker was Dave Reynolds (Epimorphics Ltd), who has been deeply involved in the Data.gov.uk initiative. Dave presented on Linked Open Data (LOD) and how LOD is used in practice [1]. The last speaker was Karen Morgenroth from the Canadian Institute for Scientific and Technical Information (CISTI) and National Research Council of Canada (NRCC). Karen introduced DataCite to the workshop members. Karen showed how DataCite makes research data citation easy and how DataCite fosters a community of practice [3]. Birte Christensen-Dalsgaard, from the Royal Library Denmark, chaired the 2nd LIBER workshop.

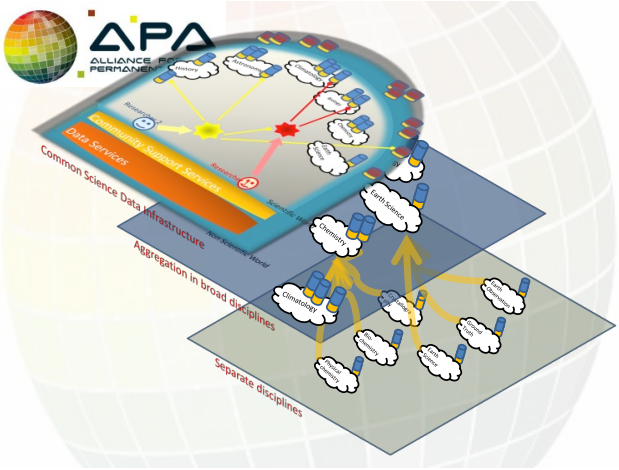
2. David Giaretta: Riding the Wave; Preservation and Adding Value by Enabling Reuse

The key message of the 2030 Vision that David presented was that a scientific e-infrastructure should support seamless access, use, reuse and trust of data. David then addressed the impact on science and society that could be achieved if the 2030 Vision should become reality and who -and why- should pay for digital preservation. After this general introduction key concepts for digital preservation were discussed, as well as more technical aspects of digital preservation. Examples of different types of digital objects were presented and the audience was introduced to the OAIS reference model[1], which is generally accepted as the *de facto* standard for building digital archives.

The OAIS model provides useful terminology for managing the functional areas of digital preservation. In addition to adequate functional terminology however, an information model is needed which details the representation of digital information, so that it can be used i.e. for archival information packaging (AIP) and distribution and reuse in a networked environment. David also gave examples of implementations of the information model in which he explained how formal descriptions of structure and semantics in an information model are used to preserve software functionality and to facilitate reuse of research data. David then illustrated the complexity and plethora of phenomena for digital preservation that arise from static and dynamic databases and rendered and non-rendered digital information. The interested reader is encouraged to read David's recent book on advanced digital preservation [6].

[1] OAIS: Open Archival Information System (OAIS), also known as ISO 14721.

David Giaretta: A layered view of the APA on a common science data infrastructure.



3. Rob Grim[2]: e-Science, Research Data and the Role of Libraries

Rob Grim illustrated the kind of problems that can be solved with *metadata management*, how libraries can use metadata management to support research and what sort of data services libraries could develop. Rob started quoting Jim Gray's definition of e-Science. According to Jim Gray e-Science is where IT meets science and is about digital *curation*, automated *capture* and *tool* development. Rob argued that metadata are crucial to all of these aspects of e-Science as the following examples may illustrate:

1 Metadata management already plays a central role in the global data infrastructure for statistical data exchange. SDMX was used to illustrate how metadata can be used to identify, discover, exchange and disseminate statistical data via the internet [14].
2 Metadata are also increasingly used to enhance data-intensive networked infrastructures. Rob gave an example where metadata content caching is used in wireless networks to optimize data retrieval.
3 In addition tools are created throughout scientific disciplines that capture critical aspects of primary data when the data are collected, generated or processed. These tools are in fact all metadata oriented. DDI 3 was then used as an example of a standard that was explicitly designed to capture and manage the digital life of research data. DDI 3 is a complex standard as it incorporates many requirements from archivists, librarians and statisticians. Despite its complexity, DDI 3 does what it is supposed to do i.e. identify, version and maintain object relations and therefore appeals to an active software developer community[4].

Rob signified four functional areas where libraries could develop services for research data. These functional areas are best described by the following general headings:

[2] See: http://www.slideshare.net/RobGrim/escience-research-data-and-libaries

1. Collection development. Libraries could get involved in archiving research data for a limited or longer period of time. Libraries could also start developing research data collections that are relevant to local research communities and which are supportive to teaching purposes, cross disciplinary projects and newly hired researchers (PHDs, etc.).
2. Research data discovery services. Subject portals for example can be dramatically improved by using custom search engine technologies that have access to metadata repositories and – registries. Negotiating an Open Metadata agreement with commercial data providers might be a prerequisite for libraries that want to develop efficient and fine grain research data discovery services.
3. Supportive environments for secure - and open access to research data. Who should have access, to what, when and what should a user be allowed to do? Academic libraries traditionally provide a state of the art environment, which gives access to distributed collections and resources. Libraries could use their knowledge of library and information systems to integrate access to research data.
4. Research data delivery services. Research datasets and supplementary materials are only useful if the user is provided with sufficient metadata and adequate documentation on what the data contains. Libraries can contribute in many ways to research data services that are valued by the research community. Curation of research data, linking research datasets to publications and providing support for research data dissemination are only a few examples of activity areas that might be of interest to academic libraries.

4. Dave Reynolds: Linked (Open) Data

Linked data is about publishing data on the web. Linked data is therefore “data you can click on”. Linked data enables easy integration, linking and reuse of data across silos and is well suited for describing things such as schools, companies, animal species, music tracks, TV programmes, etc [8]. But what about datasets? Can linked data also be used to describe environmental measurements, experimental results, and statistical analyses? Starting with a simple question “what information is relevant to the public about beaches”, Dave Reynolds gave an example of how linked data are used in practice to integrate environmental facts with end user applications.

Basically there are two approaches to publishing research data as linked data. In the first approach a research dataset is identified as a single resource that is simply identified with a URI. In addition to this, descriptive, categorical, technical and structural metadata pertaining to a dataset are usually also provided as linked data. Datasets that are published in this way support discovery services and can be easily aggregated into data catalogues.

The second approach to publishing linked research data is through fine grain publication. In this case each individual record or entity in a research dataset is identified by a URI, while the internal structure of a dataset is linked to one or more ontologies. These self-describing datasets allow for integration across datasets and the reuse of data dimensions, units and values.

One of the great advantages of linked data is that a dataset needs to be published only once and then can be consumed in many ways. Because linked data are easy to access, query and merge with disparate datasets, they are an attractive resource for both consumer and life science applications. Data.gov.uk provides many examples of linked datasets, APIs and visualizations on top of linked data.

Dave Reynolds: Example, how linked data helps

5. Karen Morgenroth: DataCite, an Introduction [3].

Karen Morgenroth from the Canadian Institute for Scientific and Technical Information (CISTI) and National Research Council of Canada introduced DataCite to the workshop members. Karen motivated clearly why research data should be registered, what the roles and responsibilities of different stakeholders are to achieve this, and how researchers and data producers can benefit from publishing research datasets. By providing a simple and recommended format for data citation, DataCite makes it easier for researchers and data producers to get credit and raise awareness for non-traditional forms of scientific output. DataCite therefore helps researchers to increase the visibility of their research output.

Karen started her presentation explaining the ideas behind the DataCite initiative. DataCite was challenged by the fact that there is no widely used method to identify and cite datasets and no effective way to link between datasets and articles. The interest in DataCite has been substantial from the beginning and DataCite now represents an impressive body of stakeholders, including datacentres, publishers, libraries, research organizations, science unions and funders. DataCite acts as a global registration agency for datasets. It uses the Digital Object Identifier (DOI) handle system to register datasets and the agency maintains the resolution infrastructure.

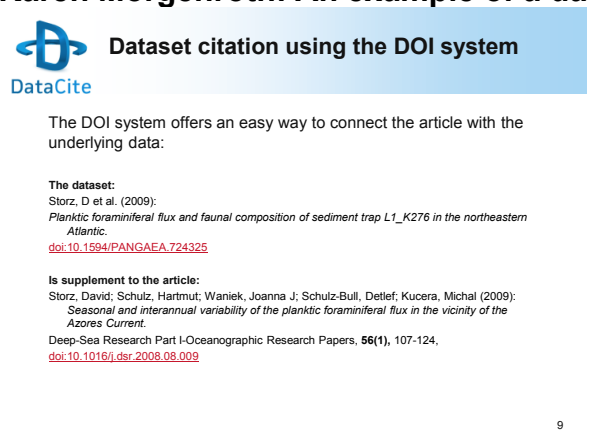
Data publishers are themselves responsible for quality assurance, content storage and providing access, generating identifiers and creating and updating metadata. Registration of datasets via DataCite simply facilitates sharing and reuse of data and provides a means for researchers to get credit for data citations. In addition, DataCite supports researchers by enabling them to locate and identify research datasets of interest. But DataCite also supports researchers and publishers by

providing a means to connect a scientific article with the underlying research data via a DOI.
Finally DataCite contributes to long-term accessibility of datasets by providing *permanent links* for research datasets.

Data Citation
DataCite recommends the following format for data citation: Creator (Publication Year): Title. Publisher. Identifier.

> Example: Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. Geological Institute, University of Tokyo. http://dx.doi.org/10.1594/PANGAEA.726855.
>
> See also data citation Dataverse Network [5, 9]

Karen Morgenroth: An example of a dataset citation with a DOI.



6. Report of the Working Groups

In preparation of this workshop three pockets of potential interest to libraries were formulated:

1. Infrastructure and access to research data (Open Access, Open Source and Open Data).
2. Technical aspects of privacy, digital identity and digital rights management for research data.
3. Research data support for cross- and multidisciplinary research.

It was also decided that separate working groups would discuss these topics during the workshop. Each working group was instructed to come up with a poster and present the main findings and recommendations of the participants. As there were no participants with sufficient expertise at the workshop to discuss the technical aspects of privacy, digital identity and digital rights management, the participants agreed to postpone the discussion for this topic. At the end of the workshop the members of the other working group evaluated each poster. Participants were stimulated to highlight the findings which they strongly agreed - or disagreed to.

Poster Infrastructure and Research Data Access

The following summary lists the issues and recommendations that were identified during the discussion by the workshop participants:

- Research data should be openly available as soon as - and whenever possible.

- The researcher data infrastructure should enable “Oracles of Trust” which allow for trusted assertions to be made over time based on research datasets and items. To build these Oracles of Trust a system for evaluating trust is needed. Such a system should be able to incorporate a cumulative body of knowledge for a research dataset. Everyone should be able to follow the lineage of annotations, annotations of annotations, etc., for a dataset.
- A potential role for the library could be in filling gaps that exist within research domains and areas that are not well covered. In addition, libraries can also have a crucial role in fostering the social infrastructure around research data. Capturing and curating research data for reuse in teaching would be an area where libraries could become (more) active for example. Another potential area of activity that was identified for libraries is providing data management - and data management planning support.
- What is the role of libraries for storing and archiving research data?
- Libraries could also play a role in the development of standards to further support research communities.
- Research data management introduces new problems for library collection management such as, what is an appropriate copy of a research dataset?

Poster Research Data Support for Cross- and Multidisciplinary Research

The following summary lists the issues and recommendations that were mentioned by the workshop participants:

- Interdisciplinary research requires ontologies. Can these types of ontologies be constructed?
- Do publishers need to deal with these topics? What should be their role?
- A quality mark for datasets adhering to some kind of standard might be needed. Such a mark should be attached at the dataset level. Quality parameters could be: a dataset is properly described, openly available, makes use of cross-discipline ontologies and is valued by peers. It was mentioned that such initiatives are already discussed for linked open datasets.
- Libraries could/should start lobbying for the importance of research data management and support the development of tenure tracks in this field.
- Libraries could in principle also contribute to funding submissions and research grant proposals.
- Can we agree - at least in each discipline - to a core set of descriptors for research datasets and data items?
- “According to the tradition, libraries need to accept all types and qualities of data. Does this same rule apply to research data?” If this is coming to us, we might need a more active role of libraries i.e. start influencing data management plans in such way that research data delivery can be aligned with library procedures.
- Data management might introduce a paradigm shift from managing largely static content to (highly) dynamic content.
- Whenever possible, detailed geographical -and time information should be attached to research datasets and items.
- After the datasets are deposited they might need additional post-processing to add structures that allow linking of disparate sources.

Summary of main findings for both working groups

At the end of the workshop a plenary wrap up was held and the issues of importance were listed which everyone could agree to. These issues are:

- Introduce a point/star system for the quality of datasets based on the associated metadata (content, context, etc), openness and adherence to standards. Those interested are referred to David Shotton who recently published an article on this subject in D-lib [15].
- Libraries should start influencing researchers to use standardized methods – through advice and help with data management plans.
- Libraries can play an important role in the social infrastructure for research data. Libraries can contribute to the use and re-use of research data, teaching facilities for research data and promoting open data access.

Epilogue

- All presentations should be accessible from the LIBER website: http://www.libereurope.eu/committee/scholarly-communication/wg-e-science
- Norbert Lossau has written an excellent overview of the research data infrastructures in Europe with recommendations for LIBER [11].
- The Nature Publishing Group recently released a linked data platform which might be of interest to readers [2].

References [7, 10, 11, 13, 15]

1. *Data.Gov.UK*. [http://data.gov.uk/].
2. *High-Level Group on Scientific Data. Riding the Wave: How Europe can gain from the rising tide of scientific data*, 2010. High Level Expert Group on Scientific Data.
3. DataCite. [http://www.DataCite.org/].
4. DDI. *Data Documentation Initiative*. [http://www.ddialliance.org/].
5. DVN. *IQSS Dataverse Network*. [http://dvn.iq.harvard.edu/dvn/].
6. Giaretta, D., *Advanced digital preservation*, 2011. Springer.
7. Grim, R. *e-Science, research data and libraries*. 2nd Liber workshop on e-Scholarschip at the IDCC [ppt] 2011; [http://www.slideshare.net/RobGrim/escience-research-data-and-libaries].
8. Hyland, B. and D. Wood, eds. *Linking government data*. ed. D. Wood. 2011, Springer: Fredericksburg, VA. 25.
9. King, G., *The Dataverse Network: An infrastructure for data sharing.* 2008.
10. LIBER. *e-Science Working Group*. [http://www.libereurope.eu/committee/scholarly-communication/wg-e-science].
11. Lossau, N., *An Overview of Research Infrastructures in Europe – and Recommendations to LIBER* 2012.
12. ODaF. *Open Data Foundation*. 2010; [http://www.opendatafoundation.org/].
13. OKF. *Open Bibliographic Data Working Group of the Open Knowledge Foundation*. [http://openbiblio.net/2011/11/21/recommendations-on-releasing-library-data-as-open-data/].
14. SDMX. *Statistical Data and Metadata eXchange*. 2010; [http://sdmx.org/].
15. Shotton, D. (2012) *The Five Stars of Online Journal Articles — a Framework for Article Evaluation*. **18**, DOI: 10.1045/january2012-shotton.