From cataloguing to digital curation: the role of libraries in data exchange

Susan K. Reilly
LIBER
Koninklijke Bibliotheek
Den Haag
+31 (0)703140160
susan.reilly@kb.nl

ABSTRACT

This paper describes the work of the Opportunities for Data Exchange (ODE) project, a project funded by the European Commission under Framework Programme 7. This project investigates issues surrounding data preservation, reuse and exchange from both sociological and technical view points.

Led by the European Organisation for Nuclear Research (CERN), the project has sought out stories of success and honorable failures. It has also brought together representatives of key stakeholder in the data preservation and sharing landscape. This has enabled dialogue between these stakeholder in order to identify opportunities for researchers, publishers and libraries to play their part in data exchange.

The growing need for research data preservation and curation services, the linking of data to publications, and increasing awareness of the potential of data sharing for innovation, presents a major opportunity for libraries to redefine their roles and embed themselves in the research process. In November 2011 ODE surveyed the 420 plus LIBER member libraries to establish what demand from researchers libraries are experiencing for support in data exchange, what roles they need to fulfill, and what new skills they need to develop and how. The results clearly emphasised the importance of the development of the role of the library in digital curation.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Data sharing

General Terms

Human Factors

Keywords

Research data sharing, digital curation, libraries

1. INTRODUCTION

Funded by the European Commission under Framework Programme 7, the Opportunities for Data Sharing[1] (ODE) project's aim was to identify, collate, interpret and deliver evidence of emerging best practices in sharing, re-using, preserving and citing data, the drivers for these changes and barriers impeding progress.

This was done in forms suited to its target audiences/stakeholders of policy makers, funders, infrastructure operators, data centres, data providers and users, libraries and publishers.

The aim of the project has been to:

- Enable operators, funders, designers and users of national and pan-European e-Infrastructures to compare their vision and explore shared opportunities
- Provide projections of potential data re-use within research and educational communities in and beyond the ERA, their needs and differences
- Demonstrate and improve understanding of best practices in the design of e-Infrastructures leading to more coherent national policies
- Document success stories in data sharing, visionary policies to enable data re-use, and the needs and opportunities for interoperability of data layers to fully enable e-Science
- Make that information available in readiness for FP8

Within this context, the stakeholder representatives in the project have worked together to engage and raise the profile of data sharing, re-use and preservation as an issue with each of our communities and to undertake further, in-depth, investigation into the issues raised.

LIBER, the Association for European Research Libraries, represents over 420 research libraries from across Europe. Through LIBER, ODE has engaged research libraries in Europe in the dialogue surrounding data exchange on issues such as linking data to publications, best practice in

[1] www.ode-project.eu

data citation and, subsequently, exploration of the role of libraries in supporting data exchange.

2. IDENTIFYING OPPORTUNITIES

The data deluge and its implications has been explored by the High level expert Group on Scientific Data in the Riding the Wave report[2]. The report outlines the need for the development of an international framework for a collaborative data infrastructure. This framework is described as broad conceptual framework which outlines how stakeholders interact with the system, including a multitude of actors, with provisions for data curation at every layer.

ODE goes some way in exploring this interaction through the identification of common issues, drivers and barriers in data exchange. One of the ways in which these are explored is through an analysis of the impact that data sharing, re-use and preservation is having on scholarly communication[3]. The aim of this analysis is to identify incentives for researchers and other stakeholders that will help to optimise the take-up of future e-Infrastructures.

One of the key areas of opportunity in terms of exploiting and proving the value of data exchange is scholarly communications. The opportunity to share and interact with research data is changing the face of scholarly communication and creating new opportunities and challenges for researchers, publishers and libraries. Publishing the underlying data of an article creates greater transparency and potentially further research, but it must also be in the interest of the data creator to publish and the data much be published in a manner which is sustainable. Three areas have been examined by ODE in relation to scholarly communications: linking data to publications, best practice in data citation, and the evolving role of libraries.

The findings of the exploration of linking data to publications were published in a report, which sought to reveal opportunities for supporting a more connected and integrated scholarly record. Four perspectives were considered, those of the researcher, who generates or reuses primary data, publishers, who provide the mechanisms to communicate research activities, and libraries & data centers, who maintain and preserve the evidence that underpins scholarly communication and the published record.

Before identifying opportunities it is necessary to look at the different layers (fig.1.) of data publication and identify issues associated with each layer.

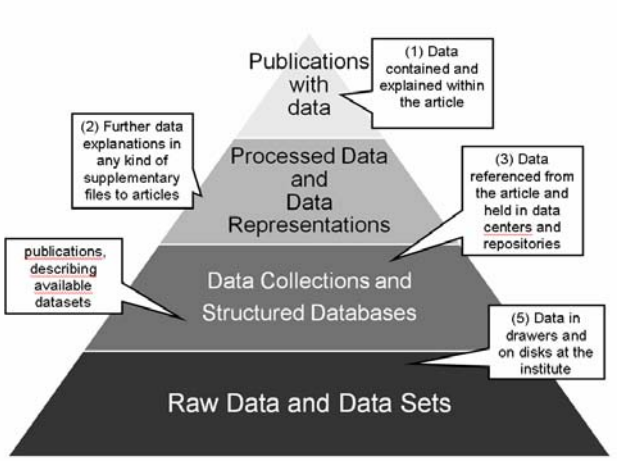

Figure 1. ODE Data Publication Pyramid

Each layer presents different challenges and opportunities e.g. one of the challenges presented by bottom layer of raw data is to encourage researchers to deposit their data in a sustainable infrastructure. The report identified opportunities for all three groups in seven key areas:

1. Libraries have the opportunity to support **availability** by helping researchers make their data available and also providing search services for data.

2. Through the provisions of support for best practice in managing data they can support **findability**.

3. As experts in metadata they can support **interpretability** through the provision of, and training in, metadescriptions.

4. By advising on the availability of subject archives and licensing for reuse libraries can help work towards ensuring the **reusability** of research data.

5. By encouraging best practice in citations through the provision of guidance and training, and through the use of persistent identifiers for data sets libraries play a role in improving the **citability** of data sets.

[2] The High Level Expert group on Scientific Data (2010), Riding the Wave, http://www.cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[3] Reilly et al. (2011) ODE report on the integration of data and publications:http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/10/ODE-ReportOnIntegrationOfDataAndPublications.pdf

6. Libraries can also take on some responsibility for the **curation** of data and provision of training on data curation

7. Contribute to the long term **preservation** of data by advocating for good data management practices and the archiving of data.

In essence, developments in linking data to publications and, more broadly, data exchange presents libraries with the opportunity to redefine their roles and become more embedded in the research process. Libraries should not underestimate their role as advocates for data sharing and for best practice in data management. This examination of linking data to publications also points to the fact that libraries are well placed to provide support for data curation across the layers of a collaborative data infrastructure.

3. REDEFINING ROLES

These seven areas of opportunity were presented to a group of librarians during a workshop at the 2011 LIBER Annual Conference in Barcelona. What emerged from this workshop was a very clear need for libraries to clarify their roles in relation to data exchange and the opportunities identified. Furthermore there is a need to understand these roles so that this can inform the identification of existing skills to be built on and new skills to be developed. The libraries were in consensus that they were in a strong position to address fragmentation in curation and archiving but there were doubts surrounding whether they were equipped to take decisions regarding what research data should be curated and archived or even what their role should be in making these decisions.

4. SURVEY OF RESEARCH LIBRARIES

The workshop established that libraries are keen to engage in data exchange but that further exploration of the types of roles libraries should play in this was needed. To follow up on this a survey was sent out to all 430 libraries in the LIBER network. The spectrum of libraries within the LIBER network covers national and state libraries, as well as university libraries and research institutes.

The survey was designed to gather evidence on the current and expected roles of libraries in regard to data management in order to prescribe steps for the evolution of these roles. This has been done through gathering answers from libraries related to the following questions:

1. What is the perceived demand from researchers for support for data management from libraries?

2. In what areas does this demand exist?

3. What support is currently in place?

4. What skills are needed to meet the demand for support?

In total 110 responses were gathered, from a mailing to LIBER members that reaches approximately 800 people (response rate 13 %). Additional responses were gathered from a dozen internationally recognized leading libraries (experts) in the field of data management support from the US and Australia. As these select few were already active in the field their responses were meant to form a benchmark.

4.1 Survey Results

The responses to the survey make it clear that librarians regard their involvement in support for research data exchange as a new and important role. For the majority, the service level is still rather low, but librarians also appear keen to develop themselves in the area of data management, archiving and curation as well as in helping their researchers find data. The survey received a response rate of nearly 20% and so can be judged as representative of the state of play across research libraries in Europe.

4.2 Demand for Support

81% of the respondents reported a demand for data management support. Considering that the response came for the broad spectrum of European research libraries and not just large university libraries this is quiet a high figure. What came out most clearly in the survey results is that libraries are nowhere near meeting the perceived demand for support (fig.2). The area where most demand was perceived was for archiving data. 80% of respondents perceived a demand in this area, yet only 41% of these respondents actually provide any sort of support services for the archiving of data.

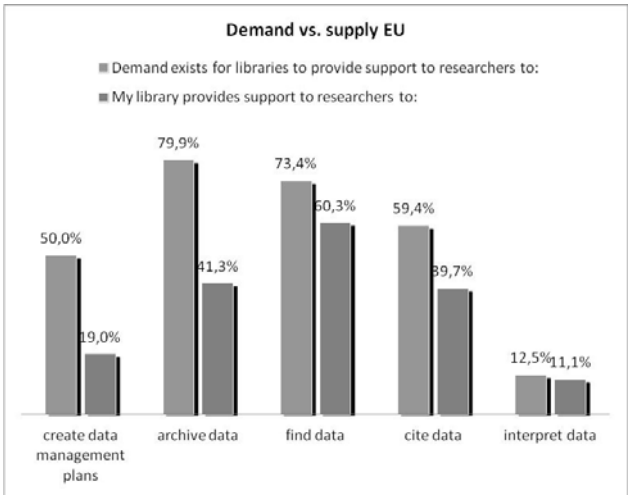

Figure 2. Demand v. Supply

When asked who should be responsible for the selection of data for archiving, respondents saw this responsibility as lying with researchers, followed by data librarians, librarians and others. Interestingly, the Expert libraries differ with this opinion in that they all agreed that only the researchers should be responsible for the selection, and no one else. There were qualifiers to this response stating that libraries should work with researchers to do this.

The majority of respondents (66%) also reported that their institutions did not have strategies in place for the preservation of research data, which is a worrying gap.

Only 6% of libraries had an internal archiving system for archiving research data and furthermore only 10% cooperated with a disciplinary data archive. That only 6% of respondent libraries have an archive for research data is not necessarily worrying. In many cases it is preferable and, arguably, more sustainable, to encourage researchers to use a disciplinary archive. What is of concern, is the fact that so few libraries seem to be collaborating with disciplinary archives. This is not just worrying from a library perspective, but also from the perspective of those funding such infrastructures. To ensure successful uptake and exploitation, such research infrastructures should be working with stakeholders such as libraries to help them to encourage and provide support to their researchers to use these collaborative infrastructures.

An encouraging figure is that a greater number of libraries are employing their traditional cataloging related skills when it comes to making sure that data remains interpretable and reusable. 39% of respondents report that they use metadata to ensure this. This shows that libraries are already adapting their existing skills to meet the increasing demand for data management support.

4.3 Developing Skills

The two key areas where skills need to be developed are IT and data curation. Responses from Europe showed that IT skills were seen to be the most important area for skills development. On the other hand, the Expert libraries strongly prioritised digital curation as an area for the development of skills. For them, IT skills came in 4th place in terms of priority. It may be that experience shows that IT skills are not as important as perceived for libraries who are actively involved in data management support.

The best means of developing all such skills, according to the libraries, is through the provision of continuing professional development. During times when budgets are contracting it is not realistic to expect to be able to recruit new skill sets externally. Instead libraries must, where possible, invest in developing the skills of existing staff. This solution may not be entirely sufficient, particularly when it comes to the need for subject specific expertise. Subject specific expertise was prioritized by 88% of Expert libraries and 67% of European libraries. Ultimately, the demand for such expertise may lead to new approaches to the professional education and recruitment of librarians.

5. CONCLUSION

It is clear that there is an opportunity and demand for libraries to provide support in digital curation. What has not been so clear is what exactly the nature of libraries' role should be in this. Perhaps wisely, libraries have realized that they can not take on full responsibility for the curation of research data. Researchers must be involved in the selection of data for archiving. If researchers are to be solely responsible for this, then libraries should begin to consider how they can support researchers to make these decisions?

On the other hand, the current support for digital curation is not sufficient. Libraries can apply their traditional skills to this area but they must also invest in developing new skills to meet demand for support and to avoid what could be a very regrettable missed opportunity. A start would be to put strategies for the preservation of research data in place, these strategies might involved the establishment of an internal archive that supports the persistent identification of data sets or they could be as simple as collaborating with disciplinary archives on behalf of their own research communities.

6. REFERENCES

[1] ParseInsight Survey (2009) http://www.parse-insight.eu/downloads/PARSE-Insight_D4-GapAnalysisFinalReport.pdf

[2] Reilly et al. (2011) ODE report on the integration of data and publications:http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/10/ODE-ReportOnIntegrationOfDataAndPublications.pdf

[3] Survey Report PARSE.Insight:http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

[4] Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, et al. (2011) Data Sharing by Scientists: Practices and Perceptions. PLoSONE 6(6): e21101. doi:10.1371/journal.pone.0021101

[5] The High Level Expert group on Scientific Data (2010), Riding the Wave, http://www.cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf