

User Engagement in Research Data Curation

Luis Martinez-Uribe¹, Stuart Macdonald²,

¹ University of Oxford, e-Research Centre, 7 Keble Road, Oxford OX1 3QG, UK
luis.martinez-uribe@oerc.ox.ac.uk

² University of Edinburgh, EDINA, 160 Causewayside, Edinburgh, EH9 1PR, Scotland
stuart.macdonald@ed.ac.uk

Abstract. In recent years information systems such as digital repositories, built to support research practice, have struggled to encourage participation partly due to inadequate analysis of the requirements of the user communities. This paper argues that engagement of users in research data curation through an understanding of their processes, constraints and culture is a key component in the development of the data repositories that will ultimately serve them. In order to maximize the effectiveness of such technologies curation activities need to start early in the research lifecycle and therefore strong links with researchers are necessary. Moreover, this paper promotes the adoption of a pragmatic approach with the result that the use of open data as a mechanism to engage researchers may not be appropriate for all disciplinary research environments.

Keywords: digital curation, research data management, open data, digital repository services, user engagement

1 Introduction

Research methods and practice, including scholarly communication, are experiencing a radical transformation in the digital age. New tools and infrastructures make possible the generation of digital research data outputs as well as new ways to use, share and reuse them. There is a growing acceptance of the importance of curating research data in order to preserve them and make them re-usable for future generations with libraries, computing services and other service units within academic institutions working together to develop digital repositories to curate this type of research output.

We believe that engagement with researchers, the user communities in this case, is crucial in order to develop systems that will meet their needs. Whilst some argue that open data is the way forward, it is not clear that it will help engage researchers with digital curation activities. Thus this paper will attempt to answer the following research questions:

- Is open data the correct concept to engage the research community?
- What other methods can be used to facilitate engagement in data curation?

2 Open Access Repositories and Researchers' Requirements – A Balancing Act

Open Access (OA) enthusiasts have written about the inevitability of 24 hours a day and 7 days a week access to all research papers and their citations “*for free, for all and forever.*” [1] Primarily led by technological developments, the increase in the overall volume of research, the increasing uncertainty about content preservation and by the strong dissatisfaction of academic libraries subjected to constant increase of journal subscription prices, digital repositories were built and employed within research institutions far and wide [2]. Being content provider as well as user and re-user of these information systems, researchers can be regarded as the key user community. Nonetheless, it has been argued no formal detailed requirements analysis has taken place in order to identify and address researchers' needs and concerns related to such scholarly communication systems [3]. As a result the user community has been overlooked in the developmental phases of technology design and implementation of the information system ultimately meant to serve them. Arguably the repository infrastructure developed was not, in most cases, built to address researchers' needs but those of libraries' and librarians. This has led to a struggle to find ways to populate repositories with researchers' output. In recent times there have been an increase in the number of institutional and research funders OA mandates, the knock-on effect from which will see the requirement of significant investment in awareness raising activities in order to highlight the benefits to researchers of using and depositing research materials in such repository systems. Such a process may have been expedited had the library and research worlds been more closely involved in a more agile digital repository design and development with an iterative requirements phase.

3 Research Data Repositories - Learning From Experience

When it comes to the research data setting we have to approach the problem from a new perspective. We have to evolve and learn from previous experiences in order to develop repository services capable of dealing with the management and curation of research data by addressing researchers' needs.

Although open data is becoming a widely used term, there is not a consistent formalisation of the concept. Murray-Rust [4] suggests that the concept of open source software can be extended to that of open data in that data should be freely available for re-use and modification without restriction. The virtues of 'Open Data' have been praised and evangelized by many since OECD's declaration back in 2003 [5] however many research communities are currently not in a position to make their data available on those terms.

The JISC-funded DISC-UK DataShare project explored a number of technical, legal and cultural issues surrounding research data in repository environments. It built on the existing collaboration of data librarians and data managers from the

Universities of Edinburgh, Oxford, Southampton and LSE and investigated mechanisms for ingesting and sharing research data in existing institutional repository systems for those researchers willing to openly share them. Project partners identified a number of barriers pertaining to the researcher and the research setting that would impact on data sharing [6], including:

- Reluctance to forfeit valuable research time to prepare datasets for deposit, e.g. anonymisation, codebook creation, formatting
- Concerns over making data available to others before it has been fully exploited
- Concerns that data might be misused or misinterpreted, e.g. by non-academic users such as journalists
- Concerns over loss of ownership, commercial or competitive advantage
- Concerns that repositories will not continue to exist over time
- Unwillingness to change working practices
- Uncertainty about ownership of IPR
- Concerns over confidentiality and data protection

It may be argued that open data is a reality in some disciplines: in crystallography there are a number of established repositories including the Cambridge Crystallographic Data Centre and Crystallography.net with further discussions taking place about a federation of crystallography data repositories [7]. Molecular biologists have been sharing data through repositories like the Protein Data Bank (PDB) since the seventies [8] and geneticists have also been sharing nucleotide sequences through GenBank from the early eighties [9]. These are great examples of communities embracing the benefits of open data but it is important to highlight that these initiatives were led by visionary researchers in those fields. There is an interesting analogy with domain specific publication repositories like arXiv or RePEc. They represented successful examples of author self-archiving repositories but this didn't translate in wide acceptance and use of open access repositories.

4 Different Approaches - Researchers' Needs Connecting Data Management and Curation

A more research inclusive and bottom up approach has been taken by data and information management activities in the Universities of Edinburgh and Oxford in order to understand better how researchers work, what are the drivers behind their information management and sharing activities and what services they require.

At Edinburgh a team of social scientists and information service specialists (respectively, from the Institute for the Study of Science, Technology and Innovation, and from Information Services at the University of Edinburgh and the Digital Curation Centre) are carrying out a RIN-funded study designed to enhance understanding of how researchers in the life sciences locate, evaluate, manage, transform and communicate information as part their research processes, in order to identify how information-related policy and practice might be improved to better meet

the needs of researchers. Information diaries were completed by over 50 life-science researchers from eight sub-disciplinary research groups. An interview schedule was constructed in order to investigate further the findings from the diaries. 24 interviews were conducted across the groups followed by focus-group discussions. An in-depth study was also employed on one of the groups. Interim findings suggests that some disciplines lend themselves more than others to open data and that there's much variety, specificity and complexity in terms of research data within the examined groups. Research data created via models/simulations, observations, and experiments are intrinsically linked with the data collection methodologies and instrumentation and as such may be better placed within a Virtual Research Environment (VRE) and/or a staging repository-type environment [10] as there are often issues surrounding the unraveling of data content when sophisticated and domain-specific proprietary systems are used. In addition, certain data cannot be considered conventionally open for example: data controversial by nature (stem cell data, brain scans); data received from industrial partners, licensed data products and the ensuing derived data products, data leading to development of patents or commercial products. Other findings include:

- Most life science researchers spend much of their time searching for and organising data however data curation and/or sharing only becomes crucial to them at certain stages of the research process.
- The groups investigated lack any obvious or explicitly appointed data/information managers, leaving individuals to manage their own information/data in a non-formal fashion.
- There is an implicit feeling across the groups surveyed that only the researchers themselves have the subject knowledge necessary to curate their own research data.
- Researchers in the life sciences express a keen sense of 'ownership' and protectiveness towards their data. However there is confusion or uncertainty about their rights with respect to data ownership

In Oxford an internal scoping study [11] on research data curation took place throughout 2008 and involved the Office of the Director of IT, Computing Services, the Library and the Oxford e-Research Centre. The aim of the study was to capture the requirements for digital repository services to manage and curate research data. A requirements gathering exercise took place and around 40 researchers across disciplines were interviewed to find out about their data management practices and capture their requirements for services to help them manage data. The findings from this exercise showed that the vast majority of researchers felt that there were potential services that could help them. The following scenarios present some of the challenges, found during the scoping study, that researchers are facing with their data and represent the types of needs that data repository services should be trying to address:

- In some cases, researchers had generated data several years ago and now could not make sense of them as they had not kept enough information on how the data was created in the first place;

- In scientific disciplines research groups require secure storage for their large volume of data generated by instruments such as electronic microscopes or by computing simulations run in GRID systems;
- Many clinical research centres compiling data for decades and spending months to migrate data formats in order to avoid format obsolescence;
- In many cases researchers want to make their articles' accompanying data available online in a sustainable way and they do not have the institutional infrastructure to do this so they published the data on their departmental website.

The scoping study is now being followed up by the JISC funded Embedding Institutional Data Curation Services in Research (EIDCSR) project. This project will attempt to address the data management and curation requirements of two research groups who produce and share data. EIDCSR involves the partners of the scoping study with Research Services and IBM to integrate research workflows with the Fedora Digital Asset Management System, long-term file storage and underpinning these efforts with policy development and economic models. A key aspect of this project will be the possibility to work with researchers from the moment they generate the data, this will ensure that the necessary and appropriate curation actions are taken early in the research lifecycle.

Oxford and Edinburgh are also both involved in the development of the Data Audit Framework¹ (DAF) which helps to establish relationships with research communities around the issues of data curation. This methodology provides organisations with the means to identify, locate, describe and assess how they are managing their research data. The methodology goes some way to enabling data auditors to identify and engage researchers regarding their research data holdings. It also provides information professionals who wish to extend their support for research data within the university community with a vehicle for engaging with researchers in addition to a focus for discussion of data curation practices. This may manifest itself through local data management training exercises to equip researchers with the skills and tools to deal with funders' data management and sharing policies. Indeed, the Edinburgh Data Audit Implementation Project [12] states that 'staff had numerous comments and suggestions for improvement of data management at different levels indicating an awareness of the issues, even where it has not been made a priority to address.'

Engagement with researchers through the activities explained above provides a valuable insight into the research process at the various stages in its lifecycle. Such activities help to gain the trust of the researchers facilitating the process of data curation within data repositories at a point early on in the research lifecycle, a fundamental key to the success of these information systems. In addition to gaining the trust of the researcher such engagement offers the opportunity to acquire the researcher's own thoughts, feelings and expectations as to how information services, policies and technologies may shape the future. Issues, such as who the technology is for, how it fits in with researchers' practices or what the purpose of the technology is, require prior consultation with those with a vested interest in the technology.

¹ Project led by HATII, University of Glasgow - <http://www.data-audit.eu/>

5 Discussion and Conclusion

In this paper we attributed the lack of researcher engagement with OA publication repositories to the fact that the main drivers behind their development were somehow distant from current research needs. This, we argue may be due to the lack of an appropriate iterative requirements analysis involving the main user community.

Research data repositories pose similar challenges. Our experience has shown that using open data as a message to engage researchers in curation activities makes it easy to become detached from current research needs in many disciplines. The heterogeneity of research practices and their datasets, some of which cannot be openly shared provides further evidence of the importance in understanding and appreciating the requirements from the different research communities. Moreover, we believe that the curation of research data requires trusted relationships achieved by working and conversing with researchers early on in their research process. This paper presents approaches from both Edinburgh and Oxford which try to articulate and understand how researchers work with data and information, the barriers they find and their priorities for services required to assist them. We argue that failure to engage with the specific needs of researchers through these initiatives, may lead to the development of data repository services that are under-exploited or indeed may not even be used.

Further work on user engagement in data curation should be pursued to explore connections with other areas such as data citation and the academic reward system, data management tools, business models as well as institutional and funders' policies.

References

1. Harnard, S.: For whom the gate tolls? How and why to free the referred research literature online through author/institution self-archiving now (2001) <http://cogprints.org/1639/>
2. Raym, C.: The case for institutional repositories: a SPARC position paper (2002)
3. Salo, D.: Innkeeper at the Roach Motel. *Library Trends* Vol. 57 No. 2 (2008)
4. Murray-Rust, P.: Open Data in science. In *Nature Proceedings* (2008)
5. OECD: OECD Principles and guidelines for access to research data from public funding. Paris (2007) <http://www.oecd.org/dataoecd/9/61/38500813.pdf>
6. Gibbs, H.: DISC-UK DataShare: State of the art review (2007) <http://www.disc-uk.org/docs/state-of-the-art-review.pdf>
7. Lyon, L. Coles, S., Duke, M. and Koch, T.: Scaling up: towards a federation of crystallography data repositories (2008) <http://eprints.soton.ac.uk/51263/>
8. Berman, H.M.: The protein Data Bank: a historical perspective. *Acta Crystallographica* 88-95 (2008)
9. Benton, D.: Recent Changes in the GenBank online service. *Nucleic Acid Research* Vol. 18 No.6 (1990)
10. Steinhart, G.: DataStaR, a data staging repository for digital research data (2008) <http://www.dcc.ac.uk/events/dcc-2008/programme/posters/DataStaR.pdf>
11. Martinez-Urbe, L.: Findings of the scoping study and research data management workshop (2008) <http://tinyurl.com/55fxgw>
12. Ekmekcioglu, Ç. and Rice, R.: Edinburgh data audit implementation project: Final report (2009) <http://ie-repository.jisc.ac.uk/283/>