Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-3

How Can Libraries and Other Academic Institutions Engage in Making Data Open?

Peter LINDE[1a], Bridgette A WESSELS[b], Thordis SVEINSDOTTIR[b], Merel NOORMAN[c]

[a] *Blekinge Institute of Technology, Sweden*
[b] *University of Sheffield, UK*
[c] *Royal Netherlands Academy of Arts and Sciences (KNAW)*

Abstract. In this paper we will address the questions of what and where the value of open access to research data might be and how libraries and related stakeholders can contribute to achieve the benefits of freely sharing data. In particular, the emphasis will be on how libraries need to acquire the competence for collaboration to train and encourage researchers and library staff to work with open data. The paper is based on the early results of the RECODE project, an EU FP7 project that addresses the drivers and barriers in developing open access to research data in Europe (http://www.recodeproject.eu).

Keywords. Open data, Libraries, Open Access

Introduction

During the last 30 years libraries have adopted to new demands while analogue media turned digital. Librarians have creatively adapted to passing fads, and/or long lived realities such as Archie, Gopher, NCSA Mosaic, FTP, SGML, XLM, Open Access, PDA etc. Today most university libraries have Institutional Repositories and a digital publishing department dedicated to supporting researchers' needs of dissemination, preservation and open access advice. Libraries do have long experience of advocacy, training and implementation of open access of publications and of dealing with digital information but now, when we are finally talking about a tipping point for scholarly Open Access documents[1], a new hot topic with a whole new set of demands on library skills, budgets and organization have arrived – Open data[2].

Open Access (OA) to research data is increasingly regarded as a positive development that should be encouraged and stimulated within the European research landscape. The European Commission is pushing for research data to be more open in its Framework Programme Horizon 2020[3], and the trend is also growing within the

[1] Peter Linde, Blekinge Institute of Technology, 37179 Karlskrona, Sweden. Peter.linde@bth.se

individual member states as well as the academic community. Several influential journals are now encouraging or requiring researchers to make the data that supports their publications freely accessible (for example all the BioMed Central journals, The Open Access Geoscience Data Journal Dataset Papers in Science, eLIFE, F1000Research etc) while national and private funding agencies list open access to research data as a condition for funding. However, achieving open access and realizing its benefits requires considerable work, as the growing literature on data sharing and open access shows.

There now seems to be a more general consensus about the value that open data can bring to science and society. According to its advocates, unrestricted and digitally facilitated access to data would enable faster progress in science through minimising duplication of effort and offering scientists a wider range of data to use for re-analysis, comparison, integration and testing. It would contribute to the quality and integrity of scientific practices, as it increases transparency and accountability. It would also improve the way science and scientific data can be used in relation to social goals, and thus enhance the value of the contribution that science makes to society. Moreover, there is a strong notion that open data will be beneficial to innovation and economic growth. The European Commission, for example, refers to open data as "an engine for innovation, growth and transparent governance[4].

But open access and the re-use of research data have proven to be a challenge in most disciplines. Many repositories, created to encourage data sharing, remain largely empty[5]). Despite the difficulties a few vanguard libraries have felt a need to support researchers in the management and dissemination of research data. We will take a closer look at some of these initiatives, which often started as 'new opportunities' projects aiming to expand library services in a time where classic university library activities like cataloguing, media acquisition, subscription services etc. are questioned or being replaced or automated. The barriers to open research data are many and it is not realistic to believe that one stakeholder can solve all the challenges single-hand. There is a strong need for cooperation inside as well as between organisations, sharing expertise and specialist knowledge.

The central question posed in this paper is: how can libraries handle this new service together with other open data stakeholders in the academic world?

The paper presents a review of policy documents, reports, scholarly literature and other relevant documents to provide an overview of current developments within the field. We provide an analysis of some of these approaches in order to identify good practices and potential barriers.[2]

In the current, very highly, competitive university climate, productivity and quality are buzz words, and increasingly funding for research is based on bibliometrics. In this environment it is becoming more important for university management to keep track of the productivity and quality of the research publications. At the same time more funders are mandating open access and universities are struggling to promote their brand in order to hire the best scientists and attract the brightest students.

[2] This paper is based on findings made in the ongoing work packages of the RECODE project[6].

In this landscape many librarians realize that their services, including repositories, is one of many that have to interconnect in order to support and make research more visible.

Today, university libraries are investigating possibilities of integrating institutional repositories with CRIS systems (Current Research Information System) usually run by university research offices or similar departments[7]. In Sweden this is being investigated on a national level where the national repository portal SwePub will possibly be integrated with the Swedish Research Councils CRIS system[8]. Universities like the University of Edinburgh have integrated all research service into one department (Information Services) which include classical library functions but also have divisions like IT-infrastructures, Digital Curation Center, the Jisc-designated national data centre (EDINA) and the Data Library[9].

In their Roadmap for Research Data the League of European Research Universities listed the library as a main source for data management and discovery[10]. It is evident that an important new role for the library going down the E-science road is to be a competent team player when it comes to build such support structures for researchers. This is best done together with other important players at the university - Research Office Services, Archive staff and Academic IT Services and of course data centre specialists.

The need for training & advocacy

Most researchers and university support staff are new to the task of open data management which implies massive amounts of advocacy and training. In the Opportunities for Data Exchange (ODE) project[11] it is spelled out: “Improving the skills and understanding of researchers in data management is essential. Training should begin in the institutions that train researchers, at the outset of postgraduate study and the latest, possibly even earlier”. It is pointed out repeatedly that discipline-focused education in data management best practice must be incorporated into student and researcher training at an early stage. So in order to play an active part in establishing open data libraries and to build competence for this, cooperation with other university stake holders is important as well as being pro-active in open data management advocacy and training.

One reason why data sharing and open access is still not the norm in most disciplines is due to the reason that researchers are reluctant to make their data public. Their concerns range from work being scooped or misused, to not having enough time or funding to make their data accessible, to maintaining the privacy and confidentiality of their research participants [5]. Researchers may also lack the expertise to share their data[12]. Scientists express a variety of concerns for the “amount of work and the time needed to make data meaningful and useful if made openly available. For instance, the time needed to annotate, create and apply metadata and document context. This extra work would take up time from other research activities such as data collection, analysis, publications and applications for funding, all of which bring clear and demonstrable rewards and benefits to scientists and their careers”[13]. Another key problem is that it requires considerable technical skills to translate data in to machine-readable formats and to use the software tools to access and analyze the data. Researchers that wish to make their data publicly and digitally available and re-usable have to become acquainted with software tools and data formats that might not easily fit their existing research practices. Re-using data, in turn, requires researchers to learn about how to

search and use data through web-based tools. It can also be difficult to find common standards and formats to share data, such that others can easily interpret and use the data. These practical barriers are also reflected in the European Commission's *Online survey on scientific information in the digital age[14]*. About 90% of the respondents in this survey disagreed with the statement: "Generally speaking, there is NO access problem to research data in Europe". Providing training to researchers and technical staff as well as creating awareness about the possibilities and limitations of data sharing will therefore be conductive to making more research data openly accessible in the various disciplines.

Academic institutions have an important role to play in training advocacy. The Commission's survey also included the question how the European Union could best contribute to access and preservation of scientific publications and data. Most respondents agreed strongly with the statements "supporting the development of a European network of repositories" and "encouraging universities/research institutes, libraries and funding bodies etc. to implement specific action"[15]. Since many funding bodies already place responsibility for data management policies and compliance with research institutions, this also increases the pressure on the academies to make data openly available.

Within the whole academic community there is a lack of professional preparation for data management and no one is really taking responsibility for the research data management function. In many ways libraries are in a good position to take on this responsibility but the standard curriculum of library schools do not prepare students for managing data. This has to change.

Different cultures and target groups

In the material reviewed it is a common observation that researchers are a very heterogeneous group. Not only discipline-wise but also between individuals within the same team. Therefore it is important to gain an understanding of the "culture" within any give set of researchers before considering how to influence their research data management behaviour[13].

Research data is different from publications. It is more diverse and often linked to project communities which calls for new ways of working, thinking and cooperating for librarians. Data diversity, tools and researcher needs should not be measured at the disciplinary level but at the research group level.

It is recommended that for advocacy and training purposes interviews, case studies and surveys are developed to understand researcher requirements and behaviour[16, 17, 20, 21, 23]. This must be the basis for developing advocacy/training materials that will motivate researchers, as well as making them understand the obligations to institutions, funders and the public. Preparing data management plans and training staff to accomplish them is new and mostly unchartered waters for universities and research institutions but there are some good examples of and reports on how to support these institutions in open data management.

Mark L. Brown and Wendy White tell the story of how University of Southampton through collaboration with UK Research Data Service and involvement in projects like the Institutional Data Management Blueprint Project (IDMB) started to improve and

formalize initiatives to support researchers at the university in managing their research data[18].

For training purposes, the use of automated- and web tools was set up. For example automated tools to support minting of DataCite DOIs and web based guidance to help interpret funders' requirements.

For data management planning service for researchers a training program was developed to engage with various groups from postgraduate researchers to senior scientists. Planning and realization of these courses, lectures, workshops and seminars were always done together with the researchers themselves.

In a consultancy report made for Jisc[19], the roles, rights, responsibilities and relationships of institutions, data centers and other stakeholders who work with data were explored. The conclusions regarding advocacy and training are very similar to the conclusions from Southampton: The importance to target and tailor measures to specific disciplines and sub-disciplines; Awareness of data curation and preservation good practice is generally low but it varies a lot between disciplines; Recommendations to data center and institutional repository staff to go out and promote their training programs with a mix of methods, seminars, workshops, lessons etc.

As reported in most of the literature an important target group for open data management advocacy and training are young scientists and students at master level and onwards. A first focus of advocacy should be on the postgraduate and the graduate student community since they are in the front line as data collectors and generators, and of course as future researchers[20].

Bottom up or top down?

The typical American data curation program is "devoid of top-level mandates and incentives, but rich with independent "bottom-up" action". A structure like this is based on enterprising individuals and makes for a slow speed of development[21]. In a recent American survey with the aim to identify current trends in research data management at research institutions only 9% of the respondents answered yes to the question "Does your institution have a DM policy"? Close to 90% agreed with the follow up statement "An institution-wide DM policy is important" which shows that university stakeholders like researchers, librarians, office of research staff, teachers etc. are keen to see such policies implemented[22].

The reason libraries have started data curation programmes at all is due to their vanguard position relating to open access publications repositories and the digital preservation initiatives early explored by university libraries. This is also said to give the library opportunity to leverage existing partnerships and engage in new ones to build skills and necessary alliances for data curation. Engaging with a few research communities as a start up pilot is a way to gain acceptance, formalization and getting program commitments from administrative levels. A successful project might well be a way of convincing university administrators of the benefits of a university wide curation policy and mandate[21].

In Southampton [19] the response to the insight that funders increasingly placed responsibility for data management policies and compliance with research institutions resulted in a bottom-up approach based on researchers needs and an incentive to design requirements for an institutional top-down approach policy and infrastructure. Their

experience with open access publishing repositories was that "researchers were open to new practice as long as it was researcher led, integrated into research workflow, reflective of discipline distinctions and supported by advice and training. Clarity over policy and responsive service support were essential".

It was very important that the institutions at the university felt that they were in command of the investments and service support regarding data management without feeling compelled by a set of requirements.

In this process the resulting data management policy was putting the responsibility for recording, maintenance, storage and security etc. and the compliance with relevant regulations on the researchers which is good news from a library viewpoint. It is sensible that the creators of the data also record it and that the library is there as a supporter of the process instead as an accountable enforcer.

Of the key components in the Southampton project, an institutional policy framework, a working institutional data registry, a one stop shop for data management advice and guidance and a sustainable business model it is the university policy on research data management that is considered the most important. In the end and because of the power balance there is a need for a formal mandate or policy from a higher university authority[18].

Librarians introduced and administer the institutional repository and the idea about open access with a great knowledge about scholarly communication issues but since they do not bring any funding into the university the library is mostly perceived as a service based unit without much influence. But in the meantime, and as a first step to a formal policy, when there is no clear guidance from government authorities and university administrations are withholding resources or initiatives on data management issues, the bottom up approach is a way to start where advocacy is the first step only.

New roles and partners

University of Southampton is one of many examples of how initiatives for data curation projects do not stop with collaboration inside the university departments. Many times necessary skills are only available through partnering with outside institutions or organizations[21, 22]

No matter how libraries approach the challenge of data curation an introduction of new skills in the library profession is sorely needed. Working in partnership with scientists' future job roles as "data librarians" must contain skills both on the technical side and the archival side of the data coin. Specialists like this will play a key role in the scholarly publication process and must be rewarded accordingly. Library schools need to introduce courses that fit these new job descriptions.

There is absolutely a need for convergence between library and archival skills in order to make university repositories a well functioning place for open data. This could also be a part of professional development and training[18]. This is also true for library professionals vis-à-vis research office professionals who are close to researchers supporting them with project applications, statistics etc. There might also be a chance for classic library roles such as liaison librarians to expand. Liaisons can help researchers depositing their data at the point of data creation. They can advice about standards applicable to the needs, create curation plans to the whole life cycle of the data in full compliance with funder mandates[23].

As stated earlier the skill levels of researchers regarding data management are variable and training is much needed. So parallel to advocacy there is a requirement for development of community skills. But since most of the expertise in data management is concentrated in data centers, there is a need to engage and formalize a flow of knowledge from data centers to institutions where staff now increasingly are being appointed to manage and develop repositories for data curation.

Since 1976, CESSDA (Consortium of European social science data archives) has served as an informal umbrella organisation for the European national data archives. The CESSDA data archives and other similar subject data archives are in a good position to work with universities libraries and negotiate with archives on training.

Sometimes there is a polarization of views regarding the role of institutional repositories for data. Data centers and data archives have a more long-term perspective than the institutional repositories, which are relatively new structures yet to prove their ability. But both data centers and libraries have a stewardship role in data curation activities. They both help and guide researchers depositing their data. Dividing the different roles on short-term, easily accessible storage taken care of by institutional repositories and long-term preservation by data centers could be one way to facilitate for better data management support and cooperation[18].

Conclusions and discussion

Underlying issue of the new roles for the libraries in open data management is of course the question about funding the new services. There is obviously a need for the university to make economic plans for the costs of storage, curation, training etc. for research data.

It can be a major problem to convince university administration to gather economic resources for developing data curation models. In fact most of the scarce funding for research data management is coming from libraries themselves[22]. Usually there is no extra seed money available inside the organization and libraries either have to reallocate internal resources or find external funding, e.g. cooperation with outside partners. Therefore the initiation of grants and funding for libraries on national or international levels will be an important factor for getting data curation to gain speed on a broader level at universities[21].

There will probably be no real increase in funding without institutional or national mandates implementing research data management plans. Bottom up practices are slow generators of change and general acceptance and will therefore have to be complemented with formal policies.

Among the major academic stakeholders in the open data eco system we have the funders of science – the councils and foundations; the creators of data – the researchers and we have the disseminators and curators of data – in this case the libraries, archives and the data centres. All these stakeholders with their organizations will need to cooperate, as the barriers are multiple and complex, that only joint forces can realize the idea of open data. Funders and policy makers need to clearly mandate data management and also earmark funds for training, infrastructure, data curation projects etc. Professional associations have to reflect on instigating new opportunities for

training of professionals. Librarians, IT-specialists and research office staff from the universities need to collaborate with archivists and curators from data centres and vice versa. Researchers need to find new priorities regarding the importance of data management, need to find ways to make data management pay career wise.

All this cooperation is already going on but it will have to spread and it has to be fuelled by governmental and academic authorities that issues policies that can facilitate cooperation and clear roadmaps for the way forward. Equally important are the non-governmental advocacy groups and other cross-professional organizations that have taken an interest in pushing the question of open data forward. Organizations like COAR, EUDAT, LIBER[24], RDA, SHERPA, SPARC, KE and many more are doing a fantastic job of advocating and informing about the importance of open data management and they are a giant resource for libraries that are about to start data curation schemes.

There is a current gap of technical knowledge and access to proper infrastructure but there is also among the libraries and librarians a lack of understanding of the complexity of the process of managing open data Using the experiences from the case studies performed in the RECODE project so far, we argue that the value of unrestricted access to research data depends significantly on the quality of the OA process. Our analysis of the values and motivations amongst researchers regarding OA showed that approaches to support and improve the development of open access to research data need to address at least the following issues:

- They should be sensitive to the different scientific practices to ensure that existing research rigour is maintained as well as facilitating OA.
- They should make the link between infrastructures, legal and ethical issues, and institutional frameworks, so that the OA ecosystem can support an appropriate approach to all types of data within their research areas.
- They need to provide safeguards for anonymity and privacy of research participants.
- They should provide ways to reference and attribute all open data correctly as part of ethical research practice.
- They need to pay attention to technological issues; such as the way technology drives the collection of vast datasets, the lack of technical infrastructure to store data and interoperability issues.
- Cultural barriers are significant, especially issues such as competition within science for reward and reputation, the lack of trust between scientists and the lack of career related rewards and prestige resulting from publishing and sharing data.

It is vital for libraries to realize that now is the time to be proactive regarding research data management – introducing professional preparation programs, starting up pilot programs, monitoring major data initiatives like DataCite, DataONE etc. and good examples of library initiatives like University of Edinburgh[9], University of York[25] University of Southampton or Purdue distributed data curation center[26] or else risk being bypassed by other players in the arena of establishing research data management programs. The role of libraries in data management training is not evident for everyone. Some researchers agree that libraries should have and increasingly important role as

data managers and experts based on their role in open access article publishing. Others argue that data centers could provide the support needed to handle the data correctly [11]. It is high time to start to reflect on these issues and to start studying experiences made so far in the urgent task of making research data openly available. If the library does not see the potential in the task of pioneering open research data, as it have in advocating open access to research publications, there is a major risk that other stakeholders quickly will fill that role and expand services visavi researchers and librarians will be left with the question, of how libraries can engage in making data open, unanswered.

References

[1] Archambault, Eric et al. Proportion of Open Access peer-Reviewed Papers at the European and World levels – 20014-2011. August 2013. Produced for the European Commission DG Research & Innovation by Science-Metrix Inc.

[2] By ”open data” we refer to research data defined as any material used as a foundation for research.

[3] Press releases database. Commission launches pilot to open up publcicly funded research data. 2013. http://europa.eu/rapid/press-release_IP-13-1257_en.htm. Visited 140131.

[4] 1 European Commission (2011). Open Data, an engine for innovation, growth and transparent governance, COM 882 final, Brussels, 12 December 2011. Retrieved from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF

[5] Nelson, Bryn. Data sharing: Empty archives. *Nature* 461, 160-163, 2009.

[6] Policy RECommendations for Open access to research Data in Europe. http://recodeproject.eu/

[7] Joint, Nicholas. Current research information systems, open access repositories and libraries. *Library Review* Vol. 57:8, 2008

[8] System för analys av svensk forskning. http://www.mynewsdesk.com/se/kungliga_biblioteket/pressreleases/system-foer-analys-av-svensk-forskning-947591. Visited 140125.

[9] Rice, Robin et al. Implementing the Research Data Management Policy: University of Edinburgh Roadmap. *International Journal of Digital Curation* Vol. 8:2, 2013.

[10] LERU roadmap for Research Data. League of European Research Universities, 2013. http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf

[11] Dallmeier-Tiessen S, et al. (2012). Compilation of Results on Drivers and Barriers and New Opportunities. Retrieved from [http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/08/ODE-CompilationResultsDriversBarriersNewOpportunities1.pdf].

[12] Borgman, Christine L. The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology*, Vol 63:6, 2012.

[13] Sveinsdottir, Thordis et al. Deliverable D1: Stakeholder Values and Ecosystems. Policy RECommednations for Open access to research Data in Europe (RECODE), 30 september 2013. http://recodeproject.eu/wp-content/uploads/2013/10/RECODE_D1-Stakeholder-values-and-ecosystems_Sept2013.pdf

[14] European Commission (2012). Online survey on scientific information in the digital age, 2012. ISBN: 978-92-79-23170-4. DOI:10.2777/7549

[15] Ibid.

[16] Lyon, Liz et al. Final report – disciplinary Approaches to Sharing, Curation, Reuse and Preservation. Jisc 2009. http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf

[17] Schmidt, Lisa, Ghering, Cynthia; Nicholson, Shawn. Digital Curation Planning at Michigan State University. *Notes on Operations* 55(2), 2011. http://staff.lib.msu.edu/nicho147/Research/DigCur_LRTS_2011.pdf

[18] Pryor, Graham and Sarah Jones and Angus Whyte. *Delivering Research Data Managment Services: Fundamentals of good practice*. Facet Publishing, 2013.

[19] Lyon, Liz. Dealing with Data: Roles, Rights, Responsibilities and Relationships – Consultancy Report, 2007.

[20] Carlson, Jake R. and Bracke, Marianne S., "Data Management and Sharing from the Perspective of Graduate Students: An Examination of Culture and Practice at the Water Quality Field Station" (2013). *Libraries Faculty and Staff Scholarship and Research.* Paper 53.
[21] Walters, Tyler. Data curation program Development in U.S. Universitites: The Georgia Institute of Technology Example. *The International Journal of Digital Curation*, Vol. 4:3, 2009.
[22] *Research Data Management* – Principles, practices, and prospects. Council on Library and Information Resources. 2013. ISBN 978-1-932326-47-5. http://www.clir.org/pubs/reports/pub160
[23] Gabridge, T. The last mile: Liason roles in curating science and engineering research data. Research Library. Issues: A bimonthly report from ARL CNL and SPARC August 2009. http://old.arl.org/bm-doc/rli 265 gabridge.pdf
[24] Chrisensen-Dalsgaard et al. Ten recommendations for libraries to get started with research data management. Final report of the LIBER working group on E-Science/Research Data Management, 2012.
[25] Research data management at the university of York. http://www.york.ac.uk/about/departments/support-and-admin/information-directorate/strategy/projects/rdm/
[26] Distributed data curation center, D2C2. Purdue University Libraries. http://d2c2.lib.purdue.edu/ Purdue University