

# **Collaboration to Data Curation: Harnessing Institutional Expertise**

Stuart Macdonald

*EDINA Data Library, University of Edinburgh, Edinburgh, Scotland*

160 Causewayside, Edinburgh EH9 1PR, Scotland, United Kingdom

Luis Martinez-Uribe

*Oxford University Computing Services and e-Research Centre, University of Oxford, Oxford, United Kingdom*

13 Banbury Road, Oxford OX2 6NN England

*(Received 19 March 2010; final version received)*

## **Abstract**

It can be argued that institutional repositories have not had the impact (Lynch 2003, Salo 2008), initially expected, on academic scholarly communications (the exception being in a few well-developed and successful instances). So why should data repositories expect to fare any better?

Firstly, data repositories can learn from publication repositories experiences and their efforts to engage researchers to accept and use these new institutional services. Secondly, they provide a technical infrastructure for storing and sharing data with the potential for providing access to complimentary research support facilities. Finally, due to the interdisciplinary expertise required to develop and maintain such systems, stronger ties will be forged between libraries, information and computing services, and researchers which will assist innovation and help to make them sustainable and embedded within academic institutional policy.

This paper, whilst aware of the diverse nature of institutional and departmental practices, aims to highlight a number of initiatives that will show how research data repository infrastructures can be effectively realised through collaboration and sharing of expertise. We argue that by employing agile community, strategic and policy judgment a robust data repository infrastructure will be part of an integrated solution to effectively manage institutional research data assets.

Keywords: data curation; research data management; data repositories, knowledge transfer and exchange

## **1. Introduction**

The recent Council for Science and Technology report ‘A Vision for UK Research’ (2010) placed emphasis on two-linked processes, namely: “focusing on excellence across the research base and second, harvesting the products of the research base” in order to remain competitive with emerging science-based economies (e.g. India, China) and to maintain the UK’s leading position in the global research marketplace. In order to sustain this position the report emphasises the need for the development of new collaborative models. These could manifest themselves as collaboration surrounding facilities where cost may be a factor, collaboration where sheer scale of effort needed can deliver both breadth and economies of scale not possible for each singular participant, collaboration at the local level which pools both resources and expertise.

Integral to the whole research base are research outputs such as publications and digital data as both evidence and the means to verify intellectual endeavour. University strategies to harvest these products have developed around the concept of digital repositories developed by the academic libraries. The first realisation of such information systems were publication repositories built to manage and disseminate research articles and aiming to provide open access to a significant proportion of newly published academic papers. The development of research data repositories has been seen as the next coherent step in the growth of repositories (Heery & Powell 2006). Nonetheless, it can be argued that institutional repositories have not had the impact, initially expected (Lynch 2003, Salo 2008), on academic scholarly communications (the exception being in a few well-developed and successful instances). So why should data repositories expect to fare any better?

Firstly, the data repository activity can learn from publication repositories experiences (Macdonald & Martinez-Urbe 2009) and their efforts to engage researchers to accept and use these new institutional services. Secondly, data repositories provide not only a technical infrastructure for storing, sharing and managing data but also access to complimentary research support facilities such as data management training and data auditing tools, and innovative utilities such as dataset citation and linking tools, and accessories to visualise and analyse heterogeneous content. Finally, due to the interdisciplinary expertise required to develop and maintain such organisational and technical systems, stronger ties will be forged between libraries, information and computing services, and researchers helping to make them sustainable and embedded within academic institutional policy.

This paper, whilst aware of the diverse nature of institutional and departmental research cultures and practices, aims to highlight a number of initiatives that will show how research data management and repository infrastructures can be effectively realised through collaborative efforts and the sharing of expertise and as such demand a prominent place on the academic research landscape by providing systematic and trusted curatorial and archival services, engaging interfaces that encourage re-use of content, in addition to addressing funder and institutional requirements regarding research data management mandates.

## **2. Strategies to harvest the products of the research base**

The following sections present activities at the University of Oxford and the University of Edinburgh that build on lessons learned from publication repositories and make use of a different set of strategies to deal with research data. These

strategies require multidisciplinary skills in areas such as information management, computing, economics, institutional governance and social dynamics supplied by such actors as departmental heads, librarians and computing staff, principal investigators, records managers, archivists and research office staff. The alignment of specialists from the aforementioned backgrounds is an important step on the route to a cohesive infrastructure to support researchers in the creation and use of data whilst ensuring the appropriate harvesting of the products of the research base.

### ***2.1 Data repository activity at Oxford***

A research data management scoping study (Martinez-Urbe 2008) directed by the Oxford Digital Repositories Steering Group throughout 2008 revealed that University staff from a range of disciplines and central departments face a variety of challenges relating to the creation and management of data. This comes at a time when research councils are increasingly developing policies that require certain levels of data sharing and curation (Jones 2009). Such policies are an important and welcome step towards a new scholarly communication landscape but in some cases they can be dislocated from the research labs or other environments where research takes place as echoed by the RIN Disciplinary Case Studies in the Life Science Project (2009) which found that data and information sharing activities are mainly driven by needs and benefits perceived as most important by practitioners rather than ‘top-down’ policies and strategies.

Central services at Oxford including computing services, library and research services together with academic departments are looking at ways of streamlining these issues for their researchers. There is an urgent need for establishing coherent institutional frameworks that support the creation, curation and reuse of data whilst

addressing research council policies and understanding the value and cost of data management and curation activities.

Since the initial study, a complete research data programme has commenced at Oxford with a range of data repository activities that are working across subject domains as well as institutional service providers and themes in order to develop a robust collaborative data repository infrastructure to support researchers with their data.

One such activity is the Embedding Institutional Data Curation Services in Research (EIDCSR) project<sup>1</sup>. Through EIDCSR central departments are working with three collaborating research groups in Life Sciences and Medicine to scope and address their data management requirements. These groups collaborate as part of a nationally-funded project and conduct research into ventricular tissue architecture, combining traditional histological and novel imaging techniques like Magnetic Resonance Imaging (MRI) and Diffusion Tensor MRI as well as with image processing and computational models for bio-mathematical simulation. By bringing together this sophisticated range of techniques and areas of expertise the collaborative groups are generating hundreds of Gigabytes of data which the funder requires them to store and made available for ten years after the completion of the project.

The EIDCSR project is also investigating institutional policy and guidance for the management of research data and records. The Research Services Office is leading this work in collaboration with the University of Melbourne following their Data Management Policy<sup>2</sup> as exemplar. The approach taken attempts to transform funders' policies into something that clarifies the responsibilities of both department and

researcher whilst pointing them to existing services and other useful information and resources. Through the experience in Melbourne and the work in Oxford it has been realised that in order to develop successful policy and guidance for research records it is crucial to understand the role such policies can play at the both university and academic department level. Moreover, policy and guidance ought to be useful, in practical terms to researchers and therefore they must be involved in their development for it is vital that the implementation of institutional policy penetrates at local or departmental level.

The EIDCSR project is also investigating the economics of research data management. Key questions in data repository development include how much it cost to manage data and who will pay for it (Blue Ribbon Task Force 2010). By participating in the KRDS2 Project<sup>3</sup> with a range of data centres, departments and institutions, detailed information has been gathered the costs required to create, manage and curate the data created by the research groups participating in EIDCSR. The results shown in the diagram below highlight the high costs of creating the specific datasets by the research groups in question. The second biggest cost, start-up curation, covers the curatorial activities undertaken as part of the EIDCSR project this includes metadata management and technical developments. These curatorial costs are expected to be lower if provided by an established institutional service. As an example of this, the back up and long-term filestore service provided by Computing Services ensures the copies of the data are kept safe for five years with a minimal cost in comparison to creation and start-up curation.

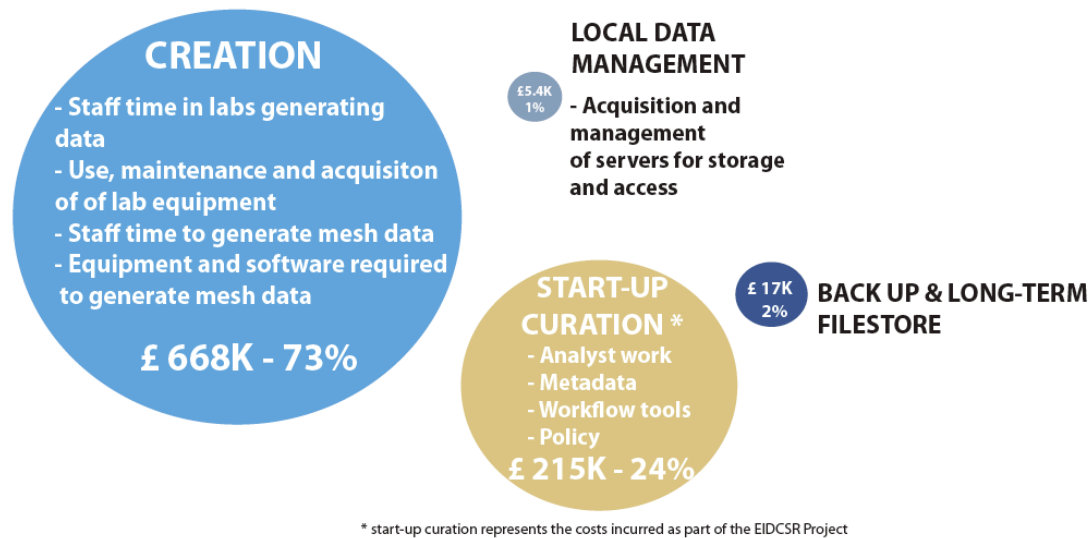


Figure 1. Data management and curation costs from the Oxford survey

Figure 2 presents the previous cost represented in time with a data lifecycle of eight years. The costs are concentrated in the first years when the data are created and reduce significantly as they progress through their lifecycle.

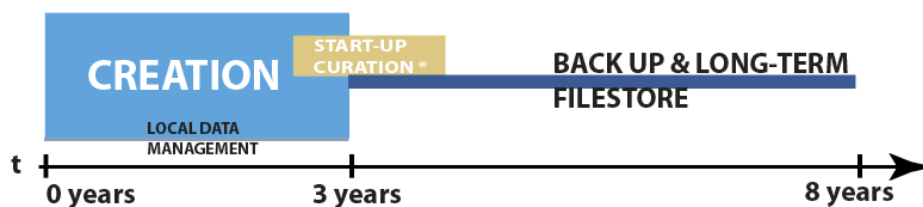


Figure 2. Data management activities placed in time.

Other data management activities that have recently started at Oxford include the Sudamih and the Admiral projects. Sudamih is an EIDCSR sister project and it shares the institutional and procedural frameworks developed through EIDCSR to scope and address the requirements of scholars in the Humanities Division. The project will pilot the provision of a database as a service to enable the creation and publication of datasets including images, text or geo-referenced data. It has also a strong emphasis in training and will gather requirements about data management training needs to then develop and pilot training modules based around existing

training resources such as DCC 101<sup>4</sup>. The Admiral project led by the Image Bioinformatics Research Group (IBRG) in Zoology is working with life science researchers to assist them with tools to locally store and annotate their data to then package them for archiving and submission into a central repository for preservation provided by the Library. Previous data curation activities from IBRG has provided extremely helpful concepts such as sheer curation or proved the usefulness of data enriched articles (Shotton et al. 2009).

Finally, Oxford is taking part in the UK Research Data Service (UKRDS) that is working towards a national infrastructure for research data management. The initial feasibility study concluded that the best approach should embed capacity, skills and training (Beagrie et al. 2009). Currently UKRDS is planning the Pathfinder implementation phase in which Oxford will play a key role with Leeds, Bristol and Leicester.

## ***2.2 Research data management stakeholders and initiatives at Edinburgh***

The DISC-UK DataShare project<sup>6</sup> (Mar 2007 – Mar 2009) sought to develop models for multi-disciplinary institutional data repositories in the UK higher education sector. It was led by EDINA and Edinburgh University Data Library in partnership with the University of Oxford and the University of Southampton. It concluded that (Rice 2009):

- Data management motivation is a better bottom-up driver for researchers than data sharing but is not sufficient to create culture change
- Data librarians, data managers and data scientists can help bridge communication between repository managers & researchers
- Institutional repositories can improve impact of sharing data over the internet.



From a local perspective the project was instrumental in developing the Edinburgh DataShare research data repository, hosted by the Data Library and contributor to JISC RepositoryNet, an interoperable network of repositories which provides UK tertiary education with access to trusted and expert information about repositories. Edinburgh DataShare shares the same DSpace software platform with the Library's Research Publication Service (comprised of the Publications Repository, a closed repository for use only in the University of Edinburgh, and the Edinburgh Research Archive which is a public open access repository) which was launched in January 2010 by Edinburgh University Library to support the implementation of the University's Open Access Publications Policy<sup>7</sup>. As such it retains the potential to interoperate and link supplementary research data produced by local researchers to corresponding publications

As a direct result of funder council requirements regarding the management and sharing of research data after the research project has been completed Edinburgh DataShare has been approached by a number of local researchers who wish to have permanent location for their completed and documented dataset(s), with an open access metadata record. Such an engagement opportunity facilitates the development of the data repository by scoping functional requirements regarding value-added visualisation and analytic tools such as multi-media viewers, licences, domain specific metadata schemes and file formats, federated access, links to remote storage, and semantification of content

The Data Library also led one of the pilot demonstrator projects of the HATII/DDC led-Data Audit Framework (DAF)<sup>8</sup> which conducting audits of departmental data collections thus engaging with the local research community in

data management practices. The primary recommendation by the Edinburgh DAF Steering Group was for the adoption of an institutional-wide research data management policy. Other key findings (Ekmekcioglu & Rice 2009) indicated that staff require practical and systematic guidance on research data management be it from research unit or school procedures, college or university-wide infrastructure and policy, or identifiable forms of support in the form of expert support staff, web pages, and discipline-specific guidelines, as well as short, focused, training opportunities. This resonates with the Oxford Scoping Study, and the RIN Disciplinary Case Studies in the Life Science Project conducted by researchers at the University of Edinburgh which indicated a lack of coherency when it came to research data guidance and training indicating local and ad hoc mechanisms were in place which reflected both scientific laboratory culture and working practices.

Using DAF as the engagement vehicle the Data Library are currently scoping generic data management training based on research data management guidance materials<sup>9</sup> developed by the Data Library and the Research Computing Service for early stage researchers in conjunction with the Postgraduate Transferable Skills Unit and the Researcher Development Programme. In November 2009 a training course was piloted in for PhD students in the School of Geosciences as part of the Postgraduate Research Students Training Programme. Initial feedback suggested that such courses should seek to strike a balance between discipline-specific and generic content

In spring 2010 a review commences at the University of Edinburgh to address the issue of managing the rapidly expanding volume and complexity of data produced

by Edinburgh researchers. Concern is both for the shorter term – ensuring competitive advantage through secure and easy-to-use access; and for the longer term – ensuring long-term access and usability for the research community into the future.

The Review is overseen by the IT Committee and the Library Committee, and will have twin tracks to look at Research Data Storage and Data Management, Curation and Preservation. The Review will look at current practice in the University of Edinburgh, review what is known about current practice in peer universities and internationally, and develop options which will include costs, feasibility, and a risk analysis of actions or inactions in this field.

Crucial to its success is the development of partnerships with researchers, heads of department, principal investigators, who have to be convinced that any research data management solution must allay fears about issues such as privacy, loss of ownership, fear of misuse, personal investment, IPR uncertainties whilst offering tangible benefits such as providing reliable access to researchers' own data, a suitable environment to adhere to funders' mandate, metadata that can increase the exposure of individual's research within the research community; and the devolution of preservation from the individual to the institution. Of equal importance are those with supporting roles such as librarians, computing services, school administrators, records managers, archivists and research office staff who have a stake in using, preserving, re-using digital data output as part of the research process.

There are a number of other services and initiatives who have a role to play on the Edinburgh repository stage, including:

- The Digital Curation Centre which enters its third phase is aiming to build strong foundations for good data curation practice across the HE sector by providing support to data custodians, with a specially devised DCC training programme aimed at encouraging the transfer of knowledge and best practice first among data custodians, then between data producers and users.
- The Edinburgh Compute Data Facility (ECDF) Storage Area Network (SAN) provides large scale storage thus offering a potential solution for hosting very large datasets through interoperation with a data repository service such as Edinburgh DataShare which could store the corresponding metadata record(s).
- The Enhancing Repository Infrastructure in Scotland (ERIS) Project , led by the University of Edinburgh, whose aim is to work in collaboration with Scottish researchers and their institutions' repository managers to motivate researchers to deposit their work in repositories and facilitate the integration of repositories in research and institutional processes. ERIS also intends to engage with research pools<sup>10</sup> to create 'virtual repositories' that represent aggregations of research outputs as collected from their participating members institutional repositories, ensuring that the practical requirements of these repositories as stated by the research pools are met. The University of Edinburgh plays a key role in a number of the Scottish Research Pools.

The Data Library and ERIS are currently appraising the potential use of Edinburgh DataShare as the type of mechanism for storing and providing shared access to research data generated by the devolved researchers who comprise the cross-institutional research pools. They also, along with colleagues from EDINA, the DCC, and the Library form the organizing committee for the Repository Fringe Conference

(held in Edinburgh since 2008) which acts as a forum for repository developers, managers, researchers, administrators, and open access practitioners to discuss and share developments in the repository world.

### **3. Conclusion**

Initially data repositories can be thought of as the technical infrastructure to deal with the creation, storage, management and curation of research data. Nonetheless, current research and practice shows that in order for a data repository to be successful, it is required to develop not only the technical infrastructure but a whole range of other institutional services. The development and implementation of these data repository services needs to be an orchestrated exercise involving a wide group of institutional actors including academics, computing services, libraries and research services.

Although this paper focuses its attention on the non-technical aspects of data repositories it is important to highlight that there are key developments on this area that promise to revolutionise the way the data is collected and analysed. With the almost exponential growth of research data output and the absence of off-the-shelf data management solutions there are those in the e-science community who are proposing 'taking the computing to the data' i.e. collaboration between the domain specialist and the computing specialist at the design phase of a data system to come up with a common language to describe those terms used in the scientific and computing domains. Jim Gray (Szalay & Blakeley 2009) called this 'Data Intensive Scalable Computing' in his informal rules that codify how to approach data engineering challenges related to large-scale scientific datasets. A more simplistic approach utilising 'Gray's Laws' in terms of scalability and connecting to the scientists

may well offer potential lessons that can be applied to research data management in the academic sphere.

In sum, the activities discussed in this paper show that insightful strategic and policy judgment can help to develop robust institutional data repository frameworks enabling a move towards seamlessness in terms of the professionals working or interacting with research data throughout the lifecycle. This consolidation of expertise within research institutions via intra- and inter-institutional and cross-facility services could be seen as part of a research data management solution in this time of the economic constraints that are impacting on academic institutions.

1. Embedding Institutional Data Curation Services in Research (EIDCSR) <http://www.eidcsr.oucs.ox.ac.uk>
2. University of Melbourne Data Management Policy <http://www.unimelb.edu.au/records/research.html>
3. Keeping Research Data Safe 2 <http://www.beagrie.com/jisc.php>
4. DCC 101 course <http://www.dcc.ac.uk/events/workshops/dcc-digital-curation-101-workshop>
5. Sheer Curation is defined at [http://en.wikipedia.org/wiki/Digital\\_curation#Sheer\\_curation](http://en.wikipedia.org/wiki/Digital_curation#Sheer_curation)
6. DataShare <http://www.disc-uk.org/datashare.html>
7. University's Open Access Publications Policy <http://tiny.cc/5845v>
8. Data Audit Framework (DAF) - <http://www.data-audit.eu/> provides organisations with the means to identify, locate, describe and assess how they are managing their research data assets via online tools and methodologies.
9. Data Library data management training materials <http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt>
10. Research Pooling is defined as the formation of strategic collaborations between universities in disciplinary or multi-disciplinary areas involving the international quality departments or individual researchers across Scotland - <http://sligachan.lib.ed.ac.uk/wordpress-mu/themes/research-pools/>

Short biographical notes on all contributors

**Stuart Macdonald**

Stuart is Associate Data Librarian at EDINA and Data Library, University of Edinburgh, promoting and developing the Data Library service for university researchers. He is part of the Edinburgh DataShare team developing the DSpace

repository for University of Edinburgh research data, and leads the development of the research data-oriented EDINA agencensus service. He is project manager for the JISC-funded AddressingHistory project and recently he was project officer for the JISC-funded DataShare project and researcher for the RIN-funded Life Sciences Case Studies Project.

Luis Martinez-Uribe

Luis is currently Data Librarian at the Centre for the Advanced Study of the Social Sciences (CEACS) from the Instituto Juan March in Madrid where he is developing a Data Library Service. At present he is also working as Data Management and Curation Consultant for the University of Oxford Computing Services in several data management projects (EIDCSR, Sudamih and UKRDS). Prior to this, Luis was Project Manager and Analyst at the Oxford e-Research Centre working in multidisciplinary R&D activities involving digital repositories, researchers, libraries and computing services. From 2001 to 2007 Luis worked as the Data Librarian at the British Library of Political and Economic Science supporting researchers at the London School of Economics (LSE).

## References

- Council for Science and Technology "A Vision for UK Research" (2010)  
<http://www.cst.gov.uk/reports/files/vision-report.pdf>
- Beagrie, N., Beagrie, R. and Rowlands, I. "Research Data Preservation and Access: The Views of Researchers" *Ariadne* 60 (2009)
- Blue Ribbon Task Force "Sustainable economics for a digital planet: ensuring long-term access to digital information" (2010)
- Ekmekcioglu, Ç and Rice, R "Edinburgh Data Audit Implementation Project: Final Report." (2009)  
<http://ie-repository.jisc.ac.uk/283/>
- Heery, R. & Powell, A. "Digital repositories roadmap: looking forward" (2006)
- Jones, S. "A report on the range of policies required for and related to digital curation" (2009)  
[http://www.dcc.ac.uk/sites/default/files/documents/reports/DCC\\_Curation\\_Policies\\_Report.pdf](http://www.dcc.ac.uk/sites/default/files/documents/reports/DCC_Curation_Policies_Report.pdf)
- Lynch, C. A. "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age" *ARL* 226 (2003): 1-7.  
[http://muse.jhu.edu/login?uri=/journals/portal\\_libraries\\_and\\_the\\_academy/v003/3.2lynch.pdf](http://muse.jhu.edu/login?uri=/journals/portal_libraries_and_the_academy/v003/3.2lynch.pdf)

- Macdonald, S. and Martinez-Urbe, L. "User Engagement in Research Data Curation". *Lecture Notes in Computer Science - Research in Advanced Technology for Digital Libraries*, 57.14 (2009)  
<http://www.springerlink.com/content/7mnq13x34717p483/>
- Martinez-Urbe, L. "Findings of the Scoping Study and Research Data Management Workshop" (2008)  
<http://ora.ouls.ox.ac.uk/objects/uuid:4e2b7e64-d941-4237-a17f-659fe8a12eb5>
- RIN "Patterns of information use and exchange: case studies of researchers in the life sciences" (2009)  
<http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/patterns-information-use-and-exchange-case-studie>
- Rice, R. "DISC-UK DataShare Project: Final Report". Edinburgh: University of Edinburgh, (2009)  
<http://ie-repository.jisc.ac.uk/336/>
- Salo, D. "Innkeeper at the Roach Motel." *Library Trends* 57.2 (2008)  
<http://minds.wisconsin.edu/handle/1793/22088>
- Shotton, D., Portwin, K., Klyne, G. and Miles, A. "Adventures in semantic publishing: exemplar semantic enhancement of a research article" *PLoS Computational Biology* 5.4 (2009)
- Szalay, A. S. and Blakeley, J. A. "Gray's laws: database-centric computing in science" *The Fourth Paradigm - Data Intensive Scientific Discovery*. Ed. T. Hey, S. Tansley, and K. Tolle. *Microsoft Research*, (2009)