Building an institutional research data management infrastructure

Sally Rumsey, The Bodleian Libraries, University of Oxford

The University of Oxford is developing an infrastructure for the management, storage and delivery of research data. This paper describes the developments at Oxford and considers the pressing need to implement such an infrastructure. The role and design of DataFinder, a catalogue for research data, is discussed.

Building an infrastructure for research data

Over the last four years there have been a number of initiatives to explore methods for dealing with research data at the University of Oxford. It began with the first stirrings of development of a data repository, DataBank, back in 2009 as part of the DataShare[1] project, followed by other JISC funded projects Admiral[2], EIDCSR[3], SUDAMIH[4], and most recently, the two UMF (University Modernisation Fund) supported projects, Dataflow[5] and ViDaaS[6]. These developments have been predominantly technical. The SUDAMIH project initiated work on non-technical elements, such as training and provision of online guidance[7]. A university-wide policy for data management is in preparation.

The JISC-funded Damaro (Data Management Roll-out for Oxford) project[8] runs until March 2013. The aim of the project is to pull together the existing developments and add new technical and support services to create a pan-university infrastructure. The project comprises four key strands: technical facilities and services, policies, training, and preparation for long-term sustainability. The technical design is a federated, modular structure: there will not be a single, monolithic, one-size-fits-all data repository for the University. Central to the technical provision is the creation of DataFinder, a semantically aware research data catalogue for registering Oxford research datasets and for easy data discovery. As well as incorporating the existing DataBank, the Bodleian Libraries' archival standard research data repository, the technical infrastructure will incorporate DataStage deposit facility (developed by the DataFlow project) and deposit from ViDaaS. The project is also collaborating with Colwiz[9] to provide an exemplar of metadata harvesting and searching using an external research collaboration platform to assist with data discovery. Including a service such as Colwiz demonstrates how independent external services can be integrated. These types of 'social' research collaboration services are gaining ground and should not be ignored.

Another strand of the Damaro project will continue discussions on the institutional data management policy, and provide user training, guidance and documentation based on the work of the SUDAMIH project. It will develop a business plan for providing and maintaining the Damaro enterprise environment for managing, preserving, and curating research data, informed by the data management policy. The Damaro project is working closely with its complementary sister project, Oxford DMPOnline[10], also funded by JISC and which will enable researchers at the University of Oxford to create, save, submit and use data management plans (DMPs), both to accompany research grant applications, and to guide the subsequent management of research data. The two projects share a steering committee. DataBank software has been released as part of the DataFlow project, and DataFinder software will be available at the end of the Damaro project.

---

1 http://www.disc-uk.org/datashare.html

2 http://imageweb.zoo.ox.ac.uk/wiki/index.php/ADMIRAL

3 http://eidcsr.oucs.ox.ac.uk/

4 http://sudamih.oucs.ox.ac.uk/

5 http://www.dataflow.ox.ac.uk/

6 http://vidaas.oucs.ox.ac.uk/

7 http://www.admin.ox.ac.uk/rdm/

8 http://damaro.oucs.ox.ac.uk/

9 http://www.colwiz.com/

10 http://imageweb.zoo.ox.ac.uk/wiki/index.php/OxfordDMPonline

Beyond the boundaries of the University

The conceptual design of the Oxford infrastructure aims to fit with both complex local federated data management, and with national and global data management activities. Local requirements include robust storage, knowledge of the existence of datasets, adequate labelling, clear data ownership and meeting funding agency requirements. Looking beyond the boundaries of the University, Oxford systems need to interact with external data repositories, ensuring that metadata can be shared, and so that easy discovery is possible. The Damaro project will provide a local deployment of this emergent infrastructure to enable outreach and dissemination at the global level. The collaboration with Colwiz will demonstrate how such local and global services can interoperate.

Metadata sharing and exchange will be enhanced by publishing metadata as linked data and by employing a sub-set of CERIF (Common European Research Information Format) for data exchange. A number of methods to support data citation are used. UUIDs (Universal Unique Identifiers) are assigned to every item in the repository and catalogue. An Oxford PURL resolver has recently been deployed to ensure link persistence. The Bodleian Libraries can mint DOIs for datasets, a service which is proving popular with data creators.

DataFinder

Discovery of datasets is paramount both internally and for external users. The project will recommend the University adopt a metadata set based on the DataCite minimum mandatory metadata kernel[11]. This will encourage researchers at the very least to provide basic details about their data. DataFinder is intended to hold metadata for Oxford datasets wherever they are located, whether internally or externally. It will provide evidence of the existence of datasets, together with minimum descriptive details of and dataset location (URL, contact details or other information). This allows Oxford data to be findable via Google, aggregators or other services. Details of location of the data will not necessarily guarantee access to the data. DataFinder will publish metadata as RDF linked data. It will have an OAI-PMH endpoint to enable harvesting and distribution of metadata. Metadata feeds will be set up to populate DataFinder wherever possible (for example from DataBank). Work will continue to automate metadata creation and deposit.

Although DataFinder will be deployed as a central institutional service for the purpose of this project, it could be implemented in a hierarchical fashion. For example, a local (departmental) instance feeds an institutional, regional, and national instance if desired. Within the University, DataFinder will be the glue that links the separate modules of the technical infrastructure and is therefore crucial to the design.

Damaro is not the final solution

The Damaro project will not result in a finished and perfected infrastructure for the University. It will present the first iteration of a modular and federated model for data creation, storage and management with supporting services. The project intends to make headway in each of the four strands, thereby building the basis of a flexible, extensible infrastructure for the future.

DataBank may not be the only central data curation service for the University. The scope of this service is yet to be defined and more central storage may be required, particularly for 'big data.' It may be the case that some academic departments choose to use the DataStage model for data management and publication via DataBank. Others may opt for direct deposit of publishable versions into DataBank. Some will want to use national and subject repositories such as DRYAD[12]. Oxford is taking the view that although data are likely to be deposited and disseminated using a number of different services, the University should provide some form of archival storage for those who require it. Who pays for such a service is still under discussion, and what cost models will be employed are as yet unknown.

The need for data repositories

There is a pressing need for a data repository at Oxford and this appears to be the case at most research active institutions. Institutions (not just libraries) are keen to provide data repository facilities and data

[11] http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf

[12] http://datadryad.org/

creators are keen to use them. The general acceptance of need for a data repository has preceded the implementation of such repositories, unlike the historical situation with institutional repositories. The following are presented as the main drivers:

- Researchers and senior university management are acutely aware of research funders' policies for deposit and open access to data in a way that took time to filter through for publications. This is partly because many major funding agencies already require researchers to submit a data management plan as part of their grant application. Applications tend to be vetted by central research services departments. There also seems more likelihood that such policies will be policed. There has certainly been an increased sense of urgency since the EPSRC released its revised data policy.[13]
- There are currently few safe archival stores for data and perhaps the concept of disappearing data is more understood. For example, data stored on a DPhil student's hard drive can become inaccessible once that student has left the institution. The need has not necessarily been the same for publications where many researchers view publishers as providing safe storage for articles.
- Researchers do not generally have easy access to other researchers' data in the same way that many do for journal articles.
- Journals are increasingly demanding a citation to research data. Some demand data deposit (for example Nucleic Acid Research[14]).
- The value of publishing datasets for peer-review and to enable research to be verified is recognised.
- Many researchers including humanities scholars are producing complex digital research data. Such datasets cannot be produced in print within a book which is a common mode of publishing for this group.
- The 'Climategate'[15] controversy and the Queen's Belfast tree rings saga,[16] coupled with the legal requirement to provide timely responses to Freedom of Information requests have scared universities into acting.
- Universities are arguably becoming more inclined to retain research data because these data are considered a manifestation of the intellectual property of the institution in a way that articles are not (at least, yet).

Much has been written about data sharing to open up the possibility of future collaborations, for furthering research and for the public good. It is so far unclear whether most academics want a data repository specifically for these reasons. According to the DCC (Digital Curation Centre), the move towards sharing is evident in 'the concerns of policy-makers, and in changes in legislation and its implementation.'[17] These changes are being driven and implemented 'through coordinated action by funders including the UK Research Councils, charities and JISC, with significant responsibilities falling to HEIs and individual researchers.'

Fit with existing data repositories

Oxford's research data systems are planned to be federated (ie united but independent). It is recognised that there are a number of existing reliable homes for data, both internally and externally. These include departmental data stores (some of which have been running successfully for decades), grid computing data services, national data services such as UKDA, and new services such as DRYAD. This means that any development of Damaro must be designed to fit with existing viable systems.

The DataFlow solution provides a neat link between local data stores (or silos) where academics can store and manipulate data and deposit into DataBank. The shunting mechanism is called DataStage and uses the SWORD2 protocol to transfer publishable data at a time chosen by the creator. It is at this stage that a

13 http://www.epsrc.ac.uk/about/standards/researchdata/Pages/default.aspx

14 http://www.oxfordjournals.org/our_journals/nar/for_authors/ed_policy.html

15 See for example 'Show Your Working': What 'ClimateGate' means,' Mike Hulme and Jerome Ravetz, BBC news website, http://news.bbc.co.uk/1/hi/8388485.stm

16 http://www.belfasttelegraph.co.uk/news/environment/queenrsquos-ordered-to-release-tree-data-14789149.html

17 Digital Curation Centre http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm#Drivers

DOI can be assigned to the dataset. This is a good example of close linkage between local and central university systems.

In recent years, publications repositories have become integrated into local research administration systems such as CRIS's (Current Research Information Systems) and administrative databases. Data repositories are likely to be considered for integration with such systems. CERIF is being increasingly cited as the standard protocol for data exchange with external bodies such as research councils. The Damaro project will expose metadata held in DataFinder in CERIF format to enable data sharing with agencies that adopt this standard.

Requirements from the outset

The Damaro project attempts to pre-empt the basic requirements for data storage and management for a large research-intensive university. The storage capacity will be estimated as the scope of central services are agreed, the systems are being built with the long-term future in mind, and the University is preparing for a data deluge. We need to be able to deal with a myriad of formats and metadata, a data model has been prepared that allows us to deal with many different entities, glued together by vocabularies as required.

It is commonly accepted that metadata is not always as good as it might be in publications repositories. It is questionable whether data repositories will fare any better for obtaining rich and accurate high quality metadata for every dataset. It is for this reason the project is will mandate minimal, yet reliable metadata for DataFinder. There is a need to be able to assign discipline specific metadata and this will present challenges for data creators and repository providers. The resources for staffing a data repository service, including providing staff to enhance metadata or for mediated deposit, are as yet unknown.

It is evident that data repositories are required now, therefore institutions need to consider what they need at the outset, including policies for smooth running of services, training, and plans for long term sustainability. There is very little existing metadata for datasets that can be repurposed: unlike publications, nearly all metadata has to be created from scratch which is a new task for most researchers and if too onerous, will not be welcomed. To make the data sets we hold interoperable, we as the service providers need to advise researchers how to describe and add metadata to their data and to make this task as automated as possible.

The work of Sansone et al at the Oxford eResearch Centre[18] illustrates how we need to work closely with researchers to learn about discipline specific vocabularies and other standards in use, so we can incorporate them into our systems. This will help make the systems more relevant to individual researchers and help us 'to empower ever more scientists to take data management and sharing into their own hands, using community standards while remaining blissfully unaware of the underlying complexities of the implementation of those standards.'[19]

Conclusions

The data environment at Oxford is rapidly developing. It is evident that there is a large amount of data already in existence and without an archival home. The Bodleian Libraries are being contacted more and more by academics wanting to find out if there is a repository for their data. The requirement to deliver a roadmap for the EPSRC has focused many minds. The Damaro project will enable the University of Oxford to progress a considerable way towards a unified infrastructure for data management. The plan is that the research data management infrastructure designed for Oxford will be comprehensive enough, and suitably modular and flexible for it to be reactive to change and to meet academic and funder preferences over time.

[18] See for example the BioScience catalog http://www.biosharing.org/standards for identifiying emerging and popular exchange formats

[19] Sansone, S.A., Rocca-Serra, P., Field, D. et al, (2012). Toward interoperable bioscience data, Nature. http://www.nature.com/ng/journal/v44/n2/pdf/ng.1054.pdf