The International Journal of Digital Curation
Issue 2, Volume 6 | 2011

An Institutional Approach to Developing Research Data Management Infrastructure

James A. J. Wilson, Luis Martinez-Uribe, Michael A. Fraser & Paul Jeffreys,
University of Oxford

Abstract

This article outlines the work that the University of Oxford is undertaking to implement a coordinated data management infrastructure. The rationale for the approach being taken by Oxford is presented, with particular attention paid to the role of each service division. This is followed by a consideration of the relative advantages and disadvantages of institutional data repositories, as opposed to national or international data centres. The article then focuses on two ongoing JISC-funded projects, 'Embedding Institutional Data Curation Services in Research' (Eidcsr) and 'Supporting Data Management Infrastructure for the Humanities' (Sudamih). Both projects are intra-institutional collaborations and involve working with researchers to develop particular aspects of infrastructure, including: University policy, systems for the preservation and documentation of research data, training and support, software tools for the visualisation of large images, and creating and sharing databases via the Web (Database as a Service).[1]

[1] This paper is based on the paper given by the authors at the 6th International Digital Curation Conference, December 2010; received December 2010, published July 2011.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Background

The University of Oxford's programme of projects to develop data management infrastructure has its roots in a cross-University committee formed in 2006 to coordinate the development of digital repositories within the University. The Oxford Digital Repositories Steering Group (ODRSG) was chaired by the Pro-Vice Chancellor for Academic Services, University Collections, and Research. The Oxford Research Archive (the University's repository for ePrints and theses) reported into the Group, as did activities related to e-learning. The ODRSG identified priorities for digital repository development including: a) to ensure interoperability between existing and planned repositories (the Group preferred to speak of a 'federated institutional repository' rather than simply an 'institutional repository' in order to reflect Oxford's devolved and federated nature as an institution); and b) to better support the management and curation of research data. The latter priority was driven partly by the Research Councils and other funding bodies increasingly requiring data management plans as a condition of funding, partly by the recognition that few University services existed to support the management of research data, and partly by new opportunities for large-scale research enabled (or potentially enabled) through the e-science agenda. All three factors are of course inextricably linked.

The ODRSG motivated the funding of an internal project, 'Scoping Digital Repository Services for Research Data Management',[2] which sought to establish exactly what was required by researchers at the University and what roles the various service groups could and should take to meet those requirements. A data management service framework was derived from the requirements and used to evaluate the current services provided at Oxford and identify gaps in provision. Following this, two further (JISC-funded) projects have commenced: 'Embedding Institutional Data Curation Services in Research' (Eidcsr);[3] and 'Supporting Data Management Infrastructure for the Humanities' (Sudamih).[4] Both of these projects have involved working with researchers identified during the scoping study to develop specific elements of infrastructure.

Approach

The University of Oxford has a highly federated structure, with the various academic divisions, the faculties within those divisions, and the supporting service groups each having a large degree of autonomy. This poses a challenge for developing a research data management infrastructure, as there is no one department in a position to 'own' the resulting set of services. Indeed, if the University is to achieve its objective of creating a solution to manage research data through all stages of the data life-cycle, from creation to long-term curation, then all of the different groups need to share a common understanding of the purpose of the enterprise and the processes required to make it work. Developing one aspect of the infrastructure in isolation and then expecting it simply to slot in with all the other aspects would not, in these circumstances, be a viable approach.

[2] Scoping Digital Repository Services for Research Data Management: http://www.ict.ox.ac.uk/odit/projects/digitalrepository/.

[3] Embedding Institutional Data Curation Services in Research (Eidcsr): http://eidcsr.oucs.ox.ac.uk/.

[4] Supporting Data Management Infrastructure for the Humanities (Sudamih): http://sudamih.oucs.ox.ac.uk/.

The data management challenge can be conceived as a sequence of steps, each of which needs to be adequately completed in order that the data itself can progress to the next step, with the intention that it can ultimately be re-used, thereby maximising its value. If the infrastructure does not exist for any given step, or the agents involved do not understand what is required of them, then the potential value of the data cannot be fully realised.

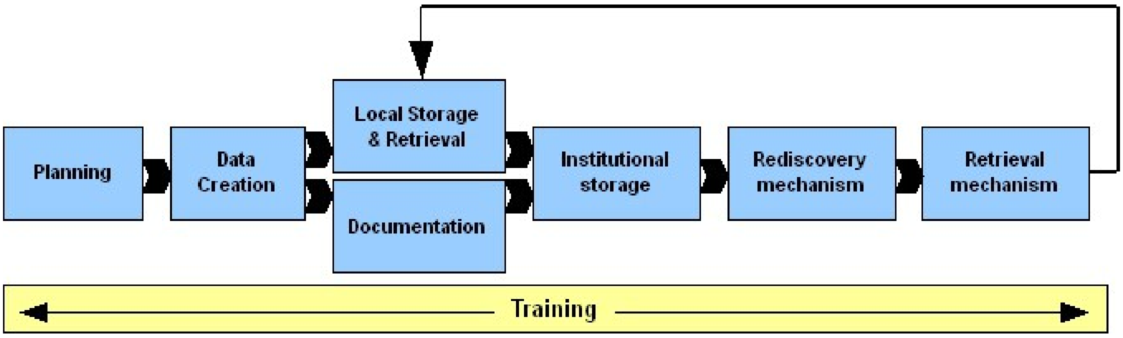

Figure 1. Required elements of an institutional data management infrastructure.

Whilst every university is likely to have a slightly different structure of departments and services, with different parts of the institution responsible for different aspects of infrastructure, the basic data management sequence is likely to be essentially the same.

With reference to Figure 1, at the University of Oxford institutional support for the planning stage of any given research project (often relating to the data management stipulations laid down by funding agencies) lies predominantly with the Research Services Office and with the research services teams within each academic division, as they are best placed to work with researchers on funding bids.

Responsibility for the data creation phase of any given project rests mostly with the researchers themselves, although each department offers training on research methods and techniques, which can help.

How researchers store and retrieve their data is, at present, up to the researchers, although the Computing Services do offer a central back-up service. Practices here vary considerably, and the interviews we have conducted suggest that many researchers do not give their local information management structures a great deal of consideration.

Documenting one’s data can take place at any point between creating the data and archiving it. As no institutional data archive yet exists, there are no requirements for researchers to document their data for later re-use. The Computing Services are working on a metadata editing system that can keep the documentation process simple and standardised.

With regards to the institutional storage, the Computing Services already provide a very secure long-term file store (the Hierarchical File Server, based on IBM Tivoli Storage Manager technology),[5] but there is no system in place as yet for associating data stored here with any metadata stored separately. Users of the HFS long-term file store facility are expected to provide the data management and curation layer; there is no University-wide service built on, or distinct from, the HFS.

The Bodleian Libraries are developing a data repository system named 'Databank', based on the Fedora digital assets management system (DAMS), which promises to offer metadata management and resource discovery services. This looks as though it will provide the ideal home for metadata, with the last two steps of the infrastructure – the discovery and retrieval mechanisms – built on top.

Researcher training is required at all stages of data management, from planning to using the retrieval interface. If this training is to be sustainable, it makes sense for the services providing and maintaining the infrastructure at any given step to take on the associated training role as well, so that training can be kept up to date as the infrastructure evolves.

Oxford is taking the approach that it is the researchers themselves who are best placed to describe their data, as it is they who understand the processes by which the data was derived, the context in which it was assembled, and any limitations which might not be immediately apparent to other researchers wishing to re-use it in the future. The role of the libraries is to maintain access to the metadata, rather than create it, although this may include the management of metadata specifically relating to preservation and curation.

The Computing Services has assumed the coordinating role at the University of Oxford for developing data management infrastructure, although in some respects this was due to the specific set of circumstances in place when the process began – at other universities it may as well be the libraries or another service department who takes the lead. At Oxford, the Computing Services was well placed to commence work due to a combination of the existing infrastructure that it supported (in particular the secure long-term file store), a long tradition of working with researchers supporting a wide variety of research projects (e.g. the OUCS Research Technologies Service provides support for the use of IT in research), and the fact that the Office of the Director of IT is embedded within the department, ensuring that strategic decisions could be made and communicated at the appropriate level. Whilst coordinating activities, the Computing Services has at all stages worked closely with other service departments and academic groups, notably the Oxford e-Research Centre and the Library Services.

The general approach being taken to the development work at Oxford may be summarised as an attempt to advance towards a coherent infrastructure on all significant fronts, using the researchers to guide and validate each strand of development as it progresses. The research communities with which the projects are working would not claim any especial expertise in data curation. Indeed an important aspect of the projects is to be able to gauge the buy-in from 'normal' researchers whose focus is on the day-to-day research itself. Whilst both the Eidcsr and Sudamih

[5] For more information about the Hierarchical File Server at Oxford University Computing Services (OUCS), see: http://www.oucs.ox.ac.uk/hfs/index.xml.

projects are guided and informed by the specific requirements of particular research groups, an important aspect of each is to assess how successfully (and economically) the outputs can be expanded to meet the needs of researchers in the University more broadly.

By working on several different aspects of data management in close coordination, the projects do not lose sight of the interrelated nature of data management activities. Unless preservation strategies can be allied with resource discovery services, for instance, the value of preservation is severely limited; without data management training and awareness of good practice, technical tools and services developed to assist with such management are unlikely to be taken up and used, or even understood by researchers on the ground.

Institutional Versus National Data Management Infrastructure

It may reasonably be asked why universities should each (or in consortia) wish to invest resources into creating and sustaining their own data management infrastructures when such infrastructure could alternatively be created at a national level, where one might expect greater economies of scale and concentration of expertise.

As the situation stands, not all academic disciplines are covered by the various national and international subject data centres, nor is it likely that every potential topic of research ever will be. Research groups within certain disciplines, such as crystallography and astronomy,[6] tend to generate data that is relatively comparable. Here, standards may be implemented that do not unduly reduce the scope for creativity and innovation, and the potential advantages of sharing data between research groups are clear to the researchers themselves. In such situations, the usefulness of national or international data repositories are self-evident: expertise in curation may be centralised, resources pooled, and services do not need to be replicated in multiple places. For other disciplines, in which the data generated is more diverse, national and international data centres can still facilitate a more limited uniformity of curation and share subject-specific expertise in a more economical manner than would be achieved by individual institutions each developing their own infrastructure. Such data centres can pool training resources and offer them to researchers in many different universities, acting as a pan-institutional resource. The UK Data Archive serves this function for the social sciences in the UK.[7] It should not be forgotten, however, that there are a great many small research groups or even lone researchers who work in diverse fields producing data with quite specific characteristics who have requirements that are not easily generalised, or in subjects areas that are too narrow to justify the costs of establishing and maintaining large data centres staffed by specialists.

[6] See, for instance, the European Virtual Observatory: http://www.euro-vo.org/pub/ and the related AstroGrid applications: http://www.astrogrid.org/, the eCrystals repository at the University of Southampton: http://ecrystals.chem.soton.ac.uk/ and standards established by the International Union of Crystallography: http://www.iucr.org/resources/cif/spec.

[7] UK Data Archive: http://www.data-archive.ac.uk/.

It is worth introducing here a distinction between 'high' and 'low' levels of curation. 'High curation' may be considered as a service requiring high levels of expertise, where subject specialists are involved during the ingest phase of data archiving, adding and cleaning descriptive metadata; 'low curation', on the other hand, would signify a greater degree of automation, with data being contributed to a repository with minimal manual intervention (Rusbridge, 2010). A national subject repository might be expected to provide a higher level of service than an institutional repository, where the emphasis would be on the preservation of a wide range of content and where it would be uneconomical to employ enough subject specialists to cover every eventuality. The infrastructure being developed at Oxford must, of necessity, tend towards the lower end of the spectrum, resting on a relatively automated set of procedures with the researchers themselves taking a larger share of responsibility for the documentation of their data, with the concomitant risk of idiosyncratic metadata and less than perfect conformity to standards. The processes involved also need to be relatively simple and unburdensome, otherwise the risk is that researchers do not bother to go through the required workflows in the first place.

Whilst the 'low curation' approach that institutional repositories necessitate incurs risks that a higher service levels avoids, universities do have certain advantages over national data centres, primarily relating to the close support they can offer researchers during the early stages of the research process. At Oxford, as at other institutions, research services teams within the academic divisions can advise researchers on data management issues during the grant proposal stage to ensure that they are aware of data management requirements and have properly considered how they will approach the issues involved. Technical help is available from the Computing Services, who can recommend how data may best be structured and stored, if necessary working closely with research teams on databases or textual mark-up.

A second argument in favour of institutional data management infrastructure relates to universities' reputation management. Universities, such as Oxford, take ownership of the data produced by their researchers (up to a point), and therefore have responsibilities for it. The exact nature of any given university's intellectual property rights regarding research data varies, but at a purely practical level it is likely to prove embarrassing for a university if the data produced by its researchers cannot withstand reasonable scrutiny.

Finally, there are potential concerns regarding the long-term sustainability of national data centres, particularly since the demise of the Arts and Humanities Data Service (AHDS) in the UK in 2008.[8] Although parts of the former AHDS have continued to operate, and the Archaeology Data Service continues to be funded at a national level, the withdrawal of much of the funding for the organisation undoubtedly created a degree of alarm. Whilst in practice it is true that institutions may also decide to remove resources dedicated to their institutional repositories during hard times, there is a school of thought that suggests that the multiple income streams that universities receive, and the longevity of the institutions themselves, offer better guarantees of survival than specialised national services subject to the vagaries of government funding decisions.

[8] Arts and Humanities Data Service: http://www.ahds.ac.uk/.

Eidcsr

The Embedding Institutional Data Curation Services in Research project (Eidcsr) began in April 2009, funded by the JISC Information Environment Programme.[9] It is due to end in December 2010, having made progress towards implementing a number of key elements of a consolidated data management infrastructure. Eidcsr is working with researchers involved in an inter-disciplinary research project developing three-dimensional models of hearts, which can be used to conduct *in-silico* experiments.

Eidcsr followed the initial scoping study, inheriting a framework where researchers are at the core of an intra-institutional collaborative network that includes the Computing Services, the Libraries, the Oxford e-Research Centre and the Research Services. The main aim of this collaboration was to address the data management requirements of the research groups involved in the 3D Heart Project, but with a view to developing expertise and infrastructure of use to other researchers beyond those immediately involved. This approach attempts to harness expertise from a range of institutional stakeholders in order to deal with a variety of research data challenges (Macdonald & Martinez-Uribe, 2010). The project was conceived to work across different areas of action: some related to the data management needs of the specific research groups and others relevant to more general institutional data management matters.

The specific data management requirements of the research groups were initially gathered using the Data Audit Framework methodology and continuously refined through meetings and one-to-one informal conversations.[10] A member of the Oxford e-Research Centre was in charge of the requirements elicitation as per the Centre's expertise in this area. Researchers' main concerns had to do with keeping their data secure, having the ability to search and browse their datasets, and accessing large image data rapidly and reliably. These requirement are being addressed through a variety of technical developments undertaken by the Computing Services and the Library.

Research Data Archiving and Access

A set of core metadata fields, applicable to a wide range of research data, have been agreed with the researchers working on the 3D Heart Project. These fields, which mostly map to Dublin Core, include a 'process' field in which researchers can record the experimental process by which their data was derived, assisting reproducibility, and a 'relationships' fields that can be used to associate datasets with publications. The basic metadata is designed to be extensible, so that metadata specific to particular research groups can be added to record additional details. Further testing with research groups is still required to ensure that the schema is applicable to different subject disciplines, but it is hoped that the core set is broad enough to meet most requirements, whilst still simple enough that researchers will not be put off by the additional demands that documentation places upon their time.

[9] JISC Information Environment Programme: http://www.jisc.ac.uk/whatwedo/programmes/inf11.aspx.

[10] Data Audit Framework (now renamed 'Data Asset Framework'): http://www.data-audit.eu/.

An archiving client for the Hierarchical File Server has also been developed in order to capture and record metadata during the data archiving process. The basic principal underlying the archiving system is that metadata records are saved into the directory structure of the data to be archived. As the data is archived to the Computing Services' existing long-term file store, the metadata is interpreted and added to a Fedora-based digital assets management system (DAMS) run by the Library. The data and metadata are linked by a unique identifier. An interface to the DAMS is currently being created, which will enable users to search and browse the metadata, placing an order to download relevant data with a designated data curator.

To enable fast access to (and annotation of) the large image datasets being created by the 3D Heart Project, a visualisation tool has been developed that only downloads in high-resolution the image tiles needed for 'magnified' viewing as required, dramatically reducing access times. The images may be annotated to assist sharing between research groups in different locations, and visual thresholding tools have also been implemented so that researchers can bring out particular features of the images they wish to view.

Institutional Data Management Policy

Another aspect of data management being addressed by Eidcsr is the development of an institutional policy. The process of developing a data management policy at a highly-devolved university, such as Oxford, requires extensive consultation with stakeholders, and the approval of the relevant committees. The management of research records and data is part of a wider programme of research integrity led by the Research Services Office, so it was natural that they took the lead in developing the data management policy. The experiences of the University of Melbourne, with whom the project consulted, showed that such policies need to be accompanied both by activities to raise awareness of what is being advised, and institutional support services to enable researchers to actually implement the recommendations. A draft policy document has been produced and is currently pending approval.

One of the recommendations of the draft policy document is to create a data management Web portal, which is now under development.[11] This will connect researchers to basic information about all aspects of data management, highlighting existing services and integrating content from Oxford and from external sources, such as the Data Curation Centre (DCC).[12]

Cost Modelling

One of the biggest challenges of the programme to implement a data management infrastructure at Oxford is to ensure that what is developed within each project phase is sustainable as an ongoing service. Both the Eidcsr and Sudamih projects are developing cost-benefit models, which are intended to be useful both within and beyond the institution. The JISC Managing Research Data Programme is helping with this by coordinating such work across a number of projects looking at data management issues in various institutional contexts.[13] By combining these case studies with the work already being undertaken by the Keeping Research Data Safe projects

[11] University of Oxford Data Management Portal [forthcoming]: http://www.admin.ox.ac.uk/rdm.
[12] Data Curation Centre: http://www.dcc.ac.uk/.
[13] JISC Managing Research Data Programme: http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx.

(KRDS),[14] it is hoped that a set of generalised models can be drawn up to assist other institutions undertaking similar exercises in future.

The Eidcsr project participated in the KRDS2 project, contributing detailed cost information about the creation, local management, and curation of the data produced by the participating researchers (Beagrie, Lavoie & Woollard, 2010). The results showed that the cost of creating the actual research datasets, including staff time as well as the use and acquisition of lab equipment, was proportionally high, representing 73% of the total cost. The costs of the admittedly limited local data management activities undertaken by the researchers, on the other hand, were modest, representing only 1% of the total. The curatorial activities undertaken as part of the Eidcsr project constituted 24% of the total, but much of this represented start-up costs which would not need to be factored into the equation were data curation infrastructure already in place. 2% of the costs stemmed from the secure storage of the very large datasets (c. 5TB) for five years on the long-term file store. Although these proportions are specific to the project, they are likely to be broadly indicative, and one would certainly expect to achieve significant economies of scale once the required infrastructure is in place.

Oxford University Computing Services, working with JISC, are currently in the process of developing a toolkit for costing IT services.[15] This advocates a step-by-step approach, beginning with determining the purpose of the service and risks associated, and proceeding through a Full Economic Costing methodology before concluding with planning for reporting and service usage monitoring. Two important principles defined by the toolkit for costing IT services are that a) any costing model has to be readily understood by those who allocate resources (in this case the Toolkit follows TRAC, the Transparent Approach to Costing, that is familiar to senior academics and administrators); and b) the costing model must be scalable, for example able to be applied to both a university-wide data management infrastructure service and to a local, department service (to ensure cost comparisons are like for like). The Eidcsr and Sudamih projects will follow the steps laid down in the toolkit and hopefully further contribute to it.

Because Oxford is in part developing its infrastructure around elements already in place, such as the HFS at the Computing Services, it should be possible to arrive at a reasonable estimate of per terabyte costs of data storage and basic curation. The HFS functions on a cost-recovery basis, and at present offers one terabyte of storage free to research projects, charging additional terabytes at a rate of £550 per terabyte per year for internally-funded projects. The full economic cost is £842 per terabyte per year, including staff time and overheads. Oxford's 'low curation' highly-automated approach should minimize staff costs, which tend to be proportionally high (Beagrie, 2010), by shifting the emphasis of data documentation and curation onto researchers rather than library staff. Although this will place some demands on the time of the researchers, we hope the minimal metadata requirements will not be onerous and discourage participation. If the process of adding metadata takes more than around thirty minutes per significant dataset, this may be a problem (similar challenges were faced by the Oxford Research Archive for ePrints and theses). We are consulting with the researchers involved in the 3D Heart Project to ensure demands are not unreasonable.

[14] Keeping Research Data Safe 2: http://www.beagrie.com/jisc.php.

[15] The Toolkit for Costing IT Services will be available via: http://www.jiscinfonet.ac.uk/flexible-service-delivery.

Sudamih

The Supporting Data Management Infrastructure for the Humanities (Sudamih) project began in October 2009 and is due to conclude at the end of March 2011. Funded by JISC as part of the Research Data Management Infrastructure Programme, Sudamih focuses on two main strands of development: the creation of a ‘Database as a Service’ (DaaS) system; and the development of training materials intended to help researchers improve their data management practices. The project is working with researchers in the Humanities Division at Oxford to address their particular concerns and requirements, although it is envisaged that the outputs of the project should be easily extendible to benefit those in other academic disciplines as well. The materials developed by the project will be made available to other institutions to use or customise as they see fit.

An important consideration during the project’s requirements-gathering interviews was that many researchers in the humanities do not think of themselves as creators of data *per se*. Although the use of ‘data’ in a narrow sense (information structured in a consistent manner for the purposes of searching and analysis) seems to be growing in the humanities, the majority of those we spoke to did not engage with data in this sense. We therefore adopted a broad definition of ‘data’, essentially encompassing all information gathered and processed during the course of research that was intended to lead to some sort of research outcome. It was emphasised by the researchers themselves that care would need to be taken when developing data management training that we did not accidentally alienate sections of our target audience by using terminology that they were unfamiliar with or which did not obviously correspond with their particular concerns. ‘Information management’ might, therefore, be a preferable term to ‘data management’ where training does not specifically concern data in the narrow sense of information in spreadsheets or databases. It also became clear that using the terminology of the data curation community to describe activities was also to be avoided. Speaking of ‘ingest’ was not a good idea, and if we were to refer to ‘metadata’ this would need to be explained.

The requirements-gathering exercise also revealed several other important concerns. Firstly, it brought home the enormous diversity of information that humanities scholars work with and the variety of practices employed to organise that data. Secondly, we found that practices relating to activities such as back-up and storage, versioning, and keeping files synchronised across multiple computers tended in many cases to be rather rudimentary. Most humanities researchers work on their own laptop and desktop computers, backing up their data onto memory sticks and hard drives. There was, on the whole, little awareness of existing central services, and even shared departmental servers tended to be used only by those with considerable experience of working on ‘narrow’ data projects, where departmental IT staff were directly engaged.

The nature of humanities data itself tends to have different characteristics from the data produced in other disciplines. When dealing with structured data, humanities researchers (with the exception of some in linguistics, history and archaeology) do not usually create structured ‘data’, but rather compile information from various existing sources, which could be manuscripts, inscriptions, previously published books and articles, newspapers, administrative records, maps, or a multitude of other sources. As a result of this, the data compiled is frequently incomplete, inconsistent, unreliable, ambiguous, and open to interpretation. If such data is to be shared, the nature of the sources and the way in which the data has been compiled needs to be well documented, otherwise misinterpretations by scholars removed from the process of gathering the data are likely to be a serious problem.

A final important characteristic of humanities data is that it does not tend to depreciate in value in the manner that some scientific data does. Whereas the researchers we are working with on the Eidcsr project estimate that in five years' time the costs of preserving their data will have been outweighed by the decreasing cost of re-creating it (and with improved imaging technology), it is clear that this is not the case with most humanities data. This is illustrated by one of the projects Sudamih is working closely with. The Roman Economy Project has been developing a database of economic activity across the Roman world, combining information about cities, demographics, patterns of trade, and sites of production. It is a large ongoing project with a significant relational database (Bowman, Wilson et al., 2010). The data that they have gathered is likely to be as much use in 50 years time as it is today, potential even more so as other data sources can be linked to it. It is therefore important that the information it contains is preserved for future generations, just as print works have been in the past. This places obvious demands on any infrastructure developed, which must offer an essentially permanent guarantee of preservation. In other situations as well, humanities scholarship often aggregates to a 'life's work' body of research, with any given researcher often wishing to go back to old datasets in order to derive new information.

Training

Many of the humanities researchers interviewed by the Sudamih Project had never previously considered data management training or what it might consist of, although most could, upon reflection, see why such training might be useful. Particular aspects of training that the researchers thought would be useful included: organising one's files so as to be able to find information quickly when required; linking notes to content; keeping track of sources; backing up; versioning; awareness of what software tools are available and which software is best for dealing with particular research challenges; and structuring data, particularly in relational databases. Several researchers also thought it would be useful if the University offered some sort of consultation service to help with funding bids or to assist with the technical aspects of database design. In fact, such consultation services are already available, suggesting that investment in publicity might, in some cases, be as effective as investment in infrastructure.

Training is arguably the aspect of work that most clearly necessitates strong intra-institutional coordination. Oxford has a well-developed training infrastructure, with various parts of the University offering different aspects: the Computing Services, the Libraries, the academic divisions, faculties and departments within the divisions, the Learning Institute, the Careers Service, and the Research Services all provide training in one shape or another. Data management training, however, does not fall neatly into the existing remits of any single group. Assistance with the writing of funding bids tends to be the domain of the Research Services, whether at the institutional or divisional levels; the Computing Services provide training on using particular software tools and structuring data; the academic divisions provide researcher training for generalisable skills such as presenting papers, managing one's doctoral thesis, publishing articles, and suchlike; the individual faculties provide training for research skills more specific to their disciplines, including dealing with particular sources, methodological approaches to research, or elements of research ethics and information handling; the Libraries deal with finding information and keeping track of sources and citations. At present, many training courses and materials touch upon data management issues, but nothing addresses data management directly.

Many of the researchers Sudamih spoke to could see the need for data management training, but warned that it might be something of a ‘hard sell’ to get researchers to put time aside in order to undertake it. One fairly typical response was that: “I’m not sure if it’s really useful enough to give up an afternoon … [but] over the long run, if it saves time, then it almost certainly is worth giving up the afternoon” (Wilson & Patrick, 2010). Given these misgivings, Oxford is taking the approach that data management training should be integrated into existing training infrastructure, as far as is possible. This is partly to ensure that such training reaches the researchers it needs to reach in a context where its importance can be appreciated, but also to ensure it is sustainable. By using existing channels, some of the challenges of requesting additional funds can be side-stepped.

Database as a Service

One of the major endeavours of the Sudamih Project is the development of a ‘Database as a Service’ system. This will take the form of a web-based interface for creating and editing relational databases, querying them, and displaying results in various formats. At present, many humanities researchers who do structure their data in databases do so via software installed locally on their own laptop machines, over which the DaaS presents several advantages: the data can be accessed via any machine with an Internet connection; back-up is automated and regular; researchers can collaboratively add and edit the data; consistent metadata can be captured to make databases easily discoverable; and the data held in the databases can be opened up to the public via simple generic search interfaces if and when desired, with little technical knowledge required of the researcher.

Conclusions and Next Steps

There are two principles underlying the University of Oxford’s institutional approach to research data management: researchers need to be at the core of development; and there must be intra-institutional collaboration amongst service providers. It is perhaps obvious that understanding the requirements of researchers, and retaining their engagement, is crucial to the longer-term success and sustainability of any institutional initiatives to better support the management of research data. If the benefits of the centrally-provided infrastructure are not self-evident to researchers, there is nothing to stop them simply ignoring it and managing (or mismanaging) their data themselves, as is largely the case at present. Less obvious may be the need for coordination between service providers, to ensure that the most appropriate support can be given at each stage of the data life-cycle, and that potential value is not lost as the data passes through each stage. At an institution such as Oxford, one cannot hope to build and sustain a data management infrastructure merely by appointing one or two professional data curators. It is a challenge that requires various service providers and researchers to arrive at a mutual understanding of data management requirements, practices and benefits, and to work accordingly. It arguably matters less which part of the organisation is tasked with coordination than whether the relevant providers are engaged in the undertaking and have a clear sense not only of current service provision and strengths but also the gaps and weaknesses.

Developing an institutional data management infrastructure takes time. Whilst the Eidcsr and Sudamih projects are developing tools, processes, policy, and training materials to address the various stages of research data management, they are both pilot projects focusing on the requirements of particular researchers within the University. Considerably more work will need to be undertaken beyond March 2011 to build a robust and sustainable infrastructure that meets the needs of researchers across all academic disciplines. Scoping work undertaken by Oxford in relation to the UK Research Data Service (UKRDS)[16] has identified several areas in which investment should now be focused. These include the need to implement aspects of institutional IT infrastructure – especially a federated, lightweight, extensible file store offered and coordinated on a cost-recovery basis and interoperable with other research-support services such as SharePoint and the Digital Asset Management System – and the need to build upon the work of Sudamih by extending research data management training to other divisions, especially through existing training facilitators.

Whilst Oxford, as an institution, is fairly clear about where its priorities lie with respect to research data management, the situation remains complex because, of course, neither Oxford nor its researchers exist in isolation. A large proportion of research within Oxford, both within the sciences and the humanities, is collaborative and often international in scope. Well-established domain-based data repositories exist for some subjects, but none at all for others. Academics, especially those early in their careers, tend to be mobile, moving between institutions and taking data (and occasionally entire research groups) with them. Whether data follows the researcher, or remains behind, there is the continued risk of fragmentation. The development of infrastructure to support data management has to reflect, as far as possible, this fluidity. It may be, as the current infrastructure planning within universities is tending, that data management is better placed in, or at least integrated with, so-called 'cloud-based' services whose elasticity means that that in some sense they are *in* but not *of* any institution.

References

Beagrie, N. (2010) *Keeping Research Data Safe factsheet: Cost issues in digital preservation of research data.* Retrieved May 5, 2011, from http://www.beagrie.com/KRDS_Factsheet_0910.pdf.

Beagrie, N., Lavoie, B. &Woollard, M. (2010) *Keeping Research Data Safe 2.* Retrieved May 5, 2011, from http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx#downloads.

Bowman, A., Wilson, A., et al. (2010) The Oxford Roman Economy Project. Retrieved May 5, 2011, from http://oxrep.classics.ox.ac.uk/.

[16] UK Research Data Service: http://www.ukrds.ac.uk/.

Macdonald, S. & Martinez-Uribe, L. (2010) Collaboration to data curation: Harnessing institutional expertise. *New Review of Academic Librarianship 16*(S1), 4-16. doi:10.1080/13614533.2010.505823

Rusbridge, C. (2010) Reflections from the Blue Ribbon Task Force. In *Proceedings of the 5th Research Data Management Forum.* Manchester: UK. Retrieved May 5, 2011, from http://www.dcc.ac.uk/events/research-data-management-forum/rdmf5-economics-applying-and-sustaining-digital-curation.

Wilson, J.A.J. & Patrick, M. (2010) *Sudamih researcher requirements report.* Retrieved May 5, 2011, from http://sudamih.oucs.ox.ac.uk/docs/Sudamih%20Researcher%20Requirements%20Report.pdf.