The International Journal of Digital Curation
Volume 2, Issue 8 | 2013

Towards a Unified University Infrastructure: The Data Management Roll-Out at the University of Oxford

James A.J. Wilson and Paul W. Jeffreys,
Damaro Project,
University of Oxford

Abstract

Since presenting a paper at the International Digital Curation Conference 2010 conference entitled 'An Institutional Approach to Developing Research Data Management Infrastructure', the University of Oxford has come a long way in developing research data management (RDM) policy, tools and training to address the various phases of the research data lifecycle. Work has now begun on integrating these various elements into a unified infrastructure for the whole university, under the aegis of the Data Management Roll-out at Oxford (Damaro) Project.

This paper will explain the process and motivation behind the project, and describes our vision for the future. It will also introduce the new tools and processes created by the university to tie the individual RDM components together. Chief among these is the 'DataFinder' – a hierarchically-structured metadata cataloguing system which will enable researchers to search for and locate research datasets hosted in a variety of different datastores from institutional repositories, through Web 2 services, to filing cabinets standing in department offices. DataFinder will be able to pull and associate research metadata from research information databases and data management plans, and is intended to be CERIF compatible. DataFinder is being designed so that it can be deployed at different levels within different contexts, with higher-level instances harvesting information from lower-level instances enabling, for example, an academic department to deploy one instance of DataFinder, which can then be harvested by another at an institutional level, which can then in turn be harvested by another at a national level.

The paper will also consider the requirements of embedding tools and training within an institution and address the difficulties of ensuring the sustainability of an RDM infrastructure at a time when funding for such endeavours is limited. Our research shows that researchers (and indeed departments) are at present not exposed to the true costs of their (often suboptimal) data management solutions, whereas when data management services are centrally provided the full costs are visible and off-putting. There is, therefore, the need to sell the benefits of centrally-provided infrastructure to researchers. Furthermore, there is a distinction between training and services that can be most effectively provided at the institutional level, and those which need to be provided at the divisional or departmental level in order to be relevant and applicable to researchers. This is being addressed in principle by Oxford's research data management policy, and in practice by the planning and piloting aspects of the Damaro Project.

International Journal of Digital Curation (2013), *2*(8), 235–246. http://dx.doi.org/10.2218/ijdc.v2i8.287

 doi:10.2218/ijdc.v2i8.287

Introduction

Over the past three and a half years, the University of Oxford has undertaken a number of projects aimed at developing aspects of research data management infrastructure for the institution and the broader community. The projects have encompassed training, software tools, service planning and institutional policies. Some tools have been designed for use by researchers whilst they are actively gathering and analysing data, and some for support departments to improve how research data is looked after over the long term. Some of the projects have looked at issues facing specific research disciplines, whilst others have tried to consider 'research' in more general terms to develop infrastructure to serve the university as a whole. Projects have involved staff from Research Services, IT Services, the Bodleian Libraries and various academic departments. In summary, our approach to building a research data management infrastructure thus far has been rather piecemeal.

Much of the funding for the projects we have undertaken has come from Jisc – the UK government-funded body charged with driving forward innovation in UK education and research. This has enabled us to work alongside other higher education institutions and to share findings and experiences, and has been absolutely vital in bringing Oxford to the situation it is now in. However, this funding route has also called for a degree of opportunism, as groups within the university have proposed projects that fit within the focus of particular calls. The university is now beginning to recognize that a more institutionally-coordinated approach to research data management is going to be required in order to put in place an infrastructure that will meet the needs of researchers, and the requirements of their funders, over the coming years.

The project tasked with initiating this process was the Jisc-funded Data Management Roll-out at Oxford (Damaro) Project, which began in October 2011.[1] The stated aim of the Damaro Project was to 'embed and integrate the outputs of a number of UMF and Jisc-funded research data management projects into an enhanced institutional infrastructure, supported by researcher training and guidance, and underpinned by a University Data Management Policy'. The university research data management policy was ratified in July 2012, and officially announced to the university on the 4th October, when it was published in the Oxford Gazette (2012). A key challenge that the university now faces is to ensure that its members are able to meet the criteria set out in the policy in practice.

The challenges faced by Oxford are shared by many other research-intensive universities, who also need to develop infrastructures for research data management – not least because of the institutional research data management requirements placed on UK institutions by the Engineering and Physical Sciences Research Council.[2] We hope that this paper may provide them, and others, with some ideas relating to the paths ahead.

[1]Damaro Project: http://damaro.oucs.ox.ac.uk/

[2]See the EPSRC Policy Framework on Research Data: http://www.epsrc.ac.uk/about/standards/researchdata/Pages/policyframework.aspx

An Institutional Data Management Infrastructure

In an article published in the International Journal of Digital Curation in 2012, we described the principles underpinning the infrastructure development at Oxford as being that 'researchers need to be at the core of development; and there must be intra-institutional collaboration amongst service providers' (Wilson et al., 2012). These principles are still very much at the core of our approach. Furthermore, our research lifecycle-based model against which data management 'interventions' are to be situated is still similar to that used to illustrate the earlier article. The responsibilities of each of the main three support services (Research Services, IT Services, and Bodleian Libraries) still lie where they were originally identified (with the project planning, 'active data', and long-term preservation phases respectively). One aspect that has changed since that earlier paper is that it is now recognized that a greater degree of overarching coordination would be beneficial. Also, whereas previously the parts of the research data lifecycle were envisaged as discrete phases, we now regard it as more helpful to think in terms of a continuum, in which researchers may be working with different parts of their data in different ways at any given time. Consequently, data management interventions need to be more appropriately aligned to this way of working.

An attempt to represent this is made in the Damaro Project poster (Figure 1), where the research data lifecycle is plotted in a less bounded manner than formerly, and the centrally provided tools and services that can support and assist the researchers at each stage are less formally mapped, better recognizing how researchers actually tend to work with their data (at least in disciplines where there are not strict and commonly-accepted workflows for data management). Below the line, the tools and services that Oxford is developing are placed roughly in alignment with the parts of the lifecycle that they are intended to support and improve. The infrastructure as a whole is underpinned by policy and training – which should affect all parts of the lifecycle – and sustainability, which is a practical requirement for the infrastructure.

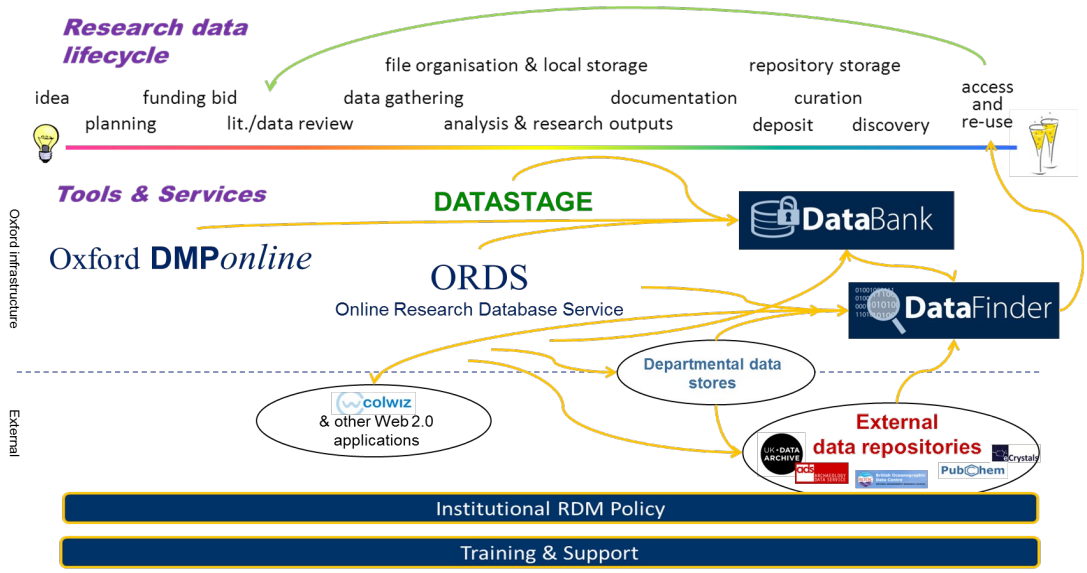

Figure 1. Damaro Project poster, illustrating aspects of the research data lifecycle and supporting infrastructure.

 doi:10.2218/ijdc.v2i8.287

The 'DataBank' and 'DataFinder' components of the infrastructure are highlighted, as these are two of the central outputs of the Damaro Project, and are required in order to 'close' the circle and enable research data generated in one project to flow back to the earlier phases of another to be re-used, maximizing the value of the research data.

DataBank will become the central institutional archive for research data produced by researchers at Oxford, whereas DataFinder will act as a catalogue or registry of such data – enabling it to be discovered, and, if the appropriate permissions are in place, to be accessed and re-used.

It is not anticipated the DataBank will be used for all, or even most, of the research data that needs to be preserved. Instead, it is intended to be used for data which has no other more appropriate repository in which it can be placed. If there are national or international subject-specific repositories for the data in question, it will be far more sensible to use these and make use of the specialized curation skills of their repository staff than to use the more generic curation services associated with the institutional archive.

DataFinder, on the other hand, should hold records for all research data produced by the university, whether the data is held locally or on systems outside the university. The DataFinder metadata schema consists of relatively few mandatory fields, essentially just the core DataCite fields required to render research data citable (Starr et al., 2011), but a larger number of 'optional' fields to improve the searchability of records and assist researchers in understanding the data and how it was gathered. DataFinder records will also record the relationships between the data and published articles, and indeed other datasets.

We know from past work with researchers that their tolerance towards entering large amounts of metadata by hand varies greatly,[3] so we are designing the system to harvest as much metadata as is possible automatically from other sources. These will include our Research Services' forthcoming research management information system,[4] the Oxford DMP (Data Management Planning) Online tool,[5] and other tools and services in which the data has been held during stages of its lifecycle, including DataBank. OAI-PMH harvesting is already being used to collect metadata about research data in data stores external to the University of Oxford.

The DataFinder system will be closely integrated with the university's existing archive of research papers, the Oxford Research Archive[6] (ORA). The interfaces of ORA, DataFinder and DataBank are all being adapted at present to offer the same presentation and structure, so that it will seem to researchers as though they are

[3]During the Jisc-funded 'Embedding Institutional Data Curation Services in Research' (EIDCSR) Project in late 2010, we created an online metadata deposit form and asked a team of researchers involved in 3D heart imaging to try it out with reference to their data. Three of the six had no qualms about filling in the form in its entirety (which was reported to take between five and 15 minutes), one said that he 'wouldn't mind, but some would', and the other two both found the process irritating, and reported that they would be particularly loath to use the system if they were required to enter the same information for multiple records. In this project, the metadata input form was shorter than that which we will be using for DataFinder. To find out more, visit the EIDCSR Project website at: http://eidcsr.oucs.ox.ac.uk/

[4]See http://www.admin.ox.ac.uk/pras/research/symplectic/ for more information.

[5]See http://datamanagementplanning.wordpress.com/ for more information.

[6]ORA: http://ora.ox.ac.uk/

interacting with a single system. We hope that this approach should reduce the need for separate training materials and help minimize the recognized problem with infrequently-used systems where users forget how to operate them between visits.

DataStage[7] and the Online Research Database Service (ORDS)[8] are two of the other components in the Oxford infrastructure. Both have been developed by the university with Jisc funding, and both are primarily intended as tools for researchers working with 'active' data. Besides offering researchers simple platforms for collaboration and data sharing, both systems are designed to capture the metadata required for long-term preservation and re-use whilst the data is still being worked with. The ORDS is intended for researchers working with data structured in databases (initially relational databases, but hopefully XML and document databases thereafter); one the other hand, DataStage is intended for unstructured data contained in files of various formats. At the end of a research project, or when data outputs are published, the metadata collected by these services can be used to automatically create an appropriate record in DataFinder, ensuring the university can keep track of the data its researchers produce, and minimizing the dependency on researchers manually inputting metadata at the moment of ingest.

All of the software underpinning these services is being made available under open source licences. It is hoped that other institutions around the world will pick up the software and adapt it to meet their own needs.[9]

Researchers' Views

Although several surveys and a host of researcher interviews have been undertaken at Oxford during the past few years, most of these have focussed on particular research disciplines and been related to particular tools or services as projects have required. It was not until November 2012 that the university decided to undertake a general survey of research data management practices and attitudes amongst researchers working with data, across all disciplines. The University of Oxford Research Data Management Survey 2012 was undertaken as part of the Damaro Project, largely as a benchmarking exercise to measure the impact of future endeavours and to better focus investment. We intend to repeat the survey in future years.

314 responses to the survey were received, with each of the four academic divisions at Oxford well represented. 25% of respondents were from the Humanities Division, 25% were from the Mathematical, Physical and Life Sciences, 29% were from the Medical Sciences, and 20% were from the Social Sciences. The rest reported themselves as 'other'. 61% of the respondents worked with textual data, 63% with numerical data, 62% with statistical data, and a surprisingly large number with images (43%). There were some significant minorities working with more specialized types of

[7]DataStage: http://www.dataflow.ox.ac.uk/index.php/about/about-datastage

[8] ORDS: http://ords.ox.ac.uk/

[9]At present, the source code for the in-development software components of the future University of Oxford Research Data Management Infrastructure is available from the following repositories: DataFinder (main programme): https://github.com/bhavanaananda/datafinder; DataFinder Data Reporter (component of DataFinder): https://github.com/asif-akram; DataBank: https://github.com/dataflow/RDFDatabank; the Database-as-a-Service software (which underpins the forthcoming ORDS service): http://code.google.com/p/ords/; DataStage: https://github.com/dataflow/DataStage.

 doi:10.2218/ijdc.v2i8.287

data: 14% of respondents worked with geospatial data in some capacity, and 14% with audio data. Perhaps less surprisingly, 76% of respondents stored at least some of their data in tables and spreadsheets, whereas 32% used relational databases and 23% used other types of database. 8% used data in .xml format.

When asked how important they regarded good data management to be for their research, an encouraging 64% responded that it was 'essential' and that their 'research would suffer significantly if [their] data were not properly managed' (other possible responses were 'important', 'helpful up to a point', and 'not important'). Given the current lack of university assistance for researchers with research data management (fewer than one in four could remember ever having received any information previously from the university relating to research data management, whether via training, induction sessions, online support material, leaflets or any other means), this suggests that there is significant demand.[10]

Awareness of existing research data management infrastructure at the university was disappointing. It is unsurprising that a large majority had never heard of DataStage given that the software has not yet been generally launched, but the fact that over 80% were unaware of the university's central research data management website after two years suggests that greater promotion is required, and also highlights the difficulties of communicating with such a diverse group within the institution. Almost half of the survey respondents were not even aware of the university's long-standing central back-up and archiving service.

On the other hand, it is apparent that when research data management services are explained to the university's researchers, demand is there. Where the in-development DataBank service was described, in an attempt to gauge likely uptake over 2013-14, it was the potential volume of data that people wished to store, rather than their obliviousness to the service, that was worrying. 23% (73) of the survey respondents said that they would like to use the service during its first year, whilst a further 46% indicated that they might, but weren't yet sure. If all those who said that they wished to deposit at least data between May 2013 and April 2014 did so, we would need at the very minimum 46TB of effectively permanent storage just for data collected in that year, and (assuming respondents data deposit requirements were around the middle of each range category rather than the minimum[11]) probably well in excess of 200TB. Furthermore, there are doubtless many more researchers within the university working with data who did not respond to our survey, and whose needs are not therefore captured. Even with the very good response rate we got, the 314 respondents only represent a small proportion of the almost 10,000 researchers, research support staff, and postgraduate research students at the university in 2012. Given the huge potential demand, the university would almost certainly need to limit the use of

[10]With regards to training, the Damaro Project undertook a second survey recently to measure more precisely the demand for particular training content amongst researchers in the Medical Sciences and the Mathematical, Physical and Life Sciences, areas not well address in previous studies at Oxford. Our findings are summarized in a blog post available at http://blogs.oucs.ox.ac.uk/damaro/2012/11/21/damaro-survey-results-research-data-management-training-for-the-sciences/, whilst a spreadsheet summarizing the survey results is available at http://damaro.oucs.ox.ac.uk/docs/RDM%20for%20sciences%20-%20training%20survey%20results.xlsx.

[11] The survey question asked researchers whether they intended to deposit any data in each three-month period, and gave the options of depositing 'less than 10 GB', 'between 10 GB and 100 GB', 'between 100 GB and 1TB', and 'more than 1TB'.

DataBank to those researchers in receipt of UK Research Council funding in the first instance.

Ultimately, the research data management infrastructure we are implementing at Oxford is intended to add value by encouraging data re-use, and here it seems as though there is considerable potential, at least if current attitudes towards data sharing can be maintained or greater liberality encouraged. 30% of survey respondents reported that, after an appropriate embargo period, they would be happy to share most or all of their research data without any restrictions, and a further 40% would be happy to share some under these terms. For many researchers, ethical, legal or commercial restrictions would prevent them from sharing some of their data. Only 48% of respondents said that none of their data was subject to ethical of privacy concerns that might impinge on their ability to share it, whereas 43% said that there were commercial or legal restrictions that might prevent them sharing at least some of their data. Perhaps understandably, some researchers were reluctant to share parts of their data completely publicly, with only 32% saying they would be happy to share all of their data without any indication of how the data would be used, and 25% saying that they would only willingly share their data in its entirety with colleagues or collaborators, but the survey did not suggest insurmountable barriers to data sharing amongst researchers in general, and attitudes to data sharing are often nuanced according to the precise circumstances of the research in question.

If attitudes to data sharing were broadly positive on the 'supply side', there seems to be plenty of demand for shared data on the other. 37% of survey respondents said that they had been inspired to undertake new or additional research as a direct result of looking at data that has been shared by researchers in the past, and this figure excludes those who responded that 'seeing existing research data may have played a part in shaping new research ideas, but has never been a particularly significant factor' (22%). Of those who said that they had been inspired by shared data, 15% said that they had found their inspirational data in a data repository. It will be interesting to measure any increase in this figure as institutional data repositories begin to become more widely available.

Whilst the Research Data Management Survey was undertaken principally for benchmarking purposes, its findings provide evidence of the potential demand amongst researchers for improved training, long-term preservation and data management infrastructure more broadly. This is likely to give further weight to the case for investment in such infrastructure at the institutional level, on top of the recent pressure from research councils in the UK for such investment. The funding requirements of Engineering and Physical Sciences Research Council (EPSRC) in particular has raised research data management up the agenda over the last year, due to their decision to hold the university rather than the individual researcher responsible for meeting the requirements of their data management policies.

Sustainability

In July 2012, a Research Data Management Working Group was formed at the University of Oxford to build a business case for continuing investment in developing the institutional infrastructure. The Working Group consists of senior staff from the Research Services, IT Services, and Bodleian Libraries, as well as others who have

 doi:10.2218/ijdc.v2i8.287

been involved in past projects to develop data management, and is focussed on ensuring that the infrastructure exists to enable researchers can comply with the new data management policy and that Oxford continues to meet the criteria required to receive funding from the major funding bodies.

A number of possible business models have been considered for the various components of the research data management infrastructure at Oxford, although at the time of writing there are still issues to be addressed.

The most obvious model would be to fund the core elements of the infrastructure (such as DataBank and DataFinder) directly from the university budget or as an indirect cost in research funding bids. Whilst these options are relatively simple, given the lack of rapid funding growth and the already proportionally quite high indirect costs levied by the university, it is far from clear whether either proposal would be successful. Where possible, therefore, we are trying to set up services on a cost-recovery basis. For such instances, the cost of sustaining the service is paid by the researcher, or rather by the funder who is paying for the research and setting the data management and preservation terms. The RCUK Common Principles on Data Policy states that:

> 'It is appropriate to use public funds to support the management and sharing of publicly-funded research data. To maximise the research benefit which can be gained from limited budgets, the mechanisms for these activities should be both efficient and cost-effective in the use of public funds.' (RCUK, n.d).

It is not yet entirely clear exactly how each of the funding councils will apply these principles in practice, but it seems reasonable to expect that bids for data management services that will be used during the active duration of a project will be given a fair hearing. By contrast, it is becoming clearer that most funders are not happy to award funding to manage data after the proposed project has finished.

One complication to the 'funder pays up front' model of research data management sustainability is that not all research is externally funded. Many researchers at Oxford and elsewhere conduct significant quantities of valuable research outside of funded research projects, particularly in the humanities. Some elements of a research data management infrastructure are likely to cost very little in proportion to the other costs of research a funder might normally pay, such as staff salaries, but may nevertheless be prohibitively expensive for an individual to bear. One can be fairly sure that research funders would not be happy to pay a cross-subsidy to finance the use of infrastructure for researchers besides those they have chosen to fund.

An additional problem encountered when trying to persuade researchers (or indeed anyone else) to use centrally-provided tools and services is that there are often alternative solutions that have no dependency on wider university staff. Not only can such options appear tempting because the researcher feels they have greater control over how processes are managed (although this may in fact be illusory), but they can also mask costs in such a way as to seem much cheaper, and therefore better value than centrally-provided university infrastructure. This is generally because services offered centrally are (or ought to be) priced so as to include all cost elements. 'Local' solutions often hide costs from the ultimate user.

A case in point is the Online Research Database Service (ORDS) that Oxford is developing. Often, when researchers realize that their research data would benefit from being structured in a database, their first instinct is to start building a database with whatever database software package came with their computer. In the past this has typically been Microsoft Access. This appears to be a 'free' option. A separate survey run as part of the VIDaaS Project revealed that half of the 34 researchers questioned had spent nothing on hardware, 41% had spent nothing on software, and five out of the seven projects who were already making their data available via a website said they had spent nothing on hosting. This suggests that these costs were not visible to the researchers, not that they didn't exist, and were being paid by somebody, somewhere (Wilson, 2011).

The ORDS service, on the other hand, has been costed to include all the hosting costs, include hardware depreciation, power and cooling, systems maintenance, and the staffing costs incurred in providing the service and updating the software. As a consequence it looks relatively expensive.

In truth, the comparison between institutionally-provided services and 'local' solutions is often unfair in the first place. Many researchers begin to collect and structure their data using software installed locally on their machines, but then find that this software does not offer quite the functionality that the researcher finds they need as their data grows or the project expands. Collaboration in particular often seems to cause difficulties: back-up regimes are frequently ad hoc, sharing data is not straightforward, and depositing data for long-term preservation and re-use becomes a major problem at the end of a project, as the data is not documented properly and is likely to be unintelligible or confusing to anyone not closely involved in assembling it. This can discourage deposit. All of these problems are either avoided or heavily mitigated by the ORDS. The VIDaaS survey revealed that 38% of researchers working with databases spent less than a day researching database software and/or consulting with technical staff before deciding how to proceed.

The research data management projects that Oxford has been involved with have looked at a number of sustainability models for the various tools and services developed so far. These have ranged from an open source software model for lightweight and easy-to-set-up solutions such as DataStage, through embedding training in activities which already have established funding streams, to designing full cost-recovery services, such as we are doing with ORDS. Particular attention has been paid to the ORDS because it constitutes a fairly 'heavyweight' solution that will be of interest to some research projects but not all, making it a good candidate for costing into research bids as a 'Small Research Facility' (in a similar way as time on a telescope or using a high-performance computer might be costed). Given that the tools and services developed at Oxford are also likely to be of interest to other research-intensive universities, we have also begun negotiations with commercial companies to gauge interest in providing these services at a national or international level, which could unlock significant economies of scale.

All of the software being developed by the University of Oxford for its research data management infrastructure is available under an open source licence. It can, therefore, be freely used by other institutions. However, the degree of customization the software will require to be successfully integrated into that institution's own infrastructure is likely to vary. DataStage is designed to be a lightweight component

 doi:10.2218/ijdc.v2i8.287

that individual researchers can install at their local project-team level to provide a shared file store for the team, with access controls and integrated metadata-collection mechanisms. The software may be hosted on cloud, central, or local networked file storage, as the principal investigator chooses. On request, or at the end of a research project, the data can be pushed to a Sword2-compliant repository, such as Oxford's DataBank. As such, there is little service to sustain in terms of DataStage itself. Provided the software is updated and obtainable, and researchers are aware of its existence, it does not require significant future investment, although building an open source development community around the software would be of substantial benefit. This, in itself, is one sustainability option, although even in this instance some continued effort is required. The developers see the software as potentially a 'loss leader' to drive dataset submissions to the DataBank service, which will require permanent staffing.[12]

The Database-as-a-Service (DaaS) software which underpins the ORDS is also published under an open source licence, but the sustainability plan for the service within Oxford is more elaborate, as a researcher could not simply install the software and start using it by themselves. It is designed very much as a centrally-provided and staffed service, closely integrated with the (cloud) data storage, back-up, and authentication mechanisms of the university. As such, it is expected to pay towards the costs and maintenance of the services upon which it depends. Further details about the costs, potential benefits, and pricing model for the ORDS can be found in Appendix A of the VIDaaS Project Final Report (Wilson, 2011).

The means by which we hope to sustain the research training and support materials for data management in the university is different again. Whilst these materials are similarly being licensed under a Creative Commons licence to encourage re-use and adaptation, our approach to ensuring their continued life within Oxford is to embed the materials within already funded research skills training programmes in departments and faculties. There is a central research data management 'hub' for the university, which is currently being managed by staff within the Research Services team on project funding, and at some point a business case will need to be made to ensure there is some sort of formal accountability for this, but similar cases ought not to be required for each departmental or divisional training programme, where responsibility for such training is already clear. It is also appropriate that research data management training be undertaken in a large part at a disciplinary level, as experience suggests that researchers relate better to such training when they have examples which they can relate to, rather than when data management is dealt with at an abstract high level. For an example of how the local expenses of running research data management training courses can be outweighed by their benefits, see the Sudamih Project Final Report (Wilson, 2011b).

The DataBank and DataFinder services are arguably more challenging than other elements of the planned infrastructure. The University Research Data Management Policy does not mandate specifically that researchers must put information about their data outputs in DataFinder, but it does state that the university has responsibility for 'providing access to services and facilities for the storage, backup, deposit and retention of research data and records that allow researchers to meet their

[12]See Dataflow Sustainability Plan, v.2.0: http://www.dataflow.ox.ac.uk/index.php/project-reports/project-sustainability/175-sustainability-plan-v20

requirements under this policy and those of the funders of their research' (University of Oxford, 2012). If all researchers with data outputs are obliged to create records for those outputs in DataFinder, or to place the data itself in DataBank (should there be no more appropriate repository), then charging them for the privilege may prove controversial. It will almost certainly act as a disincentive to researchers to declare that their projects have produced data outputs, and that would not be a good outcome for the programme. Arguing the case for central funding therefore seems sensible as far as these services are concerned, using the university's own stated commitments and the potential loss of funding income from Research Councils, such at EPSRC, as justifications. Whether this is an argument of sufficient strength to guarantee sustainability during a period of particularly tightly-controlled spending commitments remains to be seen.

Findings

Whilst this paper has in a large part described work which is on-going and incomplete, a number of conclusions can be drawn from what has been studied thus far.

Firstly, if perhaps rather obviously, it is important to understand the needs of the researchers whom the research data management infrastructure is intended to support. Whilst the primary driver behind developing the infrastructure may be the mandates of the funding agencies, if the researchers do not use the infrastructure in practice, compliance is merely theoretical.

A second point to bear in mind is that central services need to be assiduously promoted to researchers, either directly or indirectly via academic divisions and faculties, if any meaningful level of awareness, and consequently use, is to be attained. It seems that researchers tend to pay more attention to developments in their disciplines than to what their institutions are offering them. The ongoing costs of disseminating information about services need to be included in costs/benefits analyses.

Thirdly, assessing the potential uptake of a service is useful, not only with regards to capacity planning, but also when developing a business model. Institutional services, such as the ORDS, often have quite high fixed costs relative to variable costs (mostly due to the need for ongoing development and minimum staffing requirements). If a service is intended to be run on a cost recovery basis, the risks of setting prices either too low to recover costs in the event of low take-up, or setting prices too high to attain the take-up required, become greater.

Finally, our work suggests that business models are better made at the level of the individual component rather than for an infrastructure as a whole. The manner in which each component of such an infrastructure is likely to be used, at what point in the research lifecycle, by whom, and with what benefits, is likely to vary considerably. Therefore, it makes sense for component costs to be borne by the part(s) of the institution that acquire the greatest benefit from that component. This approach also ought to make it easier to add (or detach) new infrastructure components in the future, as the number of technologies available to universities increases, and knowledge of best practice improves.

References

Oxford Gazette. (2012). Notices: Policy on the Management of Research Data and Records. *Oxford Gazette, 143*(500). Retrieved from http://www.ox.ac.uk/gazette/2012-2013/4october2012-no5000/notices/#85831

Research Councils UK. (n.d). Common principles on data policy. Retrieved from http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx

Starr, J. et al. (2011). DataCite metadata schema v.2.2. Retrieved from http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf

University of Oxford. (2012). Policy on the management of research data and records. Retrieved from http://www.admin.ox.ac.uk/media/global/wwwadminoxacuk/localsites/researchdatamanagement/documents/Policy_on_the_Management_of_Research_Data_and_Records.pdf

Wilson, J.A.J. (2011). VIDaaS project report. Retrieved from http://vidaas.oucs.ox.ac.uk/docs/VIDaaS_FinalReport.pdf

Wilson, J.A.J. (2011b). Sudamih project final report. Retrieved from http://sudamih.oucs.ox.ac.uk/docs/Sudamih_FinalReport_v1.0.pdf

Wilson, J.A.J., Martinez-Uribe, L., Fraser, M.A. & Jeffreys, P. (2012). An institutional approach to developing research data management infrastructure. *International Journal of Digital Curation, 6*(2), 274-287. doi:10.2218/ijdc.v6i2.203