The International Journal of Digital Curation
Issue 1, Volume 2 | 2007

Attitudes and Aspirations in a Diverse World: The Project StORe Perspective on Scientific Repositories

Graham Pryor,
Project Manager, StORe,
University of Edinburgh

November 2006

Summary

Project StORe was conceived as an initiative to apply digital library technologies in the creation of new value for published research. Ostensibly a technical project, its primary objective was the design of middleware to enable bi-directional links between source repositories containing research data and output repositories containing research publications. Hence, researchers would be able to navigate directly from within an electronic article to the source or synthesised data from which that article was derived. To achieve a product that directly reflects user needs, a survey of researchers was conducted across seven scientific disciplines. This survey exposed the spectrum of cultural pressures, preferences and prejudices that influence the research process, as well as a range of practices in the production and management of research data. Aspects of the research environment revealed by the survey are considered in this paper in the context of repository use and, more broadly, the requirements, roles and responsibilities necessary to good data management.

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.

Introduction

When preparing a team of recent postgraduates who were about to embark on a survey of repository users, it became necessary to equip them with a glossary of terms for use when explaining their mission. Whilst our project's focus was upon the present and future functionality of source and output repositories, in the specific context of their potential interoperability, describing the gamut of issues that might arise during the survey required some definition of digital curation, which led me to the Digital Curation Centre's web pages,[1] whereupon inspired by a definition given on one such page I settled eventually on the following definition:

> "The actions needed to maintain digital data and other digital materials over their entire life-cycle and indefinitely for current and future generations of users. These actions will include not only the processes of digital archiving and preservation but also all of the processes that are essential to good data creation and management, as well as the capacity to add value to data to generate new sources of information and knowledge."

Without stretching a point, the second sentence firmly attaches the concept of digital curation to the aims and anticipated deliverables from Project StORe.

The principal aspiration of Project StORe, a two-year JISC-funded[2] project that concludes in August 2007, is to invest new value in the reports and papers that represent the intellectual products of academic research. Our proposed route to achieving this is the provision of bi-directional links between source and output repositories[3]. The benefits from such a linkage are anticipated to be improvements in opportunities for information discovery and in the curation of valuable research data. In this context the Project StORe programme is directly aligned with the ethos of digital curation.

The StORe Survey

A key deliverable from Project StORe will be a set of pilot middleware (Figure 1) designed to demonstrate the function of bi-directional links between source and output repositories. This middleware will be developed to meet the specific needs of the e-research community, as defined from a survey of the behaviours of researchers within the seven scientific disciplines represented by the project (archaeology, astronomy, biochemistry, biosciences, chemistry, physics and social sciences). The challenge, therefore, is to produce and maintain dependable long-term links between sets of data that are essentially volatile and subject to local rules governing data management and archiving. The StORe survey[4], an online questionnaire and a series of interviews undertaken between February and July 2006, confirmed that the environment in which

[1] http://www.dcc.ac.uk/

[2] For further information see http://www.jisc.ac.uk/index.cfm?name=programme_digital_repositories

[3] For the purposes of this project, **output** repositories are defined as those that contain published articles, texts or data objects. The contents of an output repository will typically include publications at a pre- or post-refereeing stage, working papers, research reports and PhD theses. **Source** repositories contain the source or primary data produced during a programme of research, and generally comprise the origins from which research publications will be developed.

[4] For details from the StORe survey, including the discipline reports, see http://jiscstore.jot.com/SurveyPhase

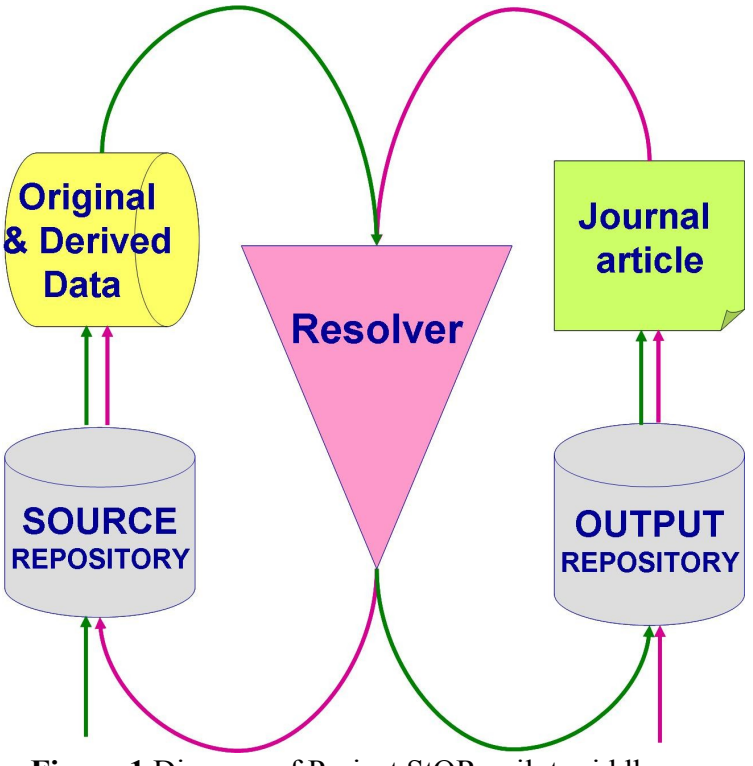

Figure 1 Diagram of Project StORe pilot middleware

we were working is diverse, sometimes colossal in terms of the data generated, and in many cases lacking standardisation or regulation.

It should not be inferred from this observation that the results from the survey were adverse in every respect. International strategies for the deposit and preservation of data were found to be embedded in several of the disciplines surveyed, these being particularly sophisticated in astronomy and the biosciences. We also found that the dual deposit of data and publications is a concept that has already become the norm for many of our survey respondents, and in all disciplines surveyed we encountered an awareness of the critical issues surrounding the appropriate assignment of metadata. Above all, when seeking views on the constraints that might govern access to data, whilst there were serious concerns over the potential risks to data ownership, the scientific research community was found to be permeated by a genuine desire to contribute to the global wealth of knowledge, and StORe's main proposition, to enable bi-directional links between source and output repositories and thereby encourage further access, was endorsed by 85% of those who responded to our questionnaire.

Nonetheless, our survey did establish that there are significant problem areas to be addressed amongst the current practices employed for data generation and administration. Cultural and organisational barriers prevail in all disciplines, which serve to deter the deposit of research data in repositories, and an inherent culture of self-sufficiency in the generation and organisation of data militates against what might be viewed as prescriptive intervention by knowledge management professionals. We found too that many researchers held serious reservations towards the voluntary deposit of research data in open access source repositories, and institutional output repositories are simply not on the agenda of most academic researchers.

Examining the Problems

In this article I will address these problem areas as portrayed within five distinct aspects of the research data experience: data generation, the use of repositories, the assignment of metadata, data ownership and the provision of support. This examination will identify some of the key modifiers to ambitions for effective data curation that were identified during the StORe survey, as well as some practical approaches to tackling them.

The persistent message we received concerning the nature of raw data generated from scientific research is that it can often be huge. Interviewees from the physics discipline proudly informed us that they may generate raw data sets of the order of petabytes (10^{15} bytes). Further, whilst many will store their data using well known formats and analyse their data using standard commercial software, it is not uncommon for physicists to use more obscure formats dictated by the nature of their particular research processes and to write their own analysis software. A typical observation by one of our physics interviewees was that it is:

> "usually impractical to access High Energy Physics experiments primary data. The expertise needed to understand it and the large size are usually prohibitive" and "the processing of data is something that can almost invariably only be done by groups very closely involved with the production of data."

A similar picture emerged within other disciplines, with particular concern in the biosciences that raw data could be misinterpreted, with disastrous consequences at a human level, and bioscientists expressed some reluctance to make their raw data available unless it could be accompanied by advice describing how they were produced, the laboratory conditions and the methodology used. A thread of consensus connecting the disciplines was articulated by the physicist who expressed a preference for accessing others' processed data rather than raw data, since this would have "taken into account measurement uncertainties and artefacts of their apparatus".

Archaeology provided an exception to this perspective. Both the questionnaire and the interviews confirmed that whilst archaeologists tend to produce highly complex data sets, often with many different data formats, the incidence of large file sizes is rare. The following bar chart (Figure 2) shows the range of file types used by the archaeologists we surveyed. One set of archaeology research data can include many if not all of these, and we found that archaeologists tend to produce more maps, plans, plots and images than other disciplines. They are sometimes linked in the form of a Geographical Information System (GIS), but this is not always the case, meaning that access and the curation of research data in archaeology presents a different challenge to that described for physics and the biosciences, requiring the maintenance of links between all the constituent elements in a heterogeneous research data 'package'. Where a GIS is employed, such a system would not necessarily represent a static collection of data, but one that may be added to and altered continually throughout the life of a project, even after deposition of the project archive in a repository.

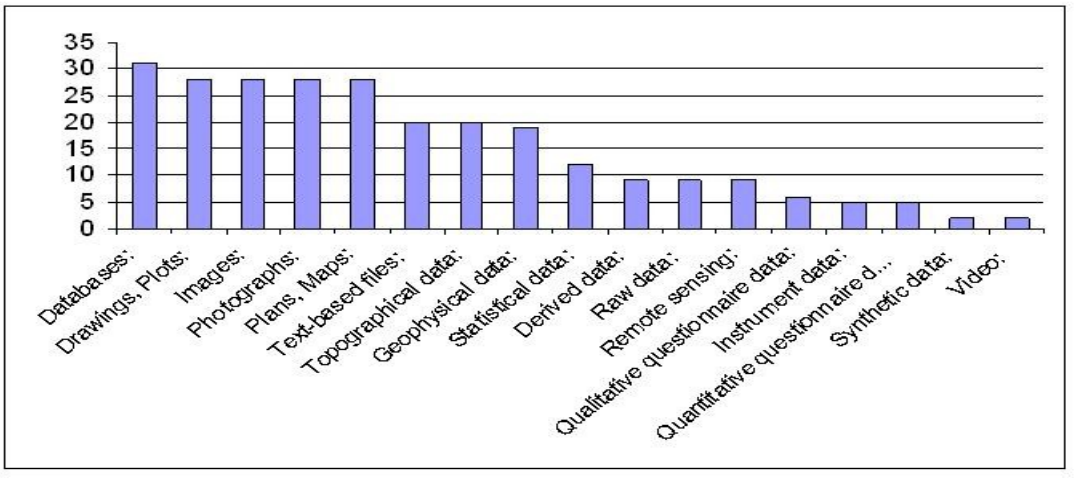

Figure 2 Graph showing range of file types used by archaeologists surveyed by Project StORe

Researchers in chemistry provided further examples of this need to sustain links between clusters of complex data sets. They also maintained that the way they have to organise their data provides a barrier to the interrogation of their unmediated research output by anyone from outside their project. Similarly to researchers in physics and the biosciences, research chemists often produce large data sets, but the majority of StORe's chemistry interviewees, most of whom belonged to the computational/theoretical chemistry community, indicated that they do not produce raw data in the same sense as other scientists. Instead, they are involved in the development of methods for testing how molecules behave in certain conditions, and although they produce data in the form of calculations and measurements testing they tend to apply their methods to other researchers' published outcomes. The data produced by the computational/theoretical chemists we interviewed is characteristically organised according to a tree structure, with many files stored in numerous individual subfolders. Access to this data by anyone from outside a specific project and by individual file was described as pointless. One of the interviewees noted:

> "First of all it would have to be everything associated with that compound. There is no point having an NMR without a picture of what it is. Then it's useful to have a synthesis scenario and say 'oh that could fit with that but I want proof' and then that really is a paper. You know you can waste a lot of time trying to follow what people have done before that isn't properly published and never have worth. It's not always, but is it worth the risk of wasting too much of your time?"

In such a situation, initiatives to preserve and manage valuable data must first identify the discrete components of a set of data, as well as who owns which parts of the set, which components are fixed and which are dynamic. In the light of these conditions, as well as in the face of physics' legendary petabytes, when considering

where to attach links from output repositories the alternative and more attractive option may be to concentrate on the higher level of order that is processed data. Although further changes to processed data are to be expected, one is likely to encounter a more accessible structure and the defining framework of interpretation.

An even more fundamental challenge revealed during the StORe survey arises from the data storage behaviours of the majority of academic researchers: the overwhelming practice amongst the researchers we surveyed is to store their data on standalone PCs or on shared network drives accessible only by project partners. A full 70% of source data in the biosciences was estimated not to be networked, and whilst high energy physicists are participating in the development of a global *Computing Grid*[5] to store and process large amounts of data, our survey indicated that others in the physics research community appear to keep their data on CDs and the hard drives of individual PCs. It seems almost facile to observe that if we do not know what data have been generated and retained or discarded, where data are stored, or what options there might be for access, then we can neither establish links nor practice data curation. One might imagine that one solution to this conundrum is the deposit of research data in source repositories, but amongst our survey constituents we discovered a limited awareness and understanding on the subject of data repositories that was not confined to any single discipline. Indeed, the extent to which scientific researchers use technology-enabled data management processes and services in general was found to vary considerably.

Nonetheless, across the seven disciplines surveyed there are well established source and output repositories, with source repositories in most disciplines catering for very specific data types. From comments made by interviewees, astronomy is particularly well served, one respondent remarking that he is "very happy with what we have in astronomy via Vizier, NED and Simbad. Please don't mess with them for the sake of some aesthetic global model." The Archaeology Data Service[6], the UK's principal digital data repository supporting research, learning and teaching in archaeology, was applauded for its careful reflection of discipline requirements, although the diversity of interests within the discipline results in a wide variety of source data being sought, with only 55% of those who had submitted data to a source repository having done so with the Archaeology Data Service. Social Sciences was another of the project's subject disciplines to portray a broad range of interests, and we found that an inclination to deposit in the UK Data Archive, the UK's largest collection of curated digital data in the social sciences and the humanities, varied according to sub-discipline, with sociology and economics exhibiting the greatest interest in using it. In the biosciences, the mandatory deposit of sequence data to GenBank or the Protein Structures Database is well established, but for other sectors of the discipline there were such references as "they don't yet exist in my field" and repeated calls were made for a repository dealing explicitly with metabolomics data. It was apparent that source repositories have often been developed within disciplines and collaborations in response to particular needs and inspirations, and usually sponsored by the research communities themselves. They are not the products of some evangelistic international movement to improve the management of knowledge; instead the development and use of data repositories represents on the whole an almost endemic culture of self-sufficiency within the academic research community.

[5] Worldwide Computing Grid, http://lcg.web.cern.ch/LCG
[6] Archaeology Data Service, http://ads.ahds.ac.uk/

In some disciplines this approach has led to the emergence of internationally sponsored strategies governing data deposit and preservation. Expressing support for the StORe programme, one astronomy researcher identified an outstanding need for easy links between source and output repositories "as part of a beginning-to-end framework that allows the tracking of source data through its entire path to publication"; another claimed that linking between source repositories is the main issue. Their aspirations are already being addressed by the Virtual Observatory (VO) Project, which is active in pursuing aims to apply standard protocols for connecting globally distributed collections.[7] The VO team at Johns Hopkins University is also working with the University of Chicago Press to consider the interconnection of output repository functions with the processes for submission of research papers for publication.

Astronomy may arguably be in the vanguard when it comes to matters of data curation and repository development, but it does not have the field to itself. Indeed, we found that the practice of dual deposit of data and research papers is already in place or planned within half the disciplines surveyed. Yet it was also evident that the culture of self-sufficiency driving these developments does not always translate into fully effective measures for accessing, organising, promulgating and curating data. During the StORe survey, respondents were asked directly for their perception of current and required functionality in source repositories and the following responses are but examples of necessary areas for improvement identified across the seven disciplines:

- Archaeology: Metadata are minimal. Data formats immensely variable - sometimes not at all obvious and with no accompanying metadata....
- Astronomy: Finding publications and data relevant to a specific need.
- Biochemistry: Older, error-prone, entries often remain in repositories even when newer, more validated, data are submitted.
- Biosciences: [There is a lack of] consistent formats, data standards.
- Chemistry: [We need] a search engine like SciFinder or Google Scholar.
- Physics: Uniformity of format for uniformity of data.
- Social Sciences: Knowing what is there - often the data are hidden behind unhelpful titles.

The contrasts exhibited in the responses to this question are more marked within disciplines than between them. Whilst one chemistry researcher announced "no problems here - we have created our own repository, to our own design", at the same time 65% of respondents from that discipline admitted never to have used a source repository.

[7] For a description of the Virtual Observatory project see http://www.arl.org/sparc/meetings/ala06/HanischPPT.pdf

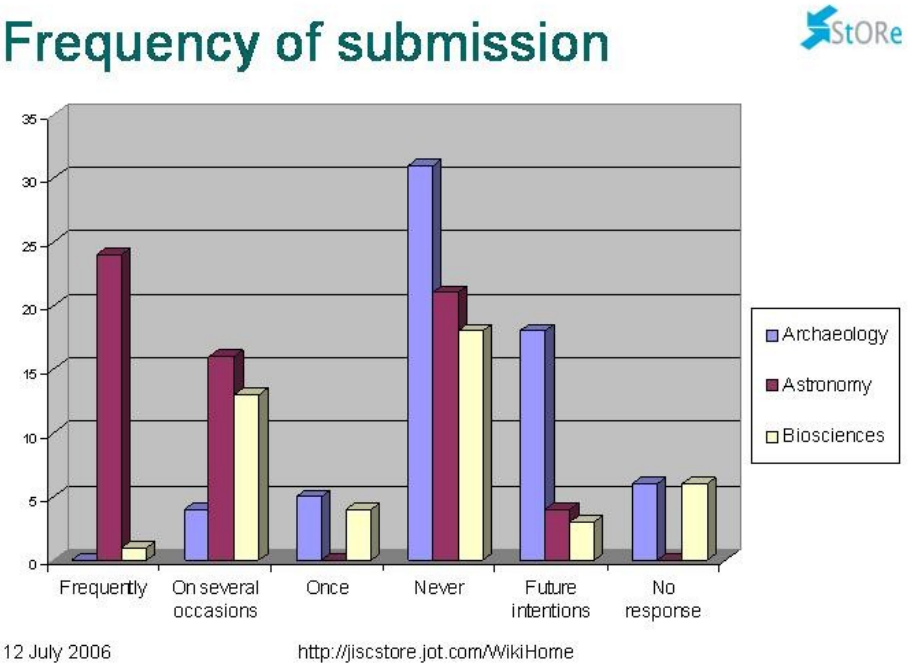

Figure 3 Graph of comparison among deposit rates for three of the subject disciplines

All respondents to the StORe questionnaire had been asked to indicate whether and how often they submitted data to certain source repositories believed to be familiar to active researchers within the subject disciplines. The list included the two repositories having mandatory deposit of sequence data and these scored relatively highly, with 51% of respondents from the biosciences indicating that they submitted data to Genbank and 47% to Protein Structures; but as can be seen from the graph, which compares the deposit rate for three of our subject disciplines, astronomy scored the highest of all through deposit in an extensive list of repositories located around the world. In contrast, whilst 45% of archaeology respondents claimed to submit to the Archaeology Data Service, their rates of submission suggested a considerably lower level of commitment, and when we assessed all seven disciplines together, two thirds of the total number of responses showed zero deposit.

When considering the deposit of research papers in output repositories, in all of the disciplines surveyed the use and apparent demand for institutional repositories were found to be negligible, and except where there is an institutional mandate, deposit in them is noticeably limited. Alternative global resources, frequently provided by publishers, often having high profiles and the benefit of relatively sophisticated structures, were described as preferred information sources or modes of dissemination. Strong scepticism was expressed concerning the quality of metadata used in institutional repositories, as well as criticism of user-unfriendly interfaces, inadequate search functions and a perceived lack of protection for intellectual property rights. Opportunities for institutions to pursue data management strategies at a corporate level would therefore seem to be inhibited, although initiatives to link institutional repositories, sharing a common structure and offering simple search features, should be expected to change the temperature of this otherwise cool reception. Interestingly, no reference was made by our survey respondents to the likely impact upon institutional repositories of future demands from the remodelled Research Assessment Exercise (RAE), which in due course will represent another agent of change.

Whilst in some instances the deposit of data into a nominated repository is a condition of grant, and in the case of the biosciences the deposit of sequence data to worldwide repositories is mandatory, we found that views about sharing or providing access to research data are polarised within almost all disciplines, the least sensitive being astronomy and physics. As observed by a social sciences respondent, researchers' attitudes to enabling access to their data will depend to a large extent on whether they are behaving as producers or users of data, with producers concerned to protect their endeavours from predatory access to source repositories by their competitors. This concern was found within each of the disciplines surveyed, and if there is a serious risk that hard-won research results might be appropriated and used by others to promote their own careers, source repositories can never be expected to offer a comprehensive collection of well organised data. Genuine dependency upon the ownership and safeguard of one's research output has to be recognised since, whilst most of our respondents did not necessarily rule out the deposit of their data in source repositories, even open access repositories, they will need to be convinced that the crucial provision of robust methods of protection are in place before they consent to deposit the fruits of their labours. Of course, protection and preservation will go hand in hand as tenets of digital curation.

Metadata, its selection, assignment and criticality of purpose proved to be the topic upon which our survey respondents achieved greatest accord, and it was encouraging to discover that in all the disciplines surveyed there is an awareness of the importance that must be attached to the appropriate assignment of metadata, if only to meet the simple demands of access and retrieval. When, during the StORe survey researchers were asked to suggest improvements to source repositories, better metadata or features that depend on high-quality metadata functionality ranked amongst the highest. Yet this topic also provided further examples of serious inconsistency both within and across the disciplines. It seems that not only is there a body of researchers who have still to grasp the purpose and importance of metadata but, where the need for good metadata is understood, this does not necessarily translate into the sufficient use of standard structures. The assignment of metadata was in too many areas found to be ad hoc, and often given consideration only in that final phase of a project or process when data are being saved. Consistent with the general culture of self-sufficiency, we also found that the self-assignment of metadata by individual researchers is commonplace, two thirds of our respondents confirming that they decide which terms to use, sometimes but not always involving reference to standard thesauri or schema. Of even greater concern was that almost one third of our survey respondents either believed no metadata were being assigned or did not know at which stage assignment took place.

It is not unreasonable to suggest that the origins for this prevailing situation are to be found in the working culture of academic research, where an admixture of self-reliance and the relentless pressure to deliver can relegate certain organisational aspects of the research information lifecycle. Certainly, the correct assignment of metadata was openly regarded as demanding, both intellectually and in terms of the time required and, where it is available, reliance on the automatic generation of metadata was found to be preferred to more laborious methods of manual assignment; although even where a significant volume of metadata generation is automated, as in the astronomy and biosciences domains, the need for improved and universal standards was acknowledged. Interestingly, so too was the need for assistance from specialists

in developing and administering metadata: across all the disciplines, a clear link was identified between the condition of metadata used and the level of support provided by information or data specialists.

A changed approach to the use of metadata appears to be the key to a range of issues encountered during our survey, made more urgent by the changing nature of research itself. Where complex research output has the potential for broad cross-discipline application and development, most notably in the biosciences, the need for some means of interpretation of data was cited as critical, and the role of metadata in enabling such features is understood. A mechanism for the constant review and update of metadata is also regarded as necessary to account for the rate and process of new discovery, and the provision of different metadata for different phases of the research information lifecycle (covering raw, processed and published data, and beyond) was clearly a concern for a number of our respondents. Providing solutions to such demands requires a fresh approach to the organisation and management of research data, one that will prescribe a new combination of skills and that represents a challenge for the specialist information/data intermediary.

Conclusions

If a large body of scientific researchers is bent on a do-it-yourself approach to data management, eschewing deposit in source repositories, inventing metadata on the fly, and concealing precious research data in the depths of their hard drives, what hope might there be not only for providing links from scholarly papers to their source data but also for the universal inculcation of sound digital curation? Of course I am exaggerating here to press my point, but the results from the StORe survey do imply that a step change is necessary, and one change that could offer a solution requires the reorientation of the traditional research team with the introduction of a new role for the information intermediary. Whilst a considerable majority of the StORe respondents do not seek assistance in the acquisition and management of data or in their use of repositories, typically providing us with such statements as “it’s my responsibility” or “the university has assigned a librarian to our department to help with searches, but I have not used her services”, it was nonetheless apparent that an unfulfilled role did exist in the provision of assistance with metadata and preservation matters. These are not skills that one should assume belong automatically to the scientific researcher, and a possible solution would be to embed in research projects a new cohort of staff having a blend of expertise in data creation and management, including metadata, and an appropriate level of subject knowledge. I am of course reminded that such an intervention may be regarded with suspicion, but what I am suggesting is a merger of skills, discipline-led, not the substitution of responsibilities. Further, the presence of information or data intermediaries embedded within projects could also provide a source of reassurance on other more troublesome aspects of data management, including the issues of rights and access, and not forgetting the creation and maintenance of long-term links between repositories! There would of course be a cost, giving rise to renewed arguments against research overheads, but when up-front investment in data management is measured against the longer-term aspirations for data preservation and exploitation it must appear slight. After all, have we not already defined a shared goal of researchers and data curators alike?: to add value to research and research data through the generation of new sources of information and knowledge.