Book Section:
Lewis, M.J. (2010) Libraries and the management of research data. In: McKnight, S, (ed.) Envisioning Future Academic Library Services. Facet Publishing , London , pp. 145-168. ISBN 978-1-85604-691-6

Libraries and the management of research data

Martin Lewis
Director of Library Services and University Librarian, The University of Sheffield

Chapter to appear in:
Envisioning Future Academic Library Services
Initiatives, ideas and challenges
Sue McKnight, editor
ISBN: 978-1-85604-691-6
Publication date: March 2010

Libraries and the management of research data

Martin Lewis
Director of Library Services and University Librarian, The University of Sheffield

Introduction

Perhaps the starting point for any discussion about libraries and research data is to ask whether managing data is actually a job for university libraries. The answer to this question is a straightforward yes and no. Yes, in the sense that data from academic research projects represents an integral part of the global research knowledge base, and so managing it should be a natural extension of the university library's current role in providing access to the published part of that knowledge base. No, because the scale of the challenge in terms of infrastructure, skills and culture change requires concerted action by a range of stakeholders, and not just university libraries.

This assessment, from the perspective of the United Kingdom (UK) in 2009, is not a prescription for inaction on the part of university libraries, however. On the contrary: libraries have a key role to play in developing both the capability and capacity of the higher education sector to manage research data assets. Some of them are already doing it; and, as for the rest of us, we need to take steps to understand the landscape even if we lack the resources to make immediate progress locally.

As with many emerging areas, the vocabulary of research data management is still evolving. In this chapter, by "research data management" we mean the storage, curation, preservation and provision of continuing access to digital research data, in other words most of the processes in the centre of the Digital Curation Centre (DCC)'s Curation Lifecycle Model, as well as the lower half of the outer circle (Fig 1). Perhaps more simply, this is not just about the storage of data, which is how the subject is sometimes represented, and how the requirement to "do something" about research data is often manifested locally.

It's worth taking a little time to reflect on how the management of research data sits alongside the other relationships the modern university library has with its academic

community. Then we'll consider what the drivers are for investing time and effort in managing research data, before looking in more detail at what contribution university libraries can and should be making.

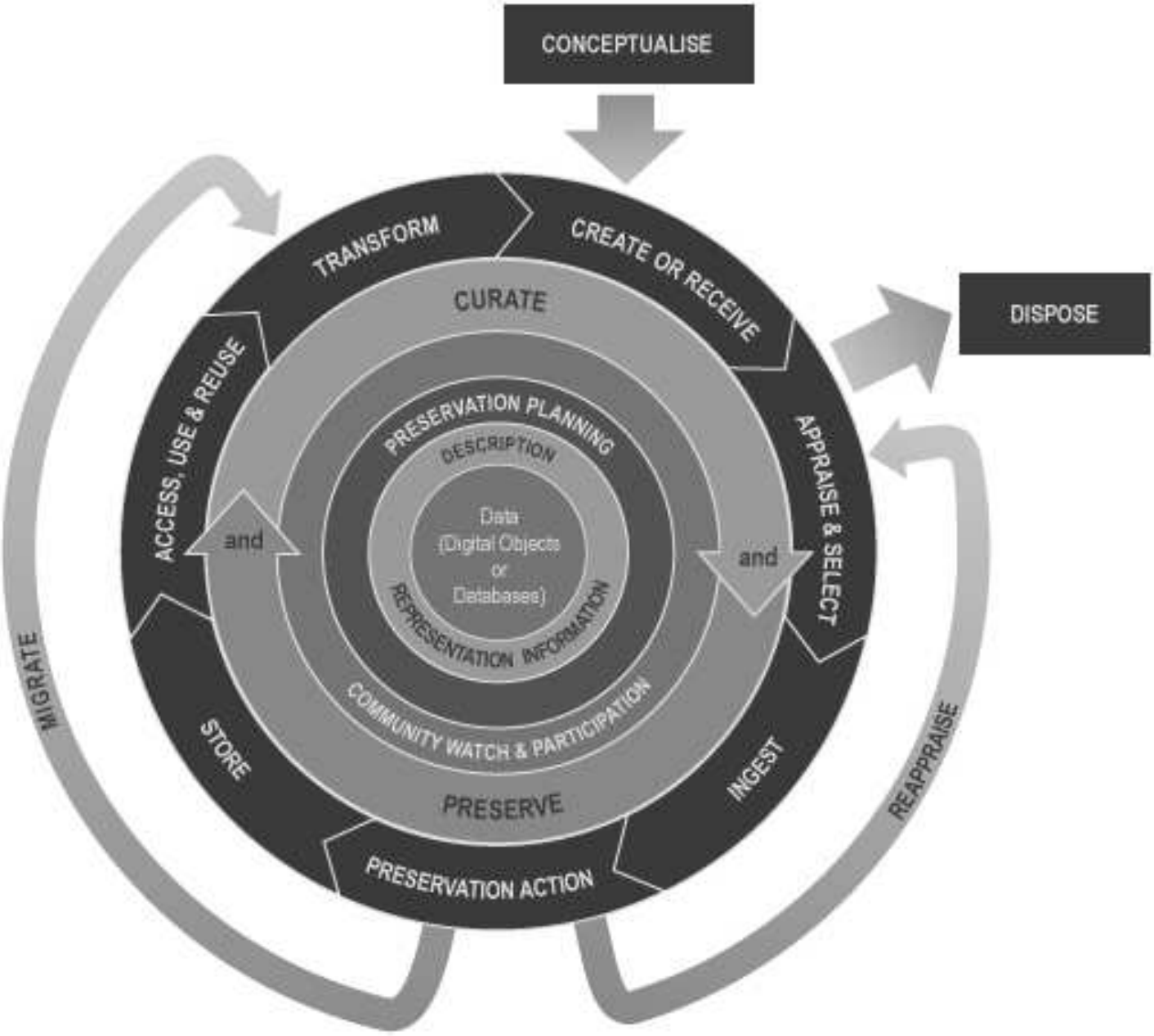

Figure 1 The Digital Curation Lifecycle model

As the other contributors to this book demonstrate, university libraries in many countries have in general been very successful at engaging with the rapidly changing learning and teaching agenda on their campuses (and off them as well). From information literacy to the development of bold new technology-enabled learning spaces, they have re-engineered the relationship with their teaching colleagues, improved the student learning experience, and raised expectations. And at the risk of over-generalisation, we might contend that (i) these successes have been evident in universities across the spectrum of research-intensiveness, from new universities without a significant research base, to the big research elite universities of the Russell Group in the UK; and (ii) that a similar general re-engagement with researchers has been notable by its absence, even in the Russell Group universities.

Despite libraries' progress over the last decade in transforming access to the research literature through provision of e-journals and resource discovery tools – and perhaps in part because of it – libraries have become more distant from their research customers, especially their STM (science, technology and medicine) research colleagues. The Research Information Network (RIN)'s report *Researchers' use of academic libraries* and their services (RIN, 2007) represents a valuable snapshot of the nature of the researcher-librarian relationship: it notes the decline in visits to the physical premises of libraries in recent years, especially by STM researchers, and the weak link in such researchers' minds between the digital content they use and the library's role in providing it. In the late 1980s and early 1990s, it was not unusual for larger university libraries to be conducting several thousand mediated online bibliographic searches per year on behalf of their researchers, the majority of them involving a detailed client interview, with the useful secondary outcome that the library liaison staff involved would have a good picture of the client's research. While no-one would suggest a return to mediated access to the research literature as a way of improving research liaison, not least since the size of the research coalface has increased enormously over the last 20 years, the challenge of re-engaging with researchers to understand their developing knowledge management needs is clear. And progress with the research data management task requires that this re-engagement takes place.

But why do we need to manage research data in the first place? Library managers contemplating multiple demands on limited resources deserve an answer to this question, even if it might seem redundant to the relatively small cadre of data managers in the workforce; and, moreover, they need to be able to articulate the answer in turn to university managers when discussing institutional approaches to the challenge.

The answer is in part a prosaic one: the volumes of data being generated by researchers are growing rapidly (there may be a case for using the word "exponential" accurately here, for a change), not least as a result of the increasing use of e-research tools (see the following section); and research funders are increasingly likely to require researchers to deposit their research data (research funders' policies on data deposit are now included in the SHERPA "Juliet" database[1] maintained by the UK Open Access project SHERPA).

More powerfully, the rewards of managing research data include significant potential benefits for academic research itself:

- The ability to share research data, minimising the need to repeat work in the laboratory, field or library
- Ensuring that research data gathered at considerable cost is not lost or inadvertently destroyed
- The retrieval, comparison and co-analysis of data from multiple sources can lead to powerful new insights
- The ability to check or repeat experiments and verify findings, particularly important amid growing national and international concern about research integrity
- New research themes – and in particular cross-disciplinary themes – can emerge from re-analysis of existing data or comparisons with new data: increasingly data may become the starting point for new research as well as representing an output from current research.

[1] http://www.sherpa.ac.uk/juliet/index.php

To this list of drivers should be added the public access argument, which is also deployed in relation to open access to published research papers: that society as a whole benefits from access to the fruits of publicly-funded research, a sentiment expressed in the Organisation for Economic Co-operation and Development (OECD)'s Principles and guidelines for access to research data from public funding, which states

> *"Sharing and open access to publicly funded research data not only helps to* maximise the research potential of new digital technologies and networks, but provides greater return from the public *investment in research."* (OECD, 2007)

Even those institutions in which research data management has not been actively discussed are likely to find it becoming more of a priority as researchers whose grants were made by funders with a requirement to manage post-project data outputs move towards the latter stages of their projects.

e-Research and research data management

Management of the data outputs of research projects is not a requirement that has just emerged in last few years: it is over 40 years since the UK Data Archive was established at the University of Essex, and many university libraries have long held collections of paper-based surveys and other data outputs. However, it is the growth of digital research data that has driven recent interest in long-term curation and storage. In the UK, the government-funded e-Science Core Programme, which ran for six years from 2001, has raised the profile of this issue, to the extent that research data management has sometimes come to be seen as a challenge exclusively linked to e-science or e-research (the term e-research is more inclusive of the non-science disciplines which are increasingly using the techniques and tools of e-science).

The e-Science Core Programme was administered by the Engineering and Physical Sciences Research Council (EPSRC) on behalf of Research Councils UK, and aimed to establish the toolkit – including infrastructure, middleware and documentation – to facilitate wider uptake of e-research. The seven Research Councils also established e-science programmes, with ringfenced funding, to promote e-science within their

disciplinary areas. The Core Programme also funded demonstrator projects to enable researchers to understand the scope and capability of e-research.

Announcing the eScience Core Programme in 2000, the then Director-General of the Research Councils, Professor Sir John Taylor, said:

> "e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it."

We can characterise e-research from the vantage point of nine years later as

- data-intensive: generating and often using large volumes of data
- collaborative: involving researchers across multiple institutions, and often transnationally
- Grid-enabled: using high-capacity networks and middleware.

Although data management was not directly addressed in the first phase of the Core Programme, the implications of large-scale e-research projects for data management were soon apparent. The term "data deluge" was used by the Core Programme's leadership to describe the challenge ahead (Hey and Trefethen, 2003). The Joint Information Systems Committee (JISC) also commissioned a report on the curation of eScience data (Lord and Macdonald, 2003) that made a number of recommendations about the need to develop national capability and capacity. It highlighted the role of the Digital Curation Centre, co-funded by the second phase of the Core Programme and by JISC, as a source of expertise and advice for the higher education sector, and made some trenchant comments about the need for a coherent national approach to the challenge:

> *"There is a lack of a government*-level, overall strategy for data stewardship and data infrastructure to which science administrators can refer, still less to support the researcher in *their evolving roles and duties with regard to data curation."* (Lord and Macdonald, 2003, p5)

The need for long-term investment appeared to have been recognised by the UK Treasury in its Science & innovation investment framework 2004-2014 (HM Treasury, 2004). This developed the concept of a national "e-infrastructure" to support world-class research and

innovation, and the Office of Science and Technology (OST), then part of the Department of Trade and Industry, was asked to lead on e-infrastructure. The OST set up a working group, with six sub-groups being asked to explore different aspects of the challenge. These were:

- Data and information creation
- Preservation and curation
- Search and navigation
- Virtual research communities
- Networks, compute and data storage
- AAA (authentication, authorisation and accounting), middleware, and DRM (digital rights management).

As can be seen, data management featured prominently in the work of the sub-groups. An opportunity to feed their work directly into the UK Government's 2007 Comprehensive Spending review was missed, however: and when the overarching report was finally published in 2007 (Pothen, 2007), it did not attempt to quantify the level of investment needed to develop and sustain a national infrastructure for the management of digital research data.

This brief historical overview of UK developments sets the scene for discussion of the UK Research Data Service feasibility study in a later section, an initiative in which higher education librarians have played a significant role.

The UK's e-Science Core Programme helped to get the UK into something of a leadership position in the early years of e-research. Since one of the key benefits of e-research is the facilitation of global collaboration, however, we should note that other countries have also been exploring and investing in e-research. The US National Science Foundation (NSF) has set out a clear vision for future investment in "cyberinfrastructure" (NSF, 2007a). Unlike the UK, it moved quickly to announce investment funds for digital research data curation through its DataNet programme, the call for which was issued in 2007 (NSF, 2007b). The call document sees a key role for what it terms "library and

archival science" in the new partnerships that it envisages for DataNets. Two DataNet projects have so far been approved: the Data Conservancy led by Johns Hopkins University Library, and the DataNetONE consortium led by the University of New Mexico. The National Science and Technology Council's Committee on Science set up an interagency working party on digital data in 2007 which has recently reported. This sets out a roadmap for a series of coordinated national activities, and includes the clear statement:

> *"*We envision a digital scientific data universe in which data creation, collection, documentation, analysis, preservation, and dissemination can be appropriately, *reliably, and readily managed. This will enhance the return on our nation's* research and development investment by ensuring that digital data realize their full potential as catalysts for progress in our global information society.*"* (Interagency Working Group on Digital Data, 2009)

Australia has also moved relatively speedily to develop an e-research road map; and has set up the Australian National Data Service (ANDS) following a report on the data management implications of e-research which is also an excellent overview of the challenge (ANDS Technical Working Group, 2007).

Closer to home, there are significant efforts on a European Union-wide (EU) basis to progress a shared understanding of and commitment to the development of a pan-European e-infrastructure. The European Strategy Forum on Research Infrastructure (ESFRI) advises the EU Council on investment in major components of the e-infrastructure, including large-scale facilities, and published a roadmap for future development in 2006 (a revised version is in preparation). The e-Infrastructure Reflection Group (e-IRG) acts as a think-tank for major European stakeholders.. It currently has a research data management Task Force, which is undertaking a survey of data management initiatives and whose report is expected shortly.

Back in the UK, the e-Science Core Programme has ended. Interest in e-research remains high, however, as evidenced by the scale of the programmes at the UK's "All Hands Meetings" organised each year by the National e-Science Centre (NeSC). Increasingly, e-research is now becoming more mainstream, as more research acquires the

characteristics of e-research; and the growth of digital data-intensive research in the humanities and social sciences has been particularly noteworthy. Moreover, librarians contemplating the research data landscape are realising that effective data management is needed for smaller-scale projects – the "long tail" of research that doesn't involve massive data volumes, but whose data outputs have the potential to inform future research.

What libraries can do about data

For those managing academic libraries or information technology (IT) services, one of the most difficult considerations relating to data management is working out what needs to be done locally, and what might best be done nationally or internationally. The absence – at the moment – of a coherent national framework for data curation in the UK does not mean that there is no provision. Many subject areas are covered by well-developed data management facilities run by national or international data centres, reflecting disciplinary differences in the academic culture around deposit and re-use of datasets, and these represent a significant asset for the UK in terms of the knowledge base of data management. These facilities include the European Bioinformatics Institute, an agency of the European Molecular Biology Laboratory based in the UK; the network of data centres run by the Natural Environment Research Council (NERC); the UK Data Archive, and the Economic and Social Data Service (ESDS) which it hosts; and the Cambridge Crystallographic Data Centre. There are, however, large gaps, particularly with the demise in 2008 of the Arts and Humanities Data Service, a development which has raised concerns about the level of trust that can be placed in external agencies as persistent guardians of research data for the long term. The need to fill these gaps was one objective of the UKRDS feasibility study.

In the meantime, there are several areas where libraries can and should be active in relation to research data. In most of these areas, they will want to work in partnership with other campus agencies, notably IT services, but also research offices and those responsible for research governance (such as a Pro-Vice Chancellor for Research). Nine such areas can be grouped handily into a pyramid, for ease of reference (fig 2), but this is intended to be neither exhaustive nor definitive. In general, the activities lower in the

pyramid are areas of early engagement, and which may be appropriate for the highest number of university libraries regardless of the scale of the research base of the parent institution.

(i) Develop library workforce data confidence

We'll consider issues about the research data management workforce below; this heading is about raising the general level of awareness of the existing academic library workforce in relation to both e-research and data management issues, with the objective of equipping staff to hold conversations with academic colleagues and research students on these topics. The target audience is primarily academic liaison librarians, but other library staff such as systems teams, repository managers and e-resource managers may also benefit from an improved level of knowledge about and understanding of the data management landscape. There are a number of ways in which this can be achieved.

First, library staff have a professional responsibility to update their own knowledge about data management. There is now a wealth of reading available on the subject, not only from the sources already mentioned, but also as a result of a number of recent studies and projects funded by bodies such as JISC and RIN in the UK. Liz Lyon's report for JISC Dealing with data (Lyon, 2007) is an excellent overview of the current state of play that articulates the policy and operational challenges of data management very clearly. RIN have set out a series of principles for data management (RIN, 2008a), and along with JISC and NERC commissioned a report on researchers' attitudes and practice in relation to data management (RIN, 2008b). JISC have also commissioned a report on the costs of data preservation (Beagrie, 2008) which in addition to providing help for managers trying to assess the resource implications of providing data management capacity also contains a helpful analysis of the different tasks involved in managing data. It's also important for library liaison staff to ensure that they are up to speed with the policies of the principal funders of research in their universities, not only in relation to open access to published outputs, but in relation to data; and that they are aware of the existing national and discipline-based data centres and repositories.

Second, there is an increasing number of externally-organised workshops and courses dealing with data management. Research Libraries UK and the Society of College, National and University Libraries (SCONUL) have organised a number of workshops aimed at academic librarians, based in part on a needs analysis (Martinez, 2007); and the DCC has organised short courses for data managers, as well as a series of international conferences on digital curation. Third, networks of professional practice are beginning to emerge in the UK, such as the DCC Associates' Network, and the Research Data Managerment Forum. There is a still a need, however, to reach out to those university library staff for whom research data is barely on the radar, and this must be a short-to-medium term priority if libraries are to become fully engaged with data management.

(ii) Provide researcher data advice

University libraries may not (yet) have in place the capacity to provide local data management for digital datasets, but once they have engaged with the issue, and once their liaison staff have enhanced their knowledge of the landscape, they can start to provide advice on data management to researchers, both informally and through the development of more formal content on library websites. Many libraries already provide advice on open access and other aspects of scholarly communication, and data management should be seen as a natural extension of this role. Quite often, academic requests for assistance may present as requests to IT services for data storage, so it is important that libraries and IT services have a joined-up approach. Such storage requests may be made rather late in the data lifecycle, but they are a way of starting to identify the research teams and individuals whose research is data-intensive. Initially the level of advice that libraries may be able to provide will be limited: as the workforce develops its confidence, it will expect to influence the way researchers approach data management before research projects start, and ideally at the proposal-writing stage.

(iii) Develop researcher data awareness

In parallel with the provision of advice to individual teams or researchers, there is a role for university libraries in raising awareness of the challenges of data management

within their institutions, and initiating discussion about it through a range of channels. In most institutions there will be a very wide range of interest in data issues, from researchers who have given the fate of the data they generate little if any thought, to those working in areas with well-established cultures of data curation. The RIN report on data publication (RIN, 2008b) highlights this diversity, and also draws attention to some of the disincentives for researchers to expend time and effort on data management. These include lack of familiarity with data management techniques, concern about the volume of requests for information/clarification, uncertainty about whether they have all the permissions needed to publish their data, anxiety about subsequent unauthorised modification or misinterpretation, and a feeling that they themselves may be able to extract further publications from the data. Libraries embarking on local data management advocacy need to consider these points carefully, and ensure that their messages are aligned with those of other institutional stakeholders, notably research administrators.

(iv) Teach data literacy to postgraduate research students

Most UK university libraries have some involvement in research training, either through formal research training programmes, or through less formal channels, although relatively few of them cover research data management (RIN, 2008c). In theory at least this should be a natural development of libraries' information literacy role, one that that is now well established and understood. Research training for postgraduate research students is a key contribution area in relation to research data management, because it presents an opportunity to influence the way in which future researchers approach data when planning their research. The term "data literacy" is often understood to mean "statistical literacy"; but for this purpose we mean developing in postgraduates an understanding of the way in which as future researchers they will generate and use data, how they need to describe it to facilitate future retrieval, how they might approach the identification of data appropriate for preservation, and what options might be open to them for the subsequent storage and curation of their data.

(v) Bring data into undergraduate research-based learning

This is a logical extension of the development of data management skills for postgraduate students. Many undergraduate programmes include a dissertation requirement that will give students experience in the generation of data, and this is an opportunity to start to develop good practice among those who progress to research careers. However, effective management of research data on a wider scale may also bring pedagogic benefits for undergraduate education, by enabling students to access and use real research data in an educational context, an approach that aligns well with the use of problem-based and inquiry-based techniques in the curriculum. Using real research data to enhance students' learning experience will also be of interest to research-intensive universities for whom provision of "research-led" learning is an important differentiator in the undergraduate marketplace.

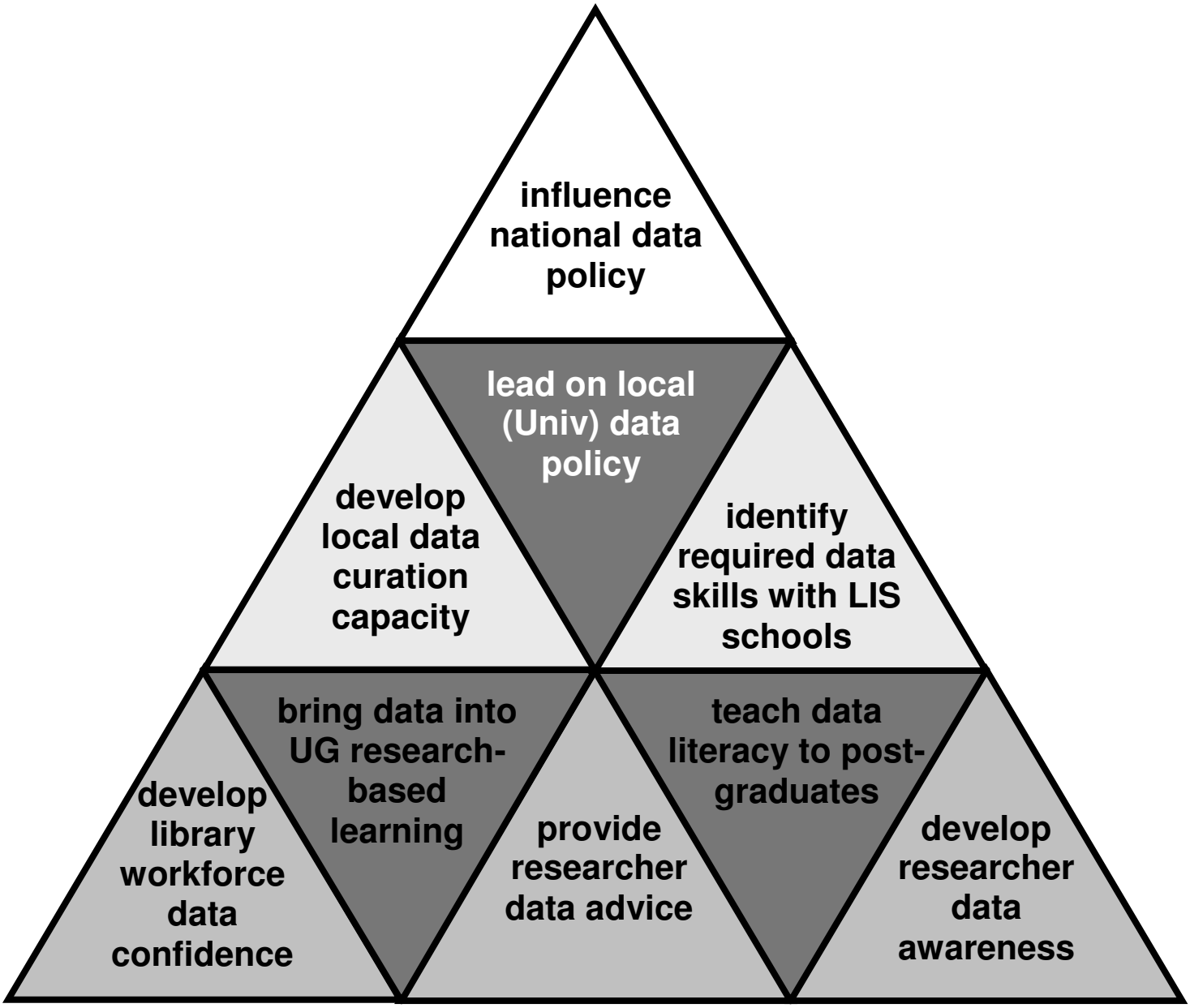

Figure 2 the research data management pyramid for libraries

(vi) Develop local data curation capacity

Assuming that a combination of advocacy and research funder policies impacts effectively on researcher behaviour, should libraries invest in actual data storage and curation capacity? The business case for such investment remains a challenging one, particularly when library budgets are under pressure from the need to sustain current services, to innovate in a wide range of areas (some of them described elsewhere in this book), and to meet the inflationary pressures associated with content procurement. However, an increasing number of case studies are available to inform decisions by library and institutional managers, many of them taking institutional repositories (IRs) as a starting point for data curation. In the UK, the DISC-UK DataShare project followed the journey from conventional IR to data repository in three big research-intensive universities (Rice, 2009), each of them using a different repository platform. Purdue University Library in the US has developed a distributed institutional approach to data curation through its Distributed Data Curation Centre (D2C2) (Brandt, 2007; Mullins, 2007). Toolkits to facilitate the introduction of a managed approach to research data are also starting to become available, among them the Data Audit Framework (Jones et al, 2008) which has been trialled at four UK universities.

(vii) Identify required data skills with LIS schools

While the existing library workforce can make a significant contribution to getting research data curation on the institutional map, even libraries with well-developed IRs are likely to find that they need additional skills in order to provide significant data curation capability locally. There is a role here for library managers in identifying the skills gap and working in partnership with library and information science (LIS) schools to develop new training and development resources to fill it. Not every university library will need or want to be active in this area, but there is a sense among many university library directors that professional practice has actually changed faster than the curricula of the LIS schools supplying new entrants to the workforce; and consequently libraries have a part to play in providing the evidence

base for the development of new data management courses. We'll return to this theme below.

(viii) Lead on local data policy

The informal contacts with researchers and other research stakeholders discussed earlier represent an opportunity for the library to exercise a degree of policy leadership more formally at University level. University research and innovation committees and even senior management teams need to understand the nature of the data management challenge and the benefits of a coherent (but not necessarily uniform) approach across the institution. They may also need to approve a business case for any investment in this area, and their commitment will be crucial in helping to bring sceptical researchers on board. In this respect, it is not only subject staff and repository managers who need to be data-literate: library directors need to be able to articulate both the challenge and the preferred solutions with their senior colleagues. The DISC-UK DataShare project has produced a policy guide for institutions embarking on the extension of their existing IRs to support data deposit (Green et al, 2009).

(ix) Influence national data policy

Librarians can and should expect to be players in their national policy arenas for research data – where these exist. Their influence has been especially apparent in Australia, where librarians are well represented on the ANDS Steering Committee; and Liz Lyon, Director of the UK's library research organisation UKOLN has been a member of the US National Science Foundation's Advisory Committee on Cyberinfrastructure. In Canada, the multiagency Research Data Strategy Working Group, led by the Canada Institute for Scientific and Technical Information, included several university library staff among its membership. The working group has recently published a detailed gap analysis of Canadian research data management provision (Research Data Strategy Working Group, 2008). UK university librarians were not heavily involved in the OST's e-infrastructure sub-groups in 2006, but they have played a major part in the UK Research Data Service feasibility study.

In the next two sections, we'll look at two key non-technical strategic challenges: funding and policy; and workforce development.

Funding and policy

It is clear from the studies conducted so far that providing effective data management throughout the data lifecycle requires non-trivial investment. The return on this investment comes from higher quality research, from easier and therefore cheaper access to existing data, from a reduction in the need to repeat data-generating investigations, and from the facilitation of new research topics and insights. But who should pay?

In the UK, this question has proved much harder to resolve than the technical challenges of data curation. In a provocative interview in 2004 in the journal of the UK's professional library association, Professor Tony Hey, then Director of the e-Science Core Programme, criticised university librarians for failing to engage with the need for long-term management of digital research data (Library and Information Update, 2004). There was some uncomfortable justification for his views, since up to that point librarians had been largely unaware of the growth of grid-enabled research and did not generally see the management of the associated data outputs as being within their professional domain. That has certainly changed: both Research Libraries UK and its US equivalent, the Association of Research Libraries (ARL), set up task forces on e-research in 2005, and few librarians would now argue that research data is an inappropriate area for professional and management attention. On the other hand, the UK Core Programme offered no funds for institution-level data management, and the e-research community probably did not appreciate the resource constraints under which university libraries worked nor the broad front on which change and innovation was taking place elsewhere in libraries, not least in support for learning and teaching.

From libraries' perspective, growing awareness of the scale of the investment needed, coupled with uncertainty about the demand from researchers for data management, and lack of confidence both about their ability to engage with researchers and in the capabilities of their workforces, has been a significant disincentive for involvement. Additional uncertainty has been generated in the UK by the patchy provision of national-

level facilities: will the disciplinary gaps be filled nationally, perhaps by the Research Councils or national agencies such as the British Library or the Joint Information Systems Committee, eventually obviating the need for major local investment? Finding the resources to initiate and develop IRs has not been straightforward for many libraries, and they may feel that extending IRs to include large volumes of data, with metadata, preservation and access challenges an order of magnitude more complex than those posed by e-prints is not a good use of their resources.

Frustrated by the failure of the OST e-infrastructure process to spark policy leadership from government or from the research councils, Research Libraries UK and its IT services equivalent, RUGIT (the Russell Universities Group IT Directors forum), developed a joint bid to the Higher Education Funding Council for England (HEFCE) in 2007, for funding for a feasibility study for a national research data management service. The bid was submitted under HEFCE's shared services programme, on the grounds that although data management was in its infancy in most universities, it would be cheaper to invest in a national framework than to have every university in England develop the necessary capability and capacity locally. The bid was successful, and the UKRDS feasibility study was completed at the end of 2008. (UK Research Data Service, 2008, 2009). The study confirmed (i) that even conservative assumptions about the cost of local research data management centres yielded significant savings for a national approach; (ii) that, rather than establishing a monolithic central agency, a UKRDS should be an enabling framework that would facilitate a mixture of appropriate local and national provision, identifying gaps and commissioning additional capacity as required, with a registry of researchers' data management plans as a core component of the service.

The final report recommended that funding should be allocated for an initial two-year "pathfinder" phase. Rather than pilots, the pathfinders would be live components of the UKRDS service involving a subset of research-intensive universities, at least one Research Council, and one of the existing national data centres. At the time of writing, a bid for the pathfinder phase is being developed by the UKRDS project team. One of the political dimensions of this challenge, and one which the UKRDS study has already encountered, is the UK's unusual "dual support system" for research funding (Adams & Bekhradnia, 2004). Dual support means that universities receive two separate streams of

public funding: one from the Research Councils, relating to specific projects and programmes, and one from HEFCE and its Welsh and Scottish equivalents, intended to provide for discretionary and "blue skies" research, but increasingly linked to the provision of basic research infrastructure. While some of the seven Research Councils top-slice their own funding in order to operate national data centres, others see data management as an infrastructural cost which should be on the university side of dual support. This continuing discussion in the UK's corridors of power demonstrates that the development of a sustainable business model for research data management is key to scaling up capacity to meet the needs of twenty-first century research.

Workforce development

While it might be heartening to hear non-librarians expressing confidence that librarians' renowned metadata skills equip them to be the research data managers of the future, knowledge of MARC, AACR and even Dublin Core does not represent a licence to curate research data. Neither does liaison librarians' knowledge of the bibliographic landscape of their territories mean that they can expect to advise scientists on data collection formats. Developing librarians' data confidence will enable them to have conversations about data with researchers, and the importance of this step should not be underestimated. However, the next level of engagement and support for research data will require new skills, or new combinations of skills, and new roles.

JISC commissioned a major study from consultancy Key Perspectives on the development of "data scientists" (Swan and Brown, 2008), a slightly unfortunate charge since as the authors note

> *"In practice, there is not yet an exact use of such terms in the data community,* and the demarcation between roles may be blurred. It will take time for a clear *terminology to become general currency."*

Swan and Brown see several differentiated but partly overlapping roles emerging to support research data management, from the data creators (the research scientists), through data scientists (data experts working closely with researchers, and often with the same domain subject background), and data managers (typically information

technologists) to data librarians (usually based in academic libraries and managing local data collections). Corrall (2008) identifies three overlapping skill domains in which the hybrid data professionals of the future will work, which she terms "context" (ie academic research), "conduit" (primarily technology) and "content" (library and information science). There are so far very few data librarians in UK universities (Swan and Brown (2008) estimate the total at five), and most of them are associated with institutions that have distinctive specialist roles or collections (Macdonald & Martinez, 2005). Data scientists and data managers can be found in national data centres, and in some cases attached to big research teams in universities.

Clearly, few if any university libraries are likely to be able to go out and recruit a team of data scientists and data managers to cover their university's disciplinary spectrum. The need for domain subject knowledge for data scientists is itself a powerful argument in favour of national-scale solutions, at least for some disciplines: large data centres are more likely to be able to create a critical workforce mass, and to be able to give their data specialists a reasonable career structure. This is already the case in some areas such as bioinformatics: the European Bioinformatics Institute has a staff bigger than most university libraries. Data scientists may also be attached to big research groups, either as permanent team members or on a per-project basis, in which case they may be supported as direct costs by research funders.

It is likely or even probable that data scientists will not come from traditional library backgrounds; they are more likely to be career researchers for whom a period as a data scientist is part of a longer-term research career track. But their posts may come into existence in part because effective liaison between the library and the research team has already highlighted the project's data management requirements, and resulted in the inclusion of a data scientist post in the grant proposal. Who might have provided that advice? Perhaps the university library's data librarian, who may also have a role in the management of locally held datasets for smaller projects, a requirement that may continue even if the large-scale gaps in national provision are plugged in the future.

This scenario implies the need for several types of training and development:

(i) award-bearing programmes, probably at Masters level, for career data scientists and data managers intending to achieve career track positions in large data centres

(ii) short-course provision, not necessarily award-bearing, but probably accredited, for career researchers interested in project-based data science and management roles

(iii) training for data librarians: in the short term the demand here is likely to be for post-qualification training from members of the existing library workforce; as the requirement for such posts increases, there may be demand for data-oriented postgraduate LIS courses for new entrants intending to specialise or retrain in data librarianship, though take-up may depend on the extent to which data librarians can (and want to) progress into more senior academic library roles. An early exemplar is the MS Specialisation in Data Curation offered by the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign (which also offers an MS in Biological Informatics).

There is also likely to be demand, in line with the data confidence theme, for some of the course content in these programmes to be available to mainstream academic library staff as continuing professional development modules.

Conclusions

It should by now be apparent that the "yes and no" with which we started is far more "yes" than "no". But there remain many questions, not all of which will find answers before library managers find themselves having to make difficult decisions about how much time and resource to invest. Anna Gold, in an excellent review article on libraries and e-research, notes

> In sum, it is fair to say there is still a substantial amount of uncertainty about the roles libraries can play in scientific data management, reflecting an environment of ongoing experimentation and negotiation (and perhaps some wishful thinking). (Gold, 2007)

This is still true, and arguably not just in relation to science data. Among the remaining uncertainties are the following:

(i) How rapidly will demand from researchers for data management grow?

(ii) Will more research funders mandate deposit of data outputs?

(iii) How will the data management requirement be funded?

(iv) Will researchers be interested in data scientist/data manager roles, and will the academic community recognise this as a mainstream research career route and not a dead end?

(v) Will data storage/curation/access capacity develop at national and international level, and how quickly?

From a UK perspective, there is a further pressing question: will a policy lead be taken by any of the major research stakeholders in a position to effect change? In 2003, as we noted earlier, Lord and Macdonald observed a lack of overall strategy for data management and the associated infrastructure. Over five years later, Professor Sir Ron Cooke, outgoing chair of JISC, commented:

> More investment and policy leadership is required for the curation of research data, including international collaboration, to build a layer of academic and scholarly resources readily available to all. This should be a priority for DIUS, RCUK and others where clear policy leadership is urgently required. (Cooke, 2008)

These questions will not be answered in the very short term. The difference librarians have made in recent years, however, is that they are now well-placed to influence many of the answers. This positions the profession to add significant value to an area that, over the course of the next decade, is set to move from being on the fringes of professional concern to being a core component of libraries' support for the academic research mission.

Acknowledgement

I am grateful to Professor Sheila Corrall, Department of Information Studies, University of Sheffield, for discussion and comments.

References

Adams, J. and Bekhradnia, B. (2004) What future for dual support? London: Higher Education Policy Institute. http://www.hepi.ac.uk/downloads/6%20Dual%20Support.pdf [accessed 2009-08-12]

ANDS Technical Working Group (2007) Towards the Australian Data Commons: a proposal for an Australian National Data Service. The Department of Education, Science and Training. http://www.pfc.org.au/pub/Main/Data/TowardstheAustralianDataCommons.pdf [accessed 2009-08-09]

Beagrie, N., Chruszcz J., and Lavoie, B. (2008) Keeping research data safe: a cost model and guidance for UK universities. Joint Information Systems Committee. http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf [accessed 2009-08-10]

Brandt, D. S. (2007) Librarians as partners in e-research: Purdue University Libraries promote collaboration. College & Research Libraries News **68** (6) 365-7, 396.

Cooke, R. (2008) On-line innovation in higher education: submission to the Rt Hon John Denham MP Secretary of State for Innovation, Universities and Skills. London: Department for Innovation Universities and Skills. ttp://www.dius.gov.uk/higher_education/shape_and_structure/he_debate/~/media/publications/O/online_innovation_in_he_131008 [accessed 2009-08-09]

Corrall, S. (2008) Research data management: professional education and training perspectives. [presentation] Research Data Management Forum, November 2008. http://www.dcc.ac.uk/events/data-forum-2008-november/presentations/07.pdf [accessed 2009-08-13]

Gold, A. (2007) Cyberinfrastructure, data, and libraries. (Parts 1 and 2) D-Lib Magazine **13** (9,10). http://www.dlib.org/dlib/september07/gold/09gold-pt1.html, http://www.dlib.org/dlib/september07/gold/09gold-pt2.html [accessed 2009-08-13]

Green, A, Macdonald, S. and Rice, R. (2009) Policy-making for research data in repositories: a guide. EDINA and [Edinburgh] University Data Library. http://www.disc-uk.org/docs/guide.pdf [accessed 2009-08-11]

H.M.Treasury. (2004) Science & innovation investment framework 2004-2014. HMSO. http://www.hm-treasury.gov.uk/spending_sr04_science.htm [accessed 2009-08-10]

Hey, T and Trefethen, A. (2003) The data deluge: an e-science perspective [in] F.Berman [et al] (eds) Grid computing: making the global infrastructure a reality. Wiley. http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/research/esci/datadeluge.pdf [accessed 2009-08-09]

Interagency Working Group on Digital Data (2009) Harnessing the power of digital data for science and society Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. Interagency Working Group on Digital Data. www.nitrd.gov/about/harnessing_power_web.pdf

Jones, S. (et al) (2008) The Data Audit Framework: a first step in the data management challenge. International journal of digital curation, **3** (2) 112-120.

Library & Information Update. (2004) Why engage in e-science? Library & Information Update. **3** (3), 25-27.

Lord, P. and Macdonald, A. (2003) e-Science Curation Report: data curation for e-Science in the UK: an audit to establish requirements for future curation and provision, prepared for the JISC Committee for the Support of Research. Joint Information Systems Committee. http://www.jisc.ac.uk/uploaded_documents/e-sciencereportfinal.pdf [accessed 2009-08-09]

Lyon, E. (2007) Dealing with data. Joint Information Systems Committee. http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report.pdf [accessed 2009-09-15]

Macdonald, S and Martinez, L. (2005) Supporting local data users in the UK academic community. Ariadne [online], 44. http://www.ariadne.ac.uk/issue44/martinez/intro.html [accessed 2009-09-17]

Martinez, L. (2007) The e-Research needs analysis survey report [for the] CURL/SCONUL Joint Task Force on e-Research. www.rluk.ac.uk/files/E-ResearchNeedsAnalysisRevised.pdf [accessed 2009-08-17]

Mullins, J.L. (2007) Enabling international access to scientific data sets: creation of the Distributed Data Curation Center (D2C2). Purdue University. [Online]. Available: http://docs.lib.purdue.edu/lib_research/85/ [accessed 2009-08-11]

National Science Foundation (2007a). Cyberinfrastructure vision for 21st century discovery. National Science Foundation. http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

National Science Foundation (2007b) Sustainable Digital Data Preservation and Access Network Partners (DataNet): Program Solicitation NSF 07-601. National Science Foundation, Office of Cyberinfrastructure. http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm [accessed 2009-08-09]

Organisation for Economic Co-operation and Development (2007) Principles and guidelines for access to research data from public funding OECD. http://www.oecd.org/dataoecd/9/61/38500813.pdf [accessed 2009-08-13]

Pothen, P. (2007) Developing the UK's e-infrastructure for science and innovation: report of the OSI e-Infrastructure Working Group. [Joint Information Systems Committee]. http://www.nesc.ac.uk/documents/OSI/report.pdf

Research Data Strategy Working Group (2008) Stewardship of Research Data in Canada: A Gap Analysis. Research Data Canada, 2008. http://data-donnees.gc.ca/docs/GapAnalysis.pdf [accessed 2009-08-13]

Research Information Network (2007) *Researchers' use of academic libraries and their* services, RIN, 2007. http://www.rin.ac.uk/files/libraries-report-2007.pdf [accessed 2009-08-13]

Research Information Network (2008a) Stewardship of digital research data: a framework of principles and guidelines. RIN.

http://www.rin.ac.uk/files/Research%20Data%20Principles%20and%20Guidelines%20full%20version%20-%20final.pdf [accessed 2009-08-13]

Research Information Network (2008b) To share or not to share: publication and quality assurance of research data outputs. RIN. http://www.rin.ac.uk/files/Data%20publication%20report,%20main%20-%20final.pdf [accessed 2009-08-13]

Research Information Network (2008c) Mind the skills gap: Information handling training for researchers. RIN. http://www.rin.ac.uk/files/Mind%20the%20skills%20gap%20REPORT%20July%2008.pdf [accessed 2009-08-13]

Rice, R. (2009) Final report [of the] DISC-UK DataShare project. Joint Information Systems Committee. http://ie-repository.jisc.ac.uk/336/1/DataSharefinalreport.pdf [accessed 2009-08-11]

Swan, A. and Brown, S. (2008) The skills, role and career structure of data scientists and curators: assessment of current practice and future needs. Key Perspectives. http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf [accessed 2009-08-12]

UK Research Data Service feasibility study (2008) Report and recommendations to HEFCE. December. http://www.ukrds.ac.uk/HEFCE%20UKRDS%20Final%20Report%20V%201.1.doc [accessed 2009-08-12]

UK Research Data Service feasibility study (2009) The data imperative: managing the *UK's* research data for future use. Joint Information Systems Committee. http://www.ukrds.ac.uk/UKRDS%20Report%20web.pdf [accessed 2009-08-12]

University of Illinois at Urbana-Champaign, Graduate School of Library and Information Science (n.d.) Master of Science--Specialization in Data Curation. http://www.lis.illinois.edu/programs/ms/data_curation.html [accessed 2009-08-13]