Managing the UK's research data: towards a UK Research Data Service

Managing the UK's research data: towards a UK Research Data Service

Foreword

"Because digital data are so easily shared and replicated and so recombinable, they present tremendous reuse opportunities, accelerating investigations already under way and taking advantage of past investments in science".

These are the words of Clifford Lynch, Executive Director of the Coalition for Networked Information in the USA, writing in *Nature* on 4 September 2008.

Lynch is well known across the world for his ability to identify and evaluate key trends in the networked information environment which have transformational power and can be translated into real-life improved services for researchers. In his *Nature* article he recognises the importance of data as a significant research resource in their own right, and his thoughts reflect a growing awareness of the need for a coherent and consistent approach to the management of data. Moreover, it is not just in 'big science' that we see the so-called data deluge. All disciplines are now subject to the creation of major amounts of raw and derived data, and attempting to manage the data in a coherent way represents a huge challenge to the research community and those who support it.

This article tells the story so far of an important project in the UK to determine the feasibility of a coherent UK-wide approach to the management of research data.

The HEFCE shared services programme (2007-08)

In 2007 the Higher Education Funding Council for England (HEFCE) called for bids from consortia of higher education institutions to undertake feasibility studies for sharing services. Members of RUGIT (Russell Group IT directors) and CURL (Consortium of University Research Libraries, now renamed RLUK, Research Libraries UK), had both been grappling with issues to do with support of research data, and they decided to submit a joint bid for funding to assess the feasibility and costs of developing and maintaining a national shared research data service for the UK. The bid succeeded in attracting a grant of £200,000 from HEFCE, together with contributions of £35,000 from JISC and £10,000 each from RUGIT and RLUK themselves. The initiative became known as UKRDS (UK Research Data Service). A wide range of stakeholders across the research spectrum were included in an online mailing list to whom regular progress reports and updates were sent throughout the project. A logo was designed and a website established (http://www.ukrds.ac.uk). The study employed consultants Serco plc and project manager John Milner to take forward the issues. In support were both a high-level steering committee chaired by Professor John Wood of Imperial College, who is chair of the JISC (Joint Information Systems Committee) Committee for Research, and a small Project Management Group consisting of members of RUGIT and RLUK chaired by the author of this article who is a member of both organisations. The study was completed within a year and was submitted to HEFCE just before Christmas 2008.

The antecedents of the UKRDS feasibility study

A research data service on a national scale was seen by the project sponsors as forming a crucial component of the UK's e-infrastructure for research and innovation. Already a Treasury report entitled *Science and Innovation Investment Framework*

2004-14 (HMSO 2004) had signalled the need for development of e-infrastructure. As a result the Office of Science and Innovation (OSI) established an e-infrastructure group with wide participation including from higher education, and in February 2007 the group published its report *Developing the UK's e-infrastructure for science and innovation* (http://www.nesc.ac.uk/documents/OSI/report.pdf). The report clearly indicated areas requiring further development, including data curation and management. Among the stakeholders interested in the development of an effective research data infrastructure for the UK are the research councils, the Department for Innovation, Universities, and Skills (DIUS), JISC, the higher education funding councils, and individual higher education institutions.

The RUGIT and RLUK feasibility study aimed to build on the OSI report by addressing the issue of the entire data management lifecycle and providing a roadmap for infrastructure development.

The challenge

The sponsors identified the main challenges to be addressed by the feasibility study as follows:

- Research data remain a substantially untapped resource beyond the originators in many cases;
- Management of research data is not consistent or coherent and only a minority of researchers have access to national or international facilities;
- Provision of skills for data curation support and training for researchers is under-developed;
- Research data are often unstructured and inaccessible to others;
- There is little consistency of policy or practice across funders and disciplines;
- Pressure is growing on higher education libraries and IT services to assist researchers with these issues and it is unlikely that the necessary data management and curation capacity can be provided locally in individual universities.

Simply put, from our observations of researchers in our universities we saw that research data were often stored on memory sticks, desktop PCs, and departmental servers and much of it was inaccessible to researchers, including sometimes the creators of the data themselves, largely because of a lack of adequate metadata and search tools. Moreover, some researchers had begun to ask either librarians for storage capacity in the institutional repository, or IT managers for space in their central data servers. Without management tools this approach cannot add any value; nor is it sustainable into the future as the amount of research data grows.

Aims of the UKRDS

So the study was established to test the feasibility of a national shared service for managing research data. This would build on past and current investment in infrastructure and on good practices where they existed, and would develop capacity for the long term. A successful UKRDS should:

- Advocate practical data management methods for researchers and funders;
- Co-ordinate dataset management in accordance with protocols established between data providers, data users, and data storage and curation facilities;
- Enable data discovery;

- Provide access to tools and expertise;
- Be a focus for the development of policies and standards;
- Provide access to effective training services and materials;
- Engage existing facilities and expertise;
- Be more cost-effective as a shared service than institutions could be acting independently.

Methodology for the study

The study involved three key areas of work. First was the assessment of researchers' requirements through the use of questionnaires, interviews and workshops conducted in four institutions whose library and IT directors agreed to be case studies: Bristol, Leeds, Leicester, and Oxford. Second was engagement with a wide range of stakeholder groups including major funding bodies, archives and libraries, existing data facility and service providers and others, including those involved in this kind of work internationally. And third was far-ranging desk research aimed particularly at assessing the UK provision in an international context. All three methodologies served to identify what already exists in the way of data management capability, what the gaps are, and where the UK fits compared to comparators and competitors abroad. The consultants were set the tasks of using the evidence gathered in these three ways, synthesising it, evaluating the feasibility or otherwise of a coherent national approach, developing a business case, and making an initial attempt to identify shared service vehicle design and articulate a suitable governance structure. The consultants and project manager reported regularly to the project management board and at key milestones also presented to the steering committee (for example an interim report was produced halfway through the study period).

Key findings of the study

The case study work involved consultation with groups representing approximately 700 researchers and showed a number of issues including:

- An increasing number of disciplines are producing electronic research data;
- There is an acknowledged difficulty for researchers in retaining or managing research data beyond the life of a project once the funding associated with the project ceases;
- Most research data are stored at faculty or department level unless there is a national data facility available;
- 21% of researchers use a national or international facility;
- Most researchers share data – only 12% do not make their data available in any way; however informal peer exchange networks within research teams and with collaborators are pre-dominant in this sharing; only 18% share via a data centre;
- Although a relatively small percentage (18%) share their data via a data centre, 43% expressed the need to access other researchers' data;
- Those who did not have access to an established national facility were particularly keen on the establishment of a UKRDS.

The engagement with stakeholders and desk research indicated that there was substantial infrastructure and expertise in the UK but that this was set up in 'islands', because each facility had been established to address a particular need, and coherence and communication between islands was limited. As the project progressed it became clear there was a growing awareness among public bodies in the UK and in a number

of other countries that research data needed to be harnessed and managed in order to exploit their potential. So although the conclusion was that the UK has a sound basis on which to build, we know that there is good work under way in Europe, Australia, Canada and the US, some of it centrally funded, and the UK needs to maintain its competitive position.

It is interesting to note that the thesis posited by the sponsors of the feasibility study in the first place (RUGIT and RLUK) turned out to be validated by the results of the case studies, engagement with stakeholders, and desk research: *there is a problem of lack of cohesion and consistency in the approach to managing data and a need to address this on a UK-wide basis.*

As for the optimum way forward, the study identified three options to choose from. One is the do-nothing option. Despite the complexity of the challenges facing a shared approach, this was not an option that the steering committee wished to consider. This would leave the current situation in place and UKRDS would not exist in any form. Some researchers and disciplines would be well provided for and others would not. Any attempt to improve the situation would require individual higher education institutions to manage their own research data lifecycle and most would be unable to provide the resources, skills and capacity to do this.

The second possible option is to create a highly centralised service. This would be invasive and expensive, creating a new monolithic institution with responsibilities in every area of data management. Not only would this be extremely complex and costly to build, but it would duplicate unnecessarily work already being done and would therefore be unlikely to receive support and buy-in from those service providers already in existence and doing a good job. In other words, it would be bound to fail and the steering committee rejected it.

The third option, the one recommended by the consultants and chosen by the steering committee as the only viable solution, is the co-operative service. In this option UKRDS would act as an enabling service, working with the many UK stakeholders and the existing facilities. Such a service would be well placed to act as a catalyst for new services and partnerships, as a centre of excellence, as a standards-guiding body and as a source of expertise and information about data management and repositories, building on best practice and facilities. The UKRDS could act essentially as a broker to connect researchers and institutions to existing facilities and centres of expertise, while also commissioning work from existing bodies to fill gaps in the areas of policies, data management planning, metadata creation, curation skills and others. The diagram headed *Communities and headline processes* shows how the UKRDS would operate in relation to other stakeholders.

Features and benefits of a UKRDS

Research costs are growing and the management of research data is a significant cost. A shared service approach holds the promise of minimising the long-term financial impact and adding value through better exploitation of the data. Central to the co-operative service model is the development of data management plans by researchers, based on the data lifecycle as described by the Digital Curation Centre (DCC) and the development of a central registry of such plans. This approach would allow a

UKRDS to maximise exploitation of existing facilities within the UK and to identify and fill gaps in current provision. A UKRDS would help to:

- Protect and extract greater value from research investment;
- Preserve opportunities for future research;
- Promote the work of the institution and researcher;
- Inform the strategic development of the research infrastructure;
- Reduce research data duplication, re-creation and errors, and unplanned data loss;
- Plan volume growth/capacity more effectively;
- Provide more opportunity for re-use, cross-reference and dataset integration;
- Target retention and disposal more appropriately;
- Share skills, giving better coverage and productivity in both service providers and researchers;
- Provide an effective focus for best practice in data curation.

Additional direct benefits to the institution, to the researcher and to funders could include:

- Guidance on which repository to get research data from and a gateway to approved service providers;
- Help with the use of data management plans to facilitate lifecycle management of datasets;
- The opportunity to inform strategic development of the research infrastructure to local and national levels, and to work with stakeholders to inform policy and resourcing of post-project long-term data management;
- Commissioning of new services to fill gaps in data management provision.

Initiatives in other countries

As mentioned above, the UK is by no means the only country in which the challenges facing researchers in managing their data are being actively discussed and addressed. In a number of other countries across the world the problem of the data deluge is a live topic and there is recognition at the highest level of the potential value of research data in enhancing the global reach and reputation of the national research base.

Thus for example *Research Data Canada* is a government initiative aimed at addressing the issues confronting researchers in the management of their data. A Research Data Strategy Working Group, consisting of representatives from universities, institutes, libraries, granting agencies and researchers, is working across a number of fronts. It follows a major national consultation on access to scientific research data conducted in 2005 which concluded that there was an "urgent need for action to propel Canada into a new and transformational data-intensive paradigm for Canadian research". The working group's concerns are uncannily similar to those which emerged from the UKRDS feasibility study. Their website (see http://data-donnees.gc.ca/eng/index.html) states:

"The research process generates huge amounts of data that are an important part of Canada's scholarly record and hold enormous potential as an additional discovery and problem-solving tool for researchers. Unfortunately, there are no nationally adopted standards or policies governing how this data is collected, catalogued or preserved. As a result this data is often inaccessible by other

researchers or structured in such a way that it can't be fully exploited by other users".

In December 2008 the group produced an important gap analysis report on the stewardship of research data in Canada (see the website previously cited for a link to the report). The report provides a statement of the ideal state of research stewardship in Canada compared to the current state with the aim of starting to fill gaps across a range of indicators including: policies, funding, roles and responsibilities, data repositories, standards, skills and training, reward and recognition systems, research and development, accessibility and preservation.

The German Research Foundation, *Deutsche Forschungsgemeinschaft*, is also working on a road map towards effective management and curation of digital research data on a national basis. A number of reports and declarations form the antecedents to this, starting with a set of DFG principles for safeguarding good scientific practice set out in 1998 which stated that primary data as the basis for publications should be securely stored for ten years in a durable form in the institution of their origin. More recently, in June 2008, the Alliance of German Science Organisations launched a national priority initiative on digital information. Priority area 4 is primary research data, and the text of the document reads:

"Even after a relatively short phase of scientific evaluation by individual researchers or small groups much of this data is forgotten and/or allowed to deteriorate. All scientific institutions therefore see an urgent need for action in order to ensure the systematic backup, archiving and provisioning of scientific data for subsequent (re)-use by third parties".

The alliance partners are now aiming to develop three areas. First, the formulation of a common data policy to promote the need for action and to demonstrate the usefulness of primary data infrastructures for researchers. Second, the fostering of cooperation between researchers and information specialists, and the funding of pilot projects which will develop subject-specific standards and methods of data curation and archiving. Third, following the pilot projects, the establishment of a system of discipline-specific, internationally networked data repositories for primary research data. Most importantly of all perhaps, the alliance partners have agreed to coordinate their funding in the area of primary research data and where necessary merge or harmonise them, with the future possibility of establishing common infrastructures for primary research data. (See http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/download/allianz_initiative_digital_information_en.pdf for an English translation of the joint infrastructure initiative signed by the members of the German research alliance).

In the USA the National Science Foundation's *Office of Cyberinfrastructure* plans to fund five large Datanets, each consisting of a very large consortium of universities, to build data stewardship capabilities. $100 million is available for five years for this programme, with a maximum of $20 million per consortium. As in the Canadian and German initiatives, the aims set out in the proposed Datanets programme show a marked similarity with the issues identified in the UKRDS project, although at this stage the subject domain in the US appears to be mainly science and engineering. In the original call, dated September 2007, there is first a recognition of the fast-growing volume of data involved in research. Then there is the recognition of the value of

research data: “Digital data are not only the output of research but provide input into new hypotheses, enabling new scientific insights and driving innovation”. There follows the now familiar description of the challenges facing today’s researchers and those who support them: how to develop the new methods, management structures and technologies to manage the diversity, size, and complexity of current and future research data. The NSF Datanets will be a set of exemplar national and global data research structure organisations designed to provide unique opportunities to communities of researchers in science and engineering. Full final proposals for the Datanets are due in May 2009 and therefore an announcement of the successful bidders can be expected later in the year.

The country which appears to be furthest along the route towards implementing a national data management service is Australia, with *ANDS*, the Australian National Data Service, which has received 24 million Australian dollars over three years starting in 2008. Much as for Canada and Germany, the Australian initiative has its antecedents in an alliance of key high-level policy bodies signing an agreement leading to the establishment of a national approach. The document in question is the *Australian code for the responsible conduct for research*, a joint publication of the Australian Government, the National Health and Medical Research Council, and the Australian Research Council. Published in 2007, it describes the responsibilities of institutions and researchers covering a range of issues and includes a section on the management of research data and primary materials. The first sentence of the data section states: “Policies are required that address the ownership of research materials and data, their storage, their retention beyond the end of the project, and appropriate access to them by the research community” (see http://www.nhmrc.gov.au/publications/synopses/_files/r39.pdf). ANDS is working in practical ways to help the various participants to fulfil their responsibilities under the code.

The model is a distributed one, where the institutions and the researchers are to retain their own research data, provide secure storage (via their institutional repositories), identify ownership, and ensure security and confidentiality of research data. ANDS is structured as four coordinated inter-related service delivery programmes: Frameworks, Utilities, Seeding the Commons, and Capability Development. The Frameworks programme will influence relevant national policies, build a common understanding of data management issues, and encourage default sharing practices. The Utilities stream will build and deliver national technical services including discovery and persistent identifier services and a data collections registry (much the same as the data plans registry proposed by the UKRDS feasibility study). Seeding the Commons is aimed at improving and standardising institutional repositories and encouraging researchers to deposit their data. And the Capability Development strand will help researchers to align their data management practices with the needs and outputs of ANDS, and learn from existing best practices.

These initiatives in other parts of the world show how the need for research data management has begun to be recognised and faced up to, notwithstanding the complexity of the issues and challenges involved and the potential costs of developing a national approach. In this regard the findings of the UKRDS feasibility study are very much in step with the thinking elsewhere. However, there are two significant

differences between the UK situation and that of some of the other countries mentioned above (notably Australia, Canada, and Germany).

The first difference is that in the UK there is as yet no coordinated approach, or agreement between key players (particularly funders), towards a national service in support of research data. This state of affairs is unfortunately likely to hold back the UK's potential to exploit its research data to the full, make the most of the country's investment in research, and increase its global reputation for high-quality research output. In October 2008 Professor Sir Ron Cooke, chair of the JISC Board, made this same point in his report to the Secretary of State for Innovations, Universities and Skills entitled *On-line innovation in higher education* (see http://www.dius.gov.uk/policy/documents/online_innovation_in_he_131008.pdf):

> *"Modern research is generating massive amounts of data and there is little indication that the UK is gearing up to deal with this… Recent developments in Australia…and in Germany … clearly demonstrate the extent to which the UK has already lost the initiative in this area".*

However, the second difference between the situation in the UK and that in other countries is potentially a great strength. It consists in the existence already of significant infrastructure and services in support of research which, if given more coherence and assisted by gap-filling through a UKRDS, could place the UK quickly and firmly ahead of its competitors in the global research market. No other country can boast as many existing benefits on which to build.

For example, the UK has a considerable number of national discipline-based data centres with considerable facilities, skills and expertise which could potentially be resourced to spread their services more widely. These include the excellent data centres of the UKDA (UK Data Archive), the NERC (Natural Environment Research Council) and the STFC (Science and Technology Facilities Council).

The Wellcome Trust's leadership on the encouragement of researchers, through the incentive of additional funding, to create data management plans, is also of major importance. Data management plans are seen by the UKRDS feasibility study (and by ANDS) as being the essential tool for ensuring lifecycle management of data, and a national registry of data management plans forms a key element of both these projects. The pioneering work of the UK's Digital Curation Centre (DCC) on the description of the data lifecycle model is another excellent building block on which to develop a national research data service and deserves to be adopted and spread throughout the research community in a coherent and coordinated way.

The infrastructure for data storage and retrieval is provided by JISC's much admired integrated information environment (IIE) and the world-class JANET network.

JISC and the Research Information Network (RIN) are also doing significant work in research and development to provide context and evidence to the issues of data management. For example, JISC's work on Data Audit Frameworks and data handling skills (relative lack of) offers significant pointers to the way forward. And the RIN has had considerable success in establishing itself as an observatory which can reach the researcher community and provide evidence of researcher needs and behaviours. It is also a body which offers foresight studies and horizon-scanning and

develops policy guidelines in a uniquely independent way. RIN's study in June 2008 entitled "To share or not to share" cast an illuminating light on researchers' current approach to and views on the publication and quality assurance of their research data and their expressed challenges and needs for the future (see the RIN website at http://www.rin.ac.uk). Both JISC and RIN are planning more studies on various aspects of data management during 2009/10.

The way forward

The feasibility study has concluded that, building on the impressive range of UK facilities and services which already support researchers, a UK Research Data Service can be developed to exploit best practice, bring coherent standards and policies, and encourage and commission the filling of gaps. There is a need to bring together a wide range of stakeholders to create the cohesion required for a national service. The study proposes that the best way to start is with a *Pathfinder phase* which will try out a UKRDS in earnest, using a limited number of researchers and players from the existing services. In effect this phase will be learning by doing. The IT and Library directors of the four case study universities (Bristol, Leeds, Leicester and Oxford) have agreed to participate in such a venture, and they will seek the support of their institutional research support services too. It will be the job of these local partnerships to seek the agreement of some real researchers and research groups to participate in the Pathfinder. The DCC would help with advice and guidance on how to implement its data management lifecycle model, and the RIN would assist with its own foresight studies and evidence base. The UKDA would assist by offering its expertise in data storage and management to researchers beyond its normal community, and it is hoped that STFC and NERC would also participate in this way. In the meantime, work would be initiated on filling gaps such as policy and standards documents, and issues such as training in data handling skills and preservation would be explored in some detail. At the time of writing is it uncertain how such a Pathfinder phase, if approved by HEFCE and other potential funders, would be financed. A good place to start would undoubtedly be the signing of an agreement at the highest level to pursue a coordinated approach to the management of research data within a national framework, as has happened in other countries as described above.

The 2008 Research Assessment Exercise has confirmed that the UK's higher education research base remains world-class and in many respects world-leading. Now the establishment of a UK-wide data management service can help to ensure that this national capacity for knowledge generation and innovation remains competitive and that UK researchers can continue to benefit from a high quality research infrastructure.