Exploring the roles and responsibilities of data centres and institutions in curating research data – a preliminary briefing.

Dr Liz Lyon, UKOLN, University of Bath

Introduction and Objectives

UKOLN is undertaking a small-scale consultancy for the JISC to investigate the relationships between data centres and institutions which may develop data repositories. The resulting direction-setting report will be used to advance the digital repository development agenda within the JISC Capital Programme (2006 – 2009), to assist in the co-ordination of research data repositories and to inform an emerging Vision and Roadmap.

The Consultancy objectives are:

- *To define how institutions (collectively and individually) and scientific data centres can together effectively achieve:*
  - *Preservation*
  - *Access – Managed and Open*
  - *Reuse – Data Citation, Data Mining and Reinterpretation*

- *To identify the mechanisms, business processes and good practice by which these functions can be achieved*

- *To facilitate dialogue between data centres, institutions and other key players and to define a collaborative way forward.*

This Workshop is intended to inform the Consultancy work, and to provide a forum in which stakeholders can initiate the preliminary identification and discussion of the key issues which need to be addressed.

Context and Vision

During the last three years, in the UK we have seen an increasing investment in institutional repositories (IR) though as yet, there are few examples of IRs containing research data, either raw or processed. The JISC has funded a number of projects which are investigating the implementation of data repositories, such as eBank and the Digital Repository Programme data cluster projects (GRADE, StORe, SPECTRa, CLADDIER and R4L). Information about these projects is available on the DigiRep Wiki[1].

The JISC also funds data services such as MIMAS, EDINA, AHDS and the UK Data Archive, which provide a range of dedicated facilities for data management and preservation, and which manage substantive collections of data. Three of these services, (MIMAS, AHDS and UK Data Archive), also receive funding from the Research Councils.

The Research Councils fund a number of data centres which provide expert curation services for the increasing volumes of data produced as a result of Research Council-funded research programmes and projects. Some (but by no means all), of these activities are e-Science projects, which are generating huge volumes of data from grid-enabled applications in disciplines and sub-disciplines such as high energy physics, genomics, aeronautical engineering and combinatorial chemistry. In addition, other organisations such as the Wellcome Trust, are producing data as outputs from their funded research programmes and are pro-actively promoting the concepts of open access data and information. This open approach is mirrored by the policies of the National Institutes for Health (NIH) in the US, on the basis that the outputs from all publicly-funded research should be openly available[2].

In June 2006, the UK Research Councils published an updated statement presenting their position with regard to (open) access to research outputs, and announced plans to assess the impact of author-pays publishing and self-archiving on research publishing[3]. This will report in 2008 when the RCUK position will be reviewed. The position with regard to open data is less clear. although some Research Councils such as the MRC, have a published Policy relating to data and preservation[4], and which is based upon the OECD Principles[5]:

"MRC expects that the valuable data arising from the research it supports will be made available to the scientific community to enable new research with as few restrictions as possible. Such data must be shared in a timely and responsible manner."

Some higher education institutions have also adopted a clear policy regarding self-archiving of research outputs, and these policies are gathered at the SHERPA JULIET service[6]. However the majority have not, and the degree of awareness of open access issues, preceding the adoption of a mandate to promote the approach, is at best, patchy.

In the case of both institutions and funders, there are a plethora of issues associated with socio-cultural, legal, technical infrastructure and funding requirements to be examined. All of these aspects need to be considered if the emerging vision of an open and data-centric research environment is to be achieved. The elements of this vision have been described in several recent publications including 2020 Science (Microsoft)[7] and a themed issue of Nature (March 2006)[8].

With the aim of realising this vision, in the US, the NSF has set up the Office of Cyberinfrastructure and the developing *Vision*[9] document includes description of a Data Cyberinfrastructure which adopts the taxonomy of data

collections defined in an earlier (2005) NSB Report on "*Long-lived digital data collections*"[10]. The importance of national and international collaboration in developing infrastructure to support data collections is recognised, and a "national digital data framework" which includes institutions and other organisations which manage data, is proposed.

In parallel, e-Infrastructure developments in the UK are progressing and the DTI announcement[11] of a Large Facilities Council to be formed in 2007 from a merger of CCLRC and PPARC, is indicative of the requirement to plan for a scaling up of e-research activity and support, in coming years. A UK data infrastructure with clear identification and understanding of the roles and responsibilities of its component services and organisations, will be an essential element in the implementation and exploitation of data-centric science within a Science Commons[12] in the 21st century.

The next three sections of this short paper attempt to raise some of the areas which need to be addressed, if this vision is to be achieved.

Socio-cultural, organisational, political and legal issues

The research community can be considered as the producers, authors, creators and re-users of data, either directly or indirectly through instrumentation or computational methods. However this community is highly diverse in awareness, practice and skills and there is a need to understand the full spectrum of research practice, workflows and associated data flows both within and between disciplines/sub-disciplines:

- What is the range of research practice involving data creation, capture, deposit, publication, citation, preservation, use and reuse within and across disciplines?

- What are the workflows? What elements are automated? Which transactions are human-mediated? Can we adequately describe the business process of e-research?

The "softer" aspects of data curation, many of which are the result of long-established social behaviours associated with disciplinary custom and practice, also need to be understood:

- What is the level of awareness of the need to deposit data in a managed archive? Awareness of long-term preservation requirements?
- Are data centres effective at promoting their services?
- Are disciplinary archives well-known and used by the community?
- How much data produced as a result of Research Council funded projects is deposited in a data centre / managed archive? (as mandated by funding?)

- Are researchers willing and able to deposit their data in a repository or data centre?
- Do they have the necessary skills?

- What are the professional development / training needs?
- How do researchers acquire these skills?
- Are these skills embedded in the student curriculum?
- Is there scope for new curation roles: data scientist?

- What are the barriers to deposit? (privacy, confidentiality, IPR, licensing, consent, maintaining/assessing quality, competition, etc.)

- Are there particular barriers associated with data in certain disciplines?
- How can more flexible licensing arrangements such as Creative Commons / Science Commons approaches assist in removing barriers?
- What mechanisms for indicating data quality can be implemented as QA guarantees for (re-)users and consumers?
- What are the incentives for deposit? (publication requirement, data validation, funding requirement, institutional mandate, nominations, prizes, research assessment metrics).

Some funding organisations such as the Wellcome Trust, NIH, MRC and AHRC have data-sharing and/or data deposit policies in place; data centres such as the San Diego Supercomputer Centre (SDSC) in the US have formal user agreements to determine allocations[13]; some higher education institutions are beginning to develop mandates for open access self-archiving of research outputs. Many institutions do not have such arrangements; indeed there is emerging evidence that many researchers are "self-sufficient" and do not use the various data services and information expertise that are available to them. The data sharing, curation and management policies of the different stakeholders need to be clarified in order to identify good practice and highlight gaps:

- What funding organisation data policies are in place?

- What data centre data/user policies are in place?

- What institutional data policies exist?

- How have the various policies been agreed? how often reviewed? What are the compliance procedures to ensure these policy requirements for data curation are met?

- Is the researcher self-sufficiency approach good-enough?

Federation models, interoperability and inter-relationships between repositories

The repository landscape is becoming increasingly complex with disciplinary repositories, national, institutional, laboratory repositories, services for particular media etc., making interoperability both within and across sectors a highly desirable, but very challenging goal. This distributed repository

landscape is characterised by different models of data flow. For example, the CERN Large Hadron Collider[14] (LHC) repository could be described as a "centralised repository" with data generated by the on-site LHC facility and locally stored, but used by researchers around the world. The new Diamond facility[15] at CCLRC in the UK may have a similar data flow.

In contrast, the protein databanks such as UniProt and the genome databases accessed through Ensembl[16] and managed by EBI, receive data deposits from many researchers as part of their professional working practice. The collected sequences are then used as a global reference collection, with a distributed annotation service adding value to the data.
In large-scale astronomy through the International Virtual Observatory Alliance[17], sections of the distributed data collections created from remote telescopes or sky surveys, are made available within the community, and selectively mined in order to make new discoveries such as identifying anomalies in the data, which might mark the presence of a failed star or brown dwarf. There are many different patterns of data use:

- What are the data flows associated with particular disciplinary data collections?
- Are there differences between small and large scale science?
- What are the data flows associated with institutional data repositories?
- How can we characterise these data flows?
- Are these data flows automated? Grid-enabled? How are they managed?
- What happens to the re-combined/re-processed results data? How are these curated and preserved?

There is a developing digital infrastructure for resource discovery, publication, curation and preservation, associated with the JISC Information Environment. Further e-infrastructure services are emerging as elements of Grid computing systems and VREs. The e-Framework potentially provides an over-arching structure for describing the diverse services and associated standards for data curation and preservation. In a complex federated and cross-sectoral landscape of repositories, there are many technical barriers to service interoperability and data sharing:

- Are there common data models and metadata schema in use within disciplines?
- Are there registries where data services and /or schema are published?
- Are there models of good practice which demonstrate the potential of data sharing?
- What metadata harmonisation and normalisation is occurring?
- What persistent identifiers are used? Domain identifiers?
- Are there established vocabularies to describe datasets in particular communities? Examples of inter-disciplinary practice?
- How is versioning managed?
- How is provenance tracked?

- How much duplication of data in repositories occurs?
- How is this managed?
- Does it matter?

Defining roles and responsibilities

There are many stakeholders who have a role(s) and associated responsibilities (R&R) related to data curation, but these are generally not clearly articulated and many questions arise:

- What are the R&R of a Research Council funded data centre? Are they documented?
- What are the R&R of an institutional data repository? Documented?
- What are the layered roles within an institution and who is responsible for what? At the level of Institution? Information Services? Department / School? Laboratory / research group? Individual?
- What are the R&R of funded services such as the Digital Curation Centre[18]?
- What are the funders responsible for?
- What is the role of organisations such as CODATA?
- Who sets the policy / policy definition for a data centre / service?
- Are formal contracts / MoU / agreements required between users and a data centre?
- Value of carrot and stick approaches?
- Is there joint planning between funders and Service providers?
- Is there joint strategic planning between Facility managers and data centre managers? Between funders of both?
- Do the current funding arrangements for data curation services provide adequate provision for sustainability to ensure long-term access and preservation to data collections?

Brief for Workshop Delegates

In summary, the charge for the workshop delegates is to:

- Gain clarity in understanding the current landscape of institutional data creation and management activity, and its relationship to active curation and preservation by data centres, data banks and other data archives.

- Identify and unpack the issues and challenges faced by funders and the community in this area.

- Begin to develop approaches and solutions to address these issues.

- Make recommendations to the JISC on ways to move forwards.

Whilst recognising that to make major progress with this ambitious agenda will require significant collaboration and more formal partnerships together with

joint forward planning by the key stakeholders, all of which will take some time and require both funding and effort, it will be useful at this early stage to highlight any "quick wins" that can immediately demonstrate the value of such partnerships.

References

[1] JISC DigiRep Wiki
http://www.ukoln.ac.uk/repositories/digirep/index/JISC_Digital_Repository_Wiki

[2] NIH Public Access Policy http://publicaccess.nih.gov/

[3] RCUK Position Statement http://www.rcuk.ac.uk/access/index.asp

[4] MRC Statement on Data Sharing and Preservation Policy http://www.mrc.ac.uk/index/strategy-strategy/strategy-science_strategy/strategy-strategy_implementation/strategy-other_initiatives/strategy-data_sharing/strategy-data_sharing_policy-link

[5] OECD Communique January 2004
http://www.oecd.org/document/15/0,2340,en_21571361_21590465_25998799_1_1_1_1,00.html

[6] SHERPA JULIET service
http://www.oecd.org/document/15/0,2340,en_21571361_21590465_25998799_1_1_1_1,00.html

[7] 2020 Science, Microsoft publication http://research.microsoft.com/towards2020science/

[8] Nature 440, 7083, 23rd March 2006 http://www.nature.com/nature/journal/v440/n7083/index.html

[9] NSF's Cyberinfrastructure Vision for 21st Century Discovery V7 http://www.nsf.gov/od/oci/ci-v7.pdf

[10] NSB Long-lived digital data collections 2005
http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf#search=%22long%20lived%20digital%20nsb%22

[11] DTI http://www.gnn.gov.uk/environment/detail.asp?ReleaseID=216698&NewsAreaID=2

[12] Science Commons http://sciencecommons.org/

[13] SDSC http://datacentral.sdsc.edu/how_to_apply.html

[14] CERN Large Hadron Collider (LHC) http://lhc.web.cern.ch/lhc/

[15] Diamond at CCLRC http://www.cclrc.ac.uk/Activity/Diamond

[16] Ensembl http://www.ensembl.org/index.html

[17] International Virtual Observatory Alliance http://www.ivoa.net/

[18] Digital Curation Centre http://www.dcc.ac.uk/