Dealing with Data: Roles, Rights, Responsibilities and Relationships

Consultancy Report

Document details

Author:	Dr Liz Lyon, UKOLN, University of Bath
Date:	19th June 2007
Version:	V1.0
Document Name:	data-consultancy-report-final.doc
Notes:	

Acknowledgement to contributors

The author would like to thank the various people, who contributed to the report by completing an interview, making a presentation, attending the workshop or commenting on previous versions. The author takes responsibility for interpreting the answers and for any change of emphasis that comes with collating the viewpoints of the various contributors.

Acknowledgement to funders

1 Executive summary

This Report explores the roles, rights, responsibilities and relationships of institutions, data centres and other key stakeholders who work with data. It concentrates primarily on the UK scene with some reference to other relevant experience and opinion, and is framed as "a snapshot" of a relatively fast-moving field. It is strategically positioned to provide a bridge between the high-level RIN Framework of Principles and Guidelines for the stewardship of research data, and practitioner-focussed technical development work[1]. For ease of cross-reference, the number(s) of the relevant RIN Principle(s) are given against each of the recommendations[1].

The Report is largely based on two methodological approaches: a consultation workshop and a number of semi-structured interviews with stakeholder representatives.

It is set within the context of the burgeoning "data deluge" emanating from e-Science applications, increasing momentum behind open access policy drivers for data, and developments to define requirements for a co-ordinated e-infrastructure for the UK. The diversity and complexity of data are acknowledged, and developing typologies are referenced.

The synthesis of the collated findings is presented in eight major categories: Co-ordination and Strategy, Policy and Planning, Practice, Technical Integration and Interoperability, Legal and Ethical Issues, Sustainability, Advocacy, and Training and Skills. The author suggests that the JISC and the wider stakeholder community need to focus their activities in these areas, and makes the following Recommendations. The ten Recommendations in bold below, are viewed as being of **highest priority** for JISC, institutions and research funding organisations.

- **Co-ordination and Strategy**

REC 1. JISC should commission a disciplinary DataSets Mapping and Gap Analysis, with associated curation and preservation support infrastructure. (1, 5)

REC 2. Research funding organisations should jointly develop a co-ordinated Data Curation and Preservation Strategy to address critical data issues over the longer term. (1, 5)

REC 3. The Strategic e-Content Alliance should consider adopting a facilitation role in promoting a cross-sectoral strategy for data curation and preservation. (1)

REC 4. JISC should develop a Data Audit Framework to enable all Universities and colleges to carry out an audit of departmental data collections, awareness, policies and practice for data curation and preservation. (1)

---

[1] For reference, the five RIN principles are:

1. The roles and responsibilities of researchers, research institutions and funders should be defined as clearly as possible, and that they should collaboratively establish a framework of codes of practice to ensure that creators and users of research data are aware of and fulfil their responsibilities in accordance with the principles set out in this document.

2. Digital research data should be created and collected in accordance with applicable international standards, and the processes for selecting those to be made available to others should include proper quality assurance.

3. Digital research data should be easy to find, and access should be provided in an environment which maximises ease of use; provides credit for and protects the rights of those who have gathered or created data; and protects the rights of those who have legitimate interests in how data are made accessible and used.

4. The models and mechanisms for managing and providing access to digital research data must be both efficient and cost-effective in the use of public and other funds.

5. Digital research data of long term value arising from current and future research should be preserved and remain accessible for current and future generations.

REC 5. The DCC should create a Data Networking Forum where directors/managers and staff from research council data centres, JISC data services and other bodies, can exchange experience and best practice. (1)

- **Policy and Planning**

REC 6. Each research funding organisation should openly publish, implement and enforce, a Data Management, Preservation and Sharing Policy. (1)

REC 7. All relevant stakeholders should identify and promote incentives to encourage the routine deposit of research data by researchers in an appropriate open access data repository. (3)

REC 8. JISC should commission a scoping study to investigate current practice, assess future potential and evaluate the curation and preservation issues associated with sharing research data through social software forums. (3)

REC 9. Each funded research project, should submit a structured Data Management Plan for peer-review as an integral part of the application for funding. (1, 2)

REC 10. Each higher education institution should implement an institutional Data Management, Preservation and Sharing Policy, which recommends data deposit in an appropriate open access data repository and/or data centre where these exist. (1, 2, 5)

- **Practice**

REC 11. JISC should extend the work of the DCC SCARP project to increase the range of in-depth Disciplinary Data Case Studies on data management, curation, sharing and preservation. (3, 5)

REC 12. JISC should commission a scoping study to evaluate the processes and issues around data and metadata capture at source from a range of instrumentation and laboratory equipment, as part of the end-to-end research workflow. (2, 3)

REC 13. JISC should commission a study of the applicability of generic data models and metadata schema application profiles for scientific data. (2, 3)

REC 14. JISC working through the Common Repository Interfaces Working Group (CRIG) and domain partners, should investigate the feasibility of repository deposit APIs for disciplinary data. (3)

REC 15. There is a need to identify and promote scaleable and sustainable operational models for data deposit, which are based on co-operative partnerships with researchers and common standards. (2, 3)

REC 16. The JISC Scholarly Communications Group should collaborate with the RIN Communications Group, to review data publishing principles and good practice. (3)

REC 17. JISC should commission work to investigate the effectiveness and applicability of different mechanisms for managing access and data-sharing across disciplines. (3)

REC 18. More work is needed to identify integrated information architectures, which link institutional repository and data centre software platforms. (3)

REC 19. All relevant stakeholders should commission a study to evaluate the re-purposing of data-sets, to identify the significant properties which facilitate re-use, and to develop and promote good practice guidelines and effective quality assurance mechanisms. (2, 3)

REC 20. JISC should initiate a survey to gather user requirements from practising researchers to inform the development of value-added tools and services to interpret, transform and re-use data held in archives and repositories. (2, 3)

- **Technical Integration and Interoperability**

REC 21. JISC should work with domain partners to identify and promote the mechanisms which have been successful in achieving intra-disciplinary consensus on community data standards. (2, 3)

REC 22. An assessment of the effectiveness of registries and other infrastructural services for the development and adoption of community data standards, is needed. (2, 3)

REC 23. JISC should commission development work to investigate the application of identifiers to datasets and produce guidelines for good practice in data citation. (2, 3)

REC 24. There is a need for technical work to determine models and best practice for version control of complex datasets. (2, 5)

REC 25. JISC should work with other stakeholders to investigate different annotation models and standards for datasets, and to develop guidelines for good practice. (2, 3)

REC 26. JISC should fund repository technical development projects which build on OAI-ORE work and create robust, bi-directional interdisciplinary links between data objects and derived resources. (2, 3)

REC 27. JISC should fund technical development projects seeking to enhance data discovery services, which operate across the entire data and information environment. (2, 3)

- **Legal and Ethical Issues**

REC 28. JISCLegal should provide enhanced advice and guidance to the research community on all aspects of IPR and other rights issues relating to data sets. (2, 3)

REC 29. Work by JISC and the research councils, on developing model licences for data, should be co-ordinated so that a minimum set of standard licences are adopted more widely. (2, 3)

- **Sustainability**

REC 30. The JISC should work in partnership with the research funding bodies and jointly commission a cost-benefit study of data curation and preservation infrastructure. (4, 5)

REC 31. The JISC should commission work to construct new economic models for preservation and data sharing infrastructure, to develop sustainable solutions. (4, 5)

- **Advocacy**

REC 32. The DCC should promote co-ordinated advocacy programmes, targeted at specific disciplines, and which address technical aspects of data collection and deposit. (2, 3)

- **Training and Skills**

REC 33. The DCC should collaborate with other parties to deliver co-ordinated training programmes and supporting materials, targeted at researchers in specific disciplines, to build workforce capacity within the sector. (3, 4)

REC 34. A study is needed to examine the role and career development of data scientists, and the associated supply of specialist data curation skills to the research community. (3, 4)

REC 35. JISC should fund a study to assess the value and potential of extending data handling, curation and preservation skills within the undergraduate and postgraduate curriculum. (4)

Key chronological dependencies within the set of Recommendations are also noted. The disciplinary DataSets Mapping and Gap Analysis (Rec 1) will inform the development of a co-ordinated Data Curation and Preservation Strategy (Rec 2) by research funding organisations. The Data Audit Framework (Rec 4) will ideally be in place to form a common basis for the development of institutional Data Management, Preservation and Sharing Policies (Rec 10).

The proposed roles, rights, responsibilities and relationships emerging from the findings have been presented in a Summary Table (overleaf).

Two new high-level data flow models, which derive from the Summary Table, and which illustrate contrasting exemplars of good practice, are described: a Domain Data Deposit Model and a Federation Data Deposit Model.

Role	Rights	Responsibilities	Relationships
Scientist: creation and use of data	Of first use. To be acknowledged. To expect IPR to be honoured. To receive data training and advice.	Manage data for life of project. Meet standards for good practice. Comply with funder / institutional data policies and respect IPR of others. Work up data for use by others.	With institution as employee. With subject community With data centre. With funder of work.
Institution: curation of and access to data	To be offered a copy of data.	Set internal data management policy. Manage data in the short term. Meet standards for good practice. Provide training and advice to support scientists. Promote the repository service.	With scientist as employer. With data centre through expert staff.
Data centre: curation of and access to data	To be offered a copy of data. To select data of long-term value.	Manage data for the long-term. Meet standards for good practice. Provide training for deposit. Promote the repository service. Protect rights of data contributors. Provide tools for re-use of data.	With scientist as "client" With user communities. With institution through expert staff. With funder of service.
User: use of 3rd party data	To re-use data (non-exclusive licence). To access quality metadata to inform usability.	Abide by licence conditions. Acknowledge data creators / curators. Manage derived data effectively.	With data centre as supplier. With institution as supplier.
Funder: set/react to public policy drivers	To implement data policies. To require those they fund to meet policy obligations.	Consider wider public-policy perspective & stakeholder needs. Participate in strategy co-ordination. Develop policies with stakeholders. Participate in policy co-ordination, joint planning & fund service delivery. Monitor and enforce data policies. Resource post-project long-term data management. Act as advocate for data curation & fund expert advisory service(s). Support workforce capacity development of data curators.	With scientist as funder. With institution. With data centre as funder. With other funders. With other stakeholders as policy-maker and funder of services.
Publisher: maintain integrity of the scientific record	To expect data are available to support publication. To request pre-publication data deposit in long-term repository.	Engage stakeholders in development of publication standards. Link to data to support publication standards. Monitor & enforce public. standards.	With scientist as creator, author and reader. With data centres and institutions as suppliers.

2   Introduction

UKOLN was asked to undertake 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 work has been completed by the UKOLN Director.

2.1   Terms of Reference and Objectives

As stated in the brief from the JISC, the consultancy objectives are:

- *To define how institutions (collectively and individually) and scientific data centres can together effectively achieve:*
    - o *Preservation*
    - o *Access – Managed and Open*
    - o *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.*

2.2   Audience

The primary audiences for the report are:

- the JISC Executive
- the Repositories, Preservation and Asset Management Advisory Group
- the relevant JISC Committees
- the wider community.

The report will also be made available from the JISC and UKOLN Web sites.

2.3   Scope

It is perhaps useful at this point to make some comment on the perceived scope and positioning of the Report. The area of work defined is part of an ambitious agenda, which perforce can only be addressed at a high level in this modest study. There is a growing volume of developmental projects and research activity underway in this context, both in the UK, in Europe, North America, Australasia and elsewhere, however to date it is somewhat fragmentary and constrained by local geographic and political boundaries.

This Report concentrates primarily on the UK scene with some reference to other relevant experience and opinion. It is inevitably "a snapshot" of a fast-moving field where the rapidly increasing creation of scientific and research data, raises very significant challenges for researchers, data managers, policy-makers and funders alike. It is strategically positioned to provide a bridge between the high-level RIN Framework of Principles and Guidelines for the stewardship of research data, and more technical development work.

The Report seeks to identify the range of issues which need to be explored in more depth and makes a number of Recommendations, some of which themselves are substantive items, for further research, investigation and discussion. It also aims to present some examples of good

practice and to describe disciplinary models and approaches, which may be transferable to other domains, sectors and contexts.

The organisations interviewed were selected to provide a broad range of views, and the physical sciences, social sciences, arts and humanities and bio-medical sciences are represented. Finally, it is acknowledged that there are gaps and a degree of variation in the depth to which some issues are explored. Where this is particularly evident, some indication has been given in the text together with a note of what additional work needs to be carried out. These pointers are also reflected in the Recommendations.

3 Methodology

The work has been largely based on two methodological approaches outlined below and supplemented by desk-based research. This report provides a synthesis of information and opinion gathered throughout the study with additional analysis and commentary.

3.1 Consultation Workshop

This Workshop was 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".*

Accordingly, an invitational workshop entitled "*Exploring the roles and responsibilities of data centres and institutions in curating research data*" was held at Rutherford Appleton Labs, STFC (WAS CCLRC), Didcot, on Tuesday 10th October 2006. The charge for the workshop delegates was 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.

The format of the programme included a number of speakers and small group discussion with a plenary session to close. The speakers represented the various stakeholder groups: data centres, data service providers, research councils / funders, institutions and publishers. The full programme and presentations are available at http://www.ukoln.ac.uk/projects/data-cluster-consultancy/ .

A short Briefing Paper was produced for the meeting, in order to raise issues and inform the discussion. This is available at http://www.ukoln.ac.uk/projects/data-cluster-consultancy/briefing-paper/briefing-final.pdf .

A total of twenty delegates attended the workshop and they are listed in the Appendix.

3.2 Interviews

Twenty semi-structured interviews were subsequently held with selected representatives of the various stakeholder groups. The outline pro-forma used as a basis for the interviews is included in the Appendix together with the list of interviewees.

4 Context and Vision

The vision of data-driven science was first articulated in November 2000 by Professor Sir John Taylor, then Director-General of the UK Research Councils:

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

"e-Science will change the dynamic of the way science is undertaken."

Further context is given in this text from the National e-Science Centre Web site:

"In the future, e-Science will refer to the large scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large scale computing resources and high performance visualisation back to the individual user scientists. The World Wide Web gave us access to information on Web pages written in html anywhere on the Internet. A much more powerful infrastructure is needed to support e-Science".

The challenge of the "data deluge" resulting from e-Science was discussed in a seminal paper by Tony Hey and Anne Trefethen[2]. It highlights the need to annotate, archive and curate data so that it and its associated programs, can be reproduced and re-used in the future.

In 2003, an eScience Data Curation Report[3] was commissioned by the JISC Committee for the Support of Research, which described a range of findings and made a number of strategic recommendations, many of which are highly pertinent to this Report. The authors Lord and Macdonald found that:

"there is a lack of government-level overall strategy for data stewardship and data infrastructure"

"provision of curation is patchy and more advanced in some disciplines than others"

"awareness of the issues….is generally low among researchers"

and recommended that:

"Strategic level advocacy for data curation is needed…."

"The mismatch of short-term funding against the long-term needs for data retention needs to be addressed…."

"Funding bodies should consider supporting research-led exemplars of curation…"

The Report also endorsed the need for a Digital Curation Centre (DCC) and this organisation was set up in 2004 funded by the JISC and EPSRC, and led by the University of Edinburgh, with partners at the University of Glasgow (HATII), UKOLN at the University of Bath and the STFC (WAS CCLRC). The DCC has been proactive in promoting data curation, providing advice and services to the community, developing tools and implementing a research programme.

The eScience Core Programme and the various application projects funded by EPSRC have developed the original vision and in July 2004, the UK Treasury, DTI and DfES published the *Science and Innovation Investment Framework 2004-2014*[4] which proposed a national e-infrastructure for research. The Office for Science and Technology (OST) set up a steering group to take forward discussion and a study was commissioned in 2005 to inform the creation of a high level "roadmap". The following is taken from the report published in February 2007[5]:

"*E-Infrastructure is the term used for the distributed computing infrastructure that provides shared access to large data collections, advanced ICT tools for data analysis, large-scale computing resources and high performance visualisation. It embraces networks, grids, data centres and collaborative environments*".

"*A national e-infrastructure needs: the means of producing, managing and preserving vast amounts of digital data; sophisticated means of accessing an ever-increasing range of electronic resources of all kinds; technologies and structures to support dynamic and virtual communities of researchers; unprecedented network, grid and computational capacity; and the necessary national services and systems to ensure safe and secure access to resources. We*

believe that these and other requirements presuppose not only a high level of integration and coordination, but also, in key areas, intervention at the policy level."

This clearly positions data centres (and data), as elements of e-infrastructure and recognises the requirement for strategic direction and policy drivers.

Following the UK lead, similar e-Science initiatives have ensued in the US through the National Science Foundation (NSF) Office of Cyberinfrastructure with its emerging *Vision*[6] and in Australia from the Australian Research Information Infrastructure Committee (ARIIC) through their e-Research Strategic Framework *Interim Report*[7] and in the *"From Data to Wisdom"* Report published by the Prime Minister's Science, Engineering and Innovation Council (PMSEIC) in December 2006. A recent report in *Nature*[8] described emerging plans in the US to bring together federal agencies such as NSF and NASA in a new Interagency Working Group on Digital Data, which will develop strategic plans for a public infrastructure for data.

The initial e-Science vision has been extended and described in two further publications: *2020 Science* (Microsoft, 2006)[9] and a themed issue of Nature (March 2006),[10] which both addressed the growing role and impact of computing and computer science in science.

In parallel, there has been a developing awareness of the challenges of managing the data outputs from "small science." This has to some degree been fuelled by the growth of social software and collaborative technologies, leading to the concept of "open source science" or "open source research". These approaches were explored in an essay by Philip Ball in *Nature*[11], the challenges of "Small Science" have been highlighted in an article "*Lost in a Sea of Science Data*" in the *Chronicle of Higher Education*[12] and the developments have been described in a keynote presentation by the author at the 2nd International DCC Conference[13]. A growing number of Web sites such as OpenWetWare[14] are facilitating the sharing of methods, outputs and resources to support this more open way of doing research.

Whilst centrally funded data centres provide expert curatorial facilities for the deposit, management and re-use of research data in some disciplines, institutional repositories are developing as an alternative location to deposit research outputs. It is fair to say that to date, most emphasis has been on the deposit of textual outputs ie eprints, however some organisations are beginning to explore the practical benefits and value of digital repositories for the storage of primary research data, and pioneering projects such as the JISC-funded eBank project[15] [16] have illustrated the potential of this approach. Examples of both types of provision and support will be described in this Report.

It is evident that a co-ordinated 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[17] in the 21st century.

4.1 Policy drivers for open access data

A number of international initiatives focussing on promoting open access approaches in scientific publishing (Budapest Open Access Initiative 2001, Bethesda Statement 2003, Berlin Declaration 2003) pre-date more recent developments to promote open access data. However these statements have been highly influential in their impact on current thinking and policy-making. In January 2004, the Organisation for Economic Co-operation and Development (OECD) published a Communique stating ten principles upon which open access to research data from public funding was proposed (the OECD Declaration)[18]. The principles include openness, transparency, legal conformity, protection of intellectual property, formal responsibility, professionalism, interoperability, quality and security, efficiency and accountability. This statement has since been developed by an international Expert Group with representation for the UK by Mark Thorley, NERC and a draft *OECD Recommendation concerning access to research data from public funding* has been produced. This document (October 2006) includes a set of *Principles and Guidelines* which provide broad policy recommendations to governmental science policy and funding bodies.

In June 2006, the UK Research Councils (RCUK) 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[19]. An Invitation to Tender for this independent study is expected to be released in April and the report will be available in 2008, when the RCUK position will be reviewed. The position with regard to open data are less clear and will be explored by the RCUK Research Outputs Group (ROG), which is a cross-council forum for discussing, co-ordinating and promoting councils' policies and activities supporting wider and better use of RC-funded outputs, including data.

The Research Information Network (RIN) has published its *Goals for Public Policy – Scholarly Communications Statement of Principles*[20] in February 2007 and is proposing a *Framework of Principles and Guidelines* for the stewardship of digital research data, which draws upon the RCUK Position Statement and the OECD Declaration, and aims to support a collective approach to avoid inconsistencies and duplication. This Framework of five principles will be published in 2007.

A comprehensive study commissioned by the RIN to investigate the policy and practice of UK research funders in managing their research outputs has been published in January 2007[21] and it provides an overview of positions of the different funder groups and includes reference to research data. A brief summary of research funder open access policies is also available at the ROARMAP site[22]: specific references to data are described in more detail in the Section 5.1.

The European Commission held a major conference in February 2007, at which the future of scientific publishing was discussed. At the event, a Petition[23] supporting free and open access to European research was handed to the Commissioner. There was much lively debate on a range of issues which included some discussion on provision of open access to data-sets and associated data curation and preservation requirements. The European Commissioner for Science and Research Janez Potocnik, announced the allocation of dedicated and substantive funding to support the establishment of digital repositories for storing scientific data and for digital preservation initiatives[24].

In addition, other organisations and charities 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. A diagram summarising OA compliance with Wellcome grants through deposit of research outputs in UK PubMedCentral/PMC is available on the Wellcome Web site[25] however there is no specific indication of best practice with regard to associated primary data.

4.2   UK Data Centres and Data Services

The UK Research Councils fund a number of data centres which provide expert curation services for the increasing volumes of primary 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.

The DTI has announced the creation of a new Science and Technology Facilities Council[26] which was formed in April 2007 from a merger of STFC (WAS CCLRC) and PPARC, and is indicative of the requirement to plan for a scaling up of e-research activity and support, in coming years.

The JISC also funds four data services including 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 (some of which are available through a licence arrangement). Three of these services, (MIMAS, AHDS and UK Data Archive), also receive funding from the Research Councils. This report will examine a small number of selected data centres and services in more depth, providing a range of disciplinary perspectives.

4.3 Institutional data repositories

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 Digital Repositories Roadmap[27] noted that:

"institutions need to invest in research data repositories"

"we need functionality and services that support curation, migration and preservation"

"the community needs to develop mechanisms that foster advocacy, mandates, funding, early adopters and demonstrators with designated responsibilities aggregated at discipline or regional level".

The report also stated that*:*

"culturally, data repositories have been the property of "scientists" and there is some tension between the data and information community. Institutional repositories could fill a gap where there is no data archive…"

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[28]. The Repositories Research Team is developing an ecology of repositories and some preliminary results were presented at the JISC Conference in March 2007[29].

Some higher education institutions have also adopted a clear policy regarding self-archiving of research outputs, and these policies are gathered at the ROARMAP service[30]. 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.

4.4 The diversity of data

Data can be viewed conceptually as a social construct evolving from a theoretical basis, and many layers of subsequent interpretation act upon and transform the data. It is crucial to recognise that data are both complex and heterogeneous, though there are a number of generic categories that can be applied. The NSF OCI *Vision*[31] 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*"[32]. These broad categories are: research collections (such as local laboratory project data), community collections (such as FlyBase) and reference collections (such as Protein Data Bank). Data collections may grow and evolve, and migrate from one category to the next over time. 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.

EBI and NERC refer to canonical data (data which has minimal variation) and episodic data (changing data e.g. in life of a cell), which may be unique in time and place e.g. climate info. A further categorisation is into raw, processed, derived data and metadata. These groupings are recognised by the new draft ESRC policy, but all are viewed as "resources". Some broad data definitions were presented by IUCR: raw (e.g. image on a plate or file), primary (e.g. structure factors), derived data (e.g. six-dimensional structural model), reflecting the domain of crystallography. Other groupings gather specific types of data: "omics", observational, simulations, multimedia, surveys, performances, computational, software etc.

5   Findings

The outcomes of the interviews are presented in this section, grouped by funding organisations, data centres and data services, digital repositories and other key stakeholders.

5.1   Funding organisations

The RIN-funded Research Funders' Policies Report published in January 2007 contains a useful summary of the broader position across institutions, funders, publishers and other stakeholders and considers outputs other than data. The current study has focused specifically on data-sets and builds on selected aspects of this report; however the landscape is constantly changing, and these findings represent only a snapshot from October 2006-March 2007.

Five funding organisations representatives were interviewed: 4 Research Councils and one major charity. Information about another research council was gathered from an associate. Of the Research Councils examined, AHRC, ESRC and NERC directly fund dedicated data centres. MRC contributes an element to the total funding of the European Bio-informatics Institute (EBI). EPSRC does not directly support any data centres. The Wellcome Trust also contributes to EBI.

5.1.1   Engineering and Physical Sciences Research Council

The Engineering and Physical Sciences Research Council (EPSRC) has a broad statement on access to research outputs[33] and the responsibility is on the grant holder to develop the best means of storing and making their data available. There is an expectation that their data will be made available in an OA repository but no guidance is given on which repository option(s) might be most appropriate and the Conditions of Grant do not describe that level of detail. The Web statement refers to the RCUK ROG study to be commissioned shortly and which is referenced earlier in Section 4.1; EPSRC is awaiting the outcomes of this study before making any changes to existing procedures and this decision has been taken at the highest level in the Funding Council.   Evaluation of a funded project through the peer-reviewed Final Report is expected to highlight any data and dissemination issues. IPR issues associated with research data outputs are also deemed to be the responsibility of the grant holder.

5.1.2   Medical Research Council

The Medical Research Council (MRC) has a Data-sharing and Preservation Policy[34] and is now focussing effort towards the community and stakeholders in order to realise the policy. The Council is in the process of developing Principles for Access to and use of MRC-funded research data. These will complement the MRC Data-sharing policy and will be supported by both general guidance and more specific guidance and information resources relevant to the different types of research that MRC funds. These Principles are currently being developed and will be shared more widely at an early stage to seek community feedback and to work towards consistency across research funders. The principles will require testing and there are specific areas where there are clear challenges, such as issues of data confidentiality, data protection, use of special licences and use of pseudonyms which have implications for technical implementation of identifiers and contextual linking. There are useful parallels with developments in other domains such as the social sciences, which are being leveraged to inform thinking.

There are clearly some issues around responsibility for retention of research data and the duration for preservation of clinical and non-clinical data. Whilst within MRC-funded units, there is a corporate responsibility to curate data, for MRC-funded researchers within institutions, the onus is on the institution to allocate the physical space. There are also issues of selectivity and identifying criteria for selection, aside from the question of which repository in which to deposit data-sets. Data centres may be used as 3rd party services in this context. There are also requirements for some flexibility around access agreements: different levels of access may be required depending on the user.

The MRC funds a number of population-based studies where huge amounts of data are collected in longitudinal studies over many years. Examples are the Avon Longitudinal Study of Parents and Children (ALSPAC)[35] based at the University of Bristol co-funded with the Wellcome Trust, and the National Survey of Health & Development (NSHD)[36]. The MRC is also funding strategic projects such as UK BioBank in partnership with the Wellcome Trust. In depth case studies of these types of study will provide vital information to inform strategic decision-making and cost-benefit studies. Over the long study period, there have been many changes in perspective, and evaluation of studies of this sort are likely to demonstrate how access to and usage of the data from a multidisciplinary basis, is beneficial to UK research. Earlier surveys of reuse of data[37] suggest that there was far less data re-use than might be desirable and that in the past, data was created without wider sharing in mind: there was a lack of information (or metadata) to enable re-use.

Figure 1 Allan Sudlow, MRC.
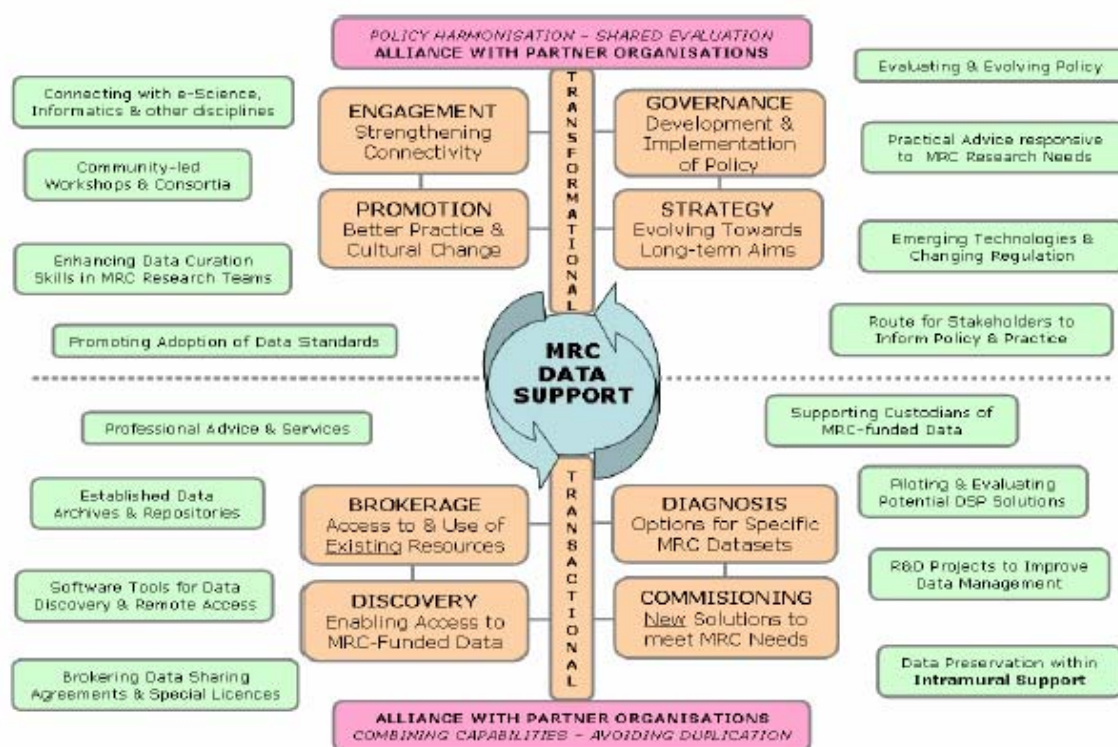
The MRC also funds a significant amount of "omics" activity where there are well-established models for large-scale data-sharing (see Section 5.2.4), however in other areas of research, models are less well-established e.g. clinical trials data and there are legal and ethical issues around the access, use and re-use of the data. A report on large-scale data sharing was published in 2004[38]. MRC is also involved with the new Diamond Light Source in partnership with STFC (Was CCLRC) and Wellcome, where the first scientists to use the synchrotron have recently begun work (February 2007).

MRC is currently exploring partnerships to deliver a range of data support services to its communities: the service will be funded for a 2-year feasibility stage with a fully-costed business plan for a 5-year forward horizon. Web-based guides on data curation, covering proposal planning, data and metadata management and long-term preservation policies will be published shortly. This highlights an important issue for funding agencies: the awareness and skill level of researchers. The Digital Curation Centre is producing a *Curation Manual*, organising professional development events and producing Briefing Papers and Case Studies. However, requirements vary across disciplines and there appears to be a need for more advocacy and

practical hands-on guidance targeted at specific disciplines and sub-disciplines. In some cases researchers are "struggling" with challenges of using a common nomenclature or ontology and have not yet thought about issues such as multiple deposit, duplication of resources or data migration between repositories.

Whilst there are incentives to carry out some degree of planning to produce the Data Preservation and Sharing Plan within any MRC proposal, which should also include costs for data management, this is seen as a pragmatic approach. Project final reports also need to include feedback on what has been implemented in this area. A mix of skills are required which may be missing within particular research teams. There are implications for training, career progression and support in this context: the role of the data scientist has been proposed as one way forward. Frequently the required skills reside in different communities and there are often cultural challenges in getting diverse groups together to discuss and work in partnership. We will return to all these points in Section 6.

5.1.3 Economic and Social Research Council

The Economic and Social Research Council (ESRC) is currently redeveloping its existing data policy which will be renamed the *Research Resources Policy* to reflect a broadening of scope to include resources other than data. The Economic and Social Data Service (ESDS) is the main repository for ESRC-funded outputs and researchers are very strongly encouraged (mandated) to offer their data to ESDS, however ESDS is not under obligation to accept it. A close working relationship is maintained between the ESRC and the ESDS with a Case Officer allocated to the partnership providing a link between ESRC planning and ESDS infrastructure. A further connection is provided by the Research Resources Board who advise on the strategic direction of the data centre and associated funding. This forms a vital link between Programme planning and support infrastructure for the outputs of programmes. ESRC strongly supports researchers depositing data in a data centre or repository to enable them to be accessed by other researchers, but recognise that the growth in institutional repositories as potential recipients of data outputs, mean that the landscape is more complex and further thought is required on best practice. It is intended that the ESRC Policy development will feed into the RCUK ROG discussions.

The ESRC recognise the need for clear and consistent messages as part of effective advocacy work. Researchers will forget, not have sufficient time, reject any additional costs and may fail to see the benefits of data deposit, and these are all viewed as barriers. Promotion and advocacy work is of great value and a range of approaches such as speaking to stakeholders at conferences, have been adopted. It is important to note that stakeholders include academics and government: the requirement to engage these parties so that the advantages and disadvantages of open access approaches are well-understood is stressed.

5.1.4 Natural Environment Research Council

The Natural Environment Research Council (NERC) has a science budget of around £370M and data are seen as essential to support the NERC mission of research, survey and monitoring. Many environmental data are episodic in nature and irreplaceable, having a unique position in time and place. The data supports knowledge transfer, income generation and the sustainability of the science base and is of huge value to mankind in particular relation to global environmental change and man's impact on the environment. Data holdings are viewed as one of NERC's core assets. There is a perceived hierarchy of drivers for data management:

Level 0 – deliver project, Level 1 - meet "good scientific practice", Level 2 - support own science (which should include QA processes to allow others to use the data), Level 3 - support employers requirements, Level 4 - support funder requirements, Level 5 - support public policy requirements, as described in the OECD Principles and Guidelines.

The NERC has a Data Management Co-ordinator responsible for co-ordinating the work of the seven designated NERC data centres[39], maintaining the Data Policy[40] and reviewing its application at both operational and strategic level, developing OA work and co-ordinating science information strategy. The term "designated" has a special meaning within the policy: it

implies responsibility for long-term curation of NERC-funded data, application of the data policy and support for it within their subject domains. The Data Policy dates from 2002, is being reviewed and will be re-published in October 2007. NERC is actively planning for developing its data infrastructure. It does not intend to create any new data centres but is formulating plans for rational data distribution, identifying gaps and looking at the longer term for certain data collections. This Review is being carried out in partnership with other relevant data centres such as EBI.

NERC is aware that capacity-building programmes with new grants are scheduled and there needs to be forward planning to assess how data produced from these new programmes will be managed, what value will it have and what partnerships are required for maintaining a cost-effective data infrastructure. The experience of studies in the "omics" area has been informative and lessons learnt suggest that data management requirements must be addressed at a much earlier stage than has previously been the case. NERC is developing a new 5 year science strategy and one element of this strategy will be a "knowledge strategy" with much more substantive statements on issues around data and information. This is part of an aim to be more strategic in thinking and planning, in order to have adequate infrastructure to support future science. The requirement to use, re-use and link large datasets ("big science") will require additional skills around curation, informatics and the ability to join data across different disciplines and sources. There is a perception that this skills base is currently not present within the community.

NERC datacentres date back to 1996 in parallel with the implementation of the data policy. Consquently they are relatively well-established with the lead centres such as the British Atmospheric Data Centre (BADC) serving large communities and with clear and accountable funding lines. This is seen as crucial in developing clarity in supporting programmes of work. Smaller centres share staff and infrastructure resulting in economies of scale. Some data centres are based within NERC Research Units and are funded out of their core budgets with primary responsibility to that centre, rather than to the community at large. The Geoscience data centre is allied to the British Geological Society and a has a much larger commercial perview than other centres because of the way in which the data and information is licensed with many commercial customers, some of whom may be the general public wanting to know about subsidence around a prospective house purchase. Licensing data are a key income generator.

The varied nature of NERC funded programmes (e.g. responsive mode, directed mode) has also led to program data centres which are funded for the life of the programme: an example is the Environmental Bioinformatics Centre at the Centre for Ecology and Hydrology at Oxford. NERC also use the Archaeology Data Service at York for data relating to science-based archaeology. This pragmatic approach to data deposit is an interesting model which crosses disciplinary funding boundaries and works to promote the most cost-effective and practical locations for data collections based on disciplinary needs. The data centres also have a role in fulfilling NERC responsibilities for environmental science FOI requirements.

Formal links between the data centres and NERC are through the Information Strategy Group (ISG) which is currently being reviewed, and which reports to the NERC Executive Board. ISG also receives reports from the Data Management Advisory Group. ISG takes a wider view which includes libraries and IT as well as the research units which host some data centres. There is a mixed funding model: some data centres receive investment from the programme over the length of the programme, others are entirely funded by the host research unit. The funding model tends to work well when data centres have a clear funding line and a clear list of deliverables (33% NERC budget of around £370M goes into its research units). The data policy states that any data outputs from NERC-funded research at institutions, have to be offered to the appropriate data centre for long term curation.

NERC runs three basic funding models for its research programmes:

1. Directed mode: these are large multi-million pound programmes which have a science steering committee with sufficient budget allocated for data management activity e.g. QUEST project (Quantifying and Understanding Earth Systems) run from the University of Bristol with whom BADC works closely with researchers around the UK to provide support based on a Data Management Plan. Support is purchased from BADC in this example and

there is one lead centre (Bristol) which acts as the primary point of contact. It was noted that this process has taken some time to be fully-implemented but that 80% NERC data are collected from this mode of operation.

2. Consortium grants: these are large consortial research proposals around £3M FeC but where NERC has not developed the science plan, which has been developed by the community. NERC recommends that sufficient resources must be allocated for data support to be purchased from appropriate data centres, however once the funding has been distributed there is less control over allocation. Evidence suggests that the earlier that data management is embedded in the research programme, the better long-term outcomes ensue. The aim is for PIs to fully cost a "data manager" post and put money aside to cover data management costs.

3. Responsive mode (standard grants): these are smaller grants run by a variety of PIs. NERC is exploring funding models to work across research areas for allocation to datacentres in a top-slice mode. Data centres need to show value for money in their services to PIs. Data deposit compliance is a difficult area: there is a big overhead in monitoring compliance. Other research councils are considering various sanctions for non-compliance, however it is deemed more effective in the longer term to win scientists' hearts and minds: requirement for publication is a real incentive in this context.

NERC expects to hold the primary copy of any data but is not concerned about duplication e.g. in an institutional repository, provided data has been offered to NERC first. Expert scientific curation is viewed as essential for secure long term curation of the data. From NERC perspective, there does not appear to be any evidence of duplication at the current time. It is most common for data to be deposited at the close of a project or programme.

The NERC Grants Handbook contains details of policies and practice and the grant is viewed as the contract between the grant holder and the data centre. Compliance is not formally overseen and it is not clear what percentage of funded data are actually deposited in NERC data centres.

NERC does not make any claim to IPR except when the data are from a NERC-funded research centre. Most data are perceived to be of little commercial value with the exception of the geological data mentioned earlier.

The "right of first use" is recognised in the NERC data policy since there is a cultural issue around the "my data" issue. Scientists are not prepared to let others use their data until they are fully published: the embargo time is not defined. This is perceived to be the greatest barrier to data sharing. Data protection, consent, confidentiality issues are considered to be less important. In general researchers with data holdings are keen to work with the data centres. Indeed. NERC is being approached by researchers asking for advice on data management. The generation of younger researchers is perceived to be much more aware. Relationships between data centres and IRs may be enhanced through linking of publications to underlying data. Demonstrating value for money is difficult: citations of reuse of data may be the most effective measure. NERC is developing a DataGrid to enhance interoperability and discovery across all data centres

The nature of data publication appears to be an issue in terms of what this really means, how it can be "published" and what it means for long-term curation. This is related to any linking between the peer-reviewed article and the dataset to maintain the scientific record – an area in which publishers also have an interest. In the omics world, a data accession number is required before data can be published. (Note however that even this mechanism is not fool-proof)[41]. This is viewed as a useful way forward so that scientific good practice to maintain the scientific record and data to support it, is recognised by the community. NERC are developing a licence to publish and are documenting roles and responsibilities in this context.

5.1.5 Wellcome Trust

The Wellcome Trust (WT) funds a wide range of research in the biomedical area and has been a leader in the drive towards making the outputs of publicly funded research openly available. It has a Policy & Advocacy Team within the Strategy Planning & Policy Unit, which is positioned in the centre of the organisation. These strong links with strategic and operational planning are

valuable in ensuring that OA activities are embedded in practice and disseminated to influence the external environment. Funded outputs and outcomes are assessed and evaluated to enable Wellcome to monitor the effect of policy transitioned into research activity. This has led to questioning around the evaluation assessment of research outputs on the Web e.g. citation analysis and the value of WT grant number quoted in acknowledgements in papers with very many authors, where Wellcome has been a funder but exact attribution is difficult to unpack. It has been recognised that the structure of research papers requires review. It is impossible currently to get a view of the impact of Wellcome Trust funding, since the term is used in many ways in papers. OA is perceived to have raised many issues associated with the ability to carry out evaluation, examine impact and inform strategy. The creation of a database or a patent, is viewed as a legitimate output of funded research; qualitative measures such as associated professional development of postgraduates and post docs, and wider social benefits are difficult to measure with the current model. Effective publishing of research outputs on the Web is seen as key to this, which may be interpreted as improving access to research, but also to use it in different ways such as making the whole text searchable, linking from a compound referenced in UKPMC to Pubchem to other articles referencing the compound; linking to gene structures, proteins, semantically linking outcomes from different research activities.

It is acknowledged that a huge cultural change is required in order to realise this vision, both amongst researchers and publishers. Researchers are perceived to have not yet embraced or fully understood the principles of OA, though younger researchers are more aware of the possibilities of the Internet and its more open approach. Publishers are operating a very wide range of rights agreements; even within a single publisher there may be multiple policies. The position is even more fragmented for data. Genomics is seen as taking the lead, setting a precedent and establishing a clear position. The preferred choice for sequence data are to ensure high quality and release on day of production (or as soon as possible thereafter): the human genome project is a good example. The cancer genome project has a delay before data release. In general this approach is supported by the community, and academics in genomics understand the concept and value of sharing data resources. Six-seven years ago there was a huge debate about releasing the human genome data, but actually the data are now seen as relatively "straightforward" and can be accessed and re-used with ease.

In contrast, population-based data and patient data are more controversial because of issues around confidentiality, privacy, IPR, 3rd party access and re-use. The Biobank Project is a good example where there are issues around where to store data, how to separate researchers from participants, 3rd party access, who should pay, access levels for different types of users (academics, pharmaceutical industry etc.), IPR and retaining rights, licensing, royalties etc. It is not clear which is the best approach for the public good. An Access and Intellectual Property policy is currently in development.

The WT Policy on Data Management and Sharing was released in January 2007[42] and includes a requirement for a Data Management Plan if the creation of a community resource is a primary goal, or if there is expected to be a significant quantity of data which could potentially be shared for added benefit. The Plan is required to take into account data quality and standards, use of public data repositories, intellectual property, protection of research participants and long term preservation and sustainability. The Trust also expects all users of data to acknowledge the sources of their data and abide by the terms and conditions under which they accessed the original data. This has implications for how datasets are cited, whether in a data centre or in a digital repository, and how access parameters are described. Guidance is provided in a Q&A document, which aims to clarify the position in certain areas such as population-based studies. Plans will be peer-reviewed as an integral part of the refereeing process. The MRC and WT are liaising on these issues. The position with regard to where to deposit researcher data when not falling within the categories listed above, is less clear. More work is needed in this area.

WT is finalising its ethics and legal governance framework relating to how data are stored, managed and accessed. A pragmatic solution is envisaged aiming to get "blanket consent" with additional scrutiny of proposals and a pilot has started to test the framework. The public have been found to be more open than expected but the Trust is funding public attitude work relating to governance of patient records and patient data with a view to informing funded work. The population research community does not have a culture of sharing data and a range of models

are being explored as a potential solution to enable further exploitation of the data. These include use of embargo periods, managed release according to need, a club / closed group model, data enclaves.

It is recognised that the infrastructure to support data sharing is hard to fund and there are many questions around long-term sustainability of such infrastructure. Various economic models may be applied but at what point do researchers pay to use it triggering the change to a cost-recovery model? There is a view that so-called "dark archives" (archives that are either completely inaccessible to users or have very limited user access), are not ideal because if data are corrupted over time, this is not realised until point of use. There is a view that institutions need to take responsibility for the outputs of their research and that research strategies should include guidance on how to manage outputs, how to preserve and archive data. There is a perception that much data resides more appropriately at the institutional level, potentially in institutional repositories. Sustainability is seen to be a key issue to be examined in more depth.

Infrastructure also requires technicians with specialist skills to maintain and curate the data. The concept of a data scientist is viewed positively, however there are concerns around their career structure. They do not usually produce research papers but may be acknowledged if the resources support the specific hypothesis of the paper. If the resource is community-focussed, then the data scientist will not be acknowledged. Recognition of the data scientist is an issue.

5.1.6 Arts and Humanities Research Council

It was not possible to interview a representative of the Council itself, however views were obtained from an academic "associate". The typical product from an AHRC Resource Enhancement grant is an online multimedia database maintained by the investigators on their own Web site and institutional system. The AHDS as the funded data centre, are able to take the underlying data but not the system. This has inevitably meant that the considerable intellectual input in the user interface and the software, is basically lost, and there may be hardware emulation issues at a later date. In addition, real concerns were expressed about the ability of most institutions or researchers to maintain the online system for more than a few years, leading to a loss of investment in resources over the medium to longer term. In many cases, the academic working with the research assistant, does not have the time to update the resource and the research assistant then leaves the institution. The potential of a network of expert centres, where members would be joint developers with the academic, and co-applicants in any research proposal, was put forward.

The format in which resources are deposited is of great importance, and the AHRC has been trying to "tighten up" the terms for grant awards for resource creation to ensure materials are created to appropriate technical standards. A "sustainability appendix" was suggested which would include collaboration with a data centre or network of expert centres. However there are related issues of maintaining the currency of resources, since the award monies must be spent within the grant period. The concept of a "dowry" to continue to maintain resources for up to ten years was proposed. A "large minority" of researchers in the arts and humanities create data, and are seeking due credit from the RAE for data publication: inclusion of a requirement to deposit in the grant letter may provide the pressure for action.

Appropriate subject expertise is seen as essential for data creation and curation and there is a challenge in dealing with multiple institutional repositories at different levels of refinement and sophistication. The JISC was viewed as having a key role in providing a national infrastructure, with generic solutions for versioning and access issues. Institutional repositories need to provide a sustainable environment for data publication, but data publication projects also need to be informed by subject expertise, to ensure that the resource is optimised from the methodological point of view. In the past, many A&H resources have not been created with re-use envisaged: there was no user testing and resources may not have been created in the most useful form. Many were constructed on a "cottage industry" basis.

The AHRC will continue to fund resource creation projects, but through the normal grants procedure, and in future, proposals will need to have a good research project attached, not just the creation of the resource. There are perceived issues around strategic co-ordination of digitisation: funding agencies need to work more closely together in this area. It was noted that

the AHRC was also planning to have a digitisation programme, and many of the JISC digitisation projects were A&H based. Co-ordination of this sort has been raised previously, for example at the British Academy e-Resources Workshop in 2006.

5.2   Data centres and data services

Interviews were held with Directors or senior representatives from five data centres. Four of the data centres and data services receive funding from multiple sources including research councils, the JISC and the Wellcome Trust. One data centre has a single research council funding source. One further data service is based within an institution and is supported by that institution, but also receives funding from The National Archives (TNA).

5.2.1   Arts and Humanities Data Service

The Arts & Humanities Data Service (AHDS)[43] is a mature data service co-funded by the JISC and the AHRC with its Executive located at Kings College, London. It comprises five distributed subject centres based at higher education institutions (HEIs), which cover Archaeology, History, Literature, Languages and Linguistics, the Performing Arts and the Visual Arts. A wide range of content is received including images, raw and processed datasets which include structured text and numeric / statistical materials, video, sound, geographic information, boundary data, CAD materials, VR files and animations. Whilst the content tends to be smaller scale when compared to some e-Science data sets, there are 5-10Tb moving image collections which are complex in nature, as well as interactive Web site materials with embedded links and multimedia.

The AHDS has roles in enhancing access, facilitating dissemination, providing managed digital storage and long-term preservation. The AHDS has preservation policies, disaster recovery plans, ingest manuals, preservation handbooks and delivery manuals. The AHDS has been operating with its own custom-built platform but is beginning to experiment with FEDORA based on its flexibility and functionality. The strategic model is to use FEDORA to manage the delivery of content whilst iRODS[44] developed by San Diego Supercomputer Centre (SDSC), will be used to manage preservation i.e. a dark archive which sits beneath the repository. The AHDS have been working with STFC (WAS CCLRC) to assist with the installation of Storage Resource Broker (also developed by SDSC), and will develop a contract and formal agreement for provision of a dark archive using the Atlas Data Store.

The ingest process is described in a manual and is a two-stage process where the subject centre negotiates with the content owners and prepares materials for dissemination and deposit. Preservation is carried out centrally by the Executive in a documented set of procedures. The latest version of a resource is usually made available to the public but earlier versions are accessible but kept in the dark archive. All materials given by a depositor are kept "as received" and there always at least two versions (for dissemination and preservation).

For the deposit of research data, AHDS currently works and negotiates with the individual researcher rather than with the institution, although the institution may sign the licence giving permission. Discussions are continuing around this area with the view that it may be beneficial to work more at the institution / departmental level. Whilst for AHRC or JISC funded data the position is clear, for data funded from elsewhere or unfunded datasets, then there is value in negotiating with the institution especially in the case of smaller arts institutes, and informal discussions are taking place to pursue these ideas.

An interesting partnership model with researchers is being explored within the Stormont Papers Project[45] where the data capture has been undertaken at The Centre for Data Digitisation and Analysis at Queen's University Belfast and the materials are then made available by the AHDS. This is viewed as an attractive and scaleable model, which is well-suited to the FEDORA platform, since tools can be customised to work with different types of content e.g. data-sets. Data capture costs can be built into the budget for each grant to enable the customisation to happen, providing a potentially sustainable model across federations of FEDORA repositories in the future.

Most of the data which is deposited with the AHDS, does not currently get deposited in an IR, however a recent example of an image collection from the visual arts was deposited at the AHDS but the organisation (Imperial War Museum), subsequently implemented an IR and deposited the collection in that location too. AHDS has Waiver of Deposit Policy forms for the AHRC and the British Academy. This was recently instigated when a researcher was developing materials within DSpace and wished to keep them there. However there is a view that AHDS would seek to harvest the metadata about that resource since the AHDS provides a cross-search service and a detailed catalogue record is required to support this process. This is an "odd case" currently but is likely to become more common for text and images in particular. In some cases if the institution is delivering the collection from a server and has taken a preservation copy then it may be more efficient simply to have a metadata record with a persistent link back to the institutional site. The AHDS would then take over delivery of the collection if the institution can no longer provide support, but how does the AHDS know that the collection is no longer available? In order to implement a formal policy in this area, reliable mechanisms are required to manage the process. Checking updates and managing version control would be more difficult when materials are held on a remote server.

Currently such instances are managed on an ad hoc basis with an almost "tacit policy" of avoiding duplication because they occur only infrequently. But it is clear that such examples will increase with the growth of IRs and it is recognised that a policy will be required in the future. IRs are currently viewed as being less suited to managing more complex materials, but these are also aspects which may change with further technical development and experience. For example, crawler software may be needed to alert data centres to updated versions of resources held on remote sites. In this context, the SHERPA DP2 Project is specifically looking at providing a preservation layer for IRs working with research data. Clearly there are many opportunities and models for productive partnerships between data centres and data services, and institutions with IRs, but there is also a need for clear delineation of roles and responsibilities, which will need to be explored on a case-by-case basis using specific examples to avoid confusion over vocabulary and use of generic terms.

The AHDS Deposit licence does not transfer IPR to AHDS but it remains with the researcher or owner. The licence is in perpetuity to disseminate and preserve the material and make associated changes required, and may be terminated with 6 months notice on either side. The AHDS is very careful to ensure that there are no copyright issues over materials deposited. Material would be rejected if there were any outstanding issues and this is the primary reason for rejecting materials. The onus is on the researcher to obtain the required clearances; certain disciplines such as the performing arts are particularly challenging in this regard.

The successful operation of the AHDS relies to some extent on good co-operation with individual researchers and their willingess to deposit: not all AHRC-funded researchers will release their data. There are challenges in acquiring materials in an adequately structured and documented state since the time and effort required to curate the resources are considerable. The guidance requests fairly comprehensive metadata for re-use but some researchers will simply offer the Final Report or a badly photocopied set of papers. This suggests that there is a lack of curatorial responsibility amongst the research community. It also highlights a requirement for funding organisations to take seriously the need to preserve research data with individuals with the requisite skills to curate to professional standards. This has implications for strategic planning and budgeting and also for the professional development and career progression of data curators.

There are also tensions between the need to use and re-use data and the requirement for long-term preservation to agreed standards. More advocacy is required to convince researchers that producing data to appropriate standards is good use of their time and beneficial for research, learning and other applications. There is also a perceived training requirement or provision of technical support. Terms such as "research technologists" or "digital technicians" with strong technical skills to assist researchers may be needed, if only for a finite period of time or for certain disciplines. The AHRC-funded ICT Methods Network may have a role to play in this regard.

A more collaborative and co-ordinated approach is needed with partnerships between JISC, the research councils and institutions to determine high level strategy for data curation and preservation, building on the UK's leading position. The DCC is well-positioned as a catalyst and facilitator in this context. There is also huge scope for greater cross-sectoral activity since much of the material for arts and humanities research comes from outside of higher education, from the cultural heritage sector where there is less funding. Links with TNA and MLA should be leveraged and there is perhaps a role for the Strategic e-Content Alliance here. Some TNA work in digitising collections has been completed in commercial partnerships and as a result, significant research data are still not accessible for academic purposes e.g. 1891 census.

5.2.2   UK Data Archive

The UK Data Archive (UKDA) which receives funding from the JISC and the ESRC, is a designated place of deposit for the National Archives and is located at the University of Essex. It is a mature service dating back some 35 years and from the start has embraced a remit which emphasises providing access to data simultaneously with secure storage and preservation. The UKDA fulfils an official requirement for TNA and NDAD for storing machine readable public records and records of government, and for ESRC, for storing electronic records generated from ESRC funding. The UKDA distinguishes itself from a digital repository as a result of these stringent requirements from funding bodies and national institutions.

The UKDA stores and makes accessible long-term statistics and database materials, but also images, sound and textual information. It is the lead partner for the ESDS and it houses the AHDS History Subject Centre, which manages a wider range of formats. Both raw and processed data are stored in disciplines such as economics, politics, sociology, history, geography: it is estimated that 80% data could be used to inform policy. There are also potential intersects between the social science and medical research areas especially within public health and demographic epidemiology. The UKDA is working with the MRC to bring more cohort data into the collection.

The main file format is generic for numeric data (SPSS or SAS or Stata or Excel or tab delimited text) and most of the data are numeric and coded. Materials range from statistical material to relational databases with 2-100 tables, to images, vector graphics, and sound, although the latter are diminishing in quantity. Paper is also ingested with machine-readable records, since a number of data collections may have paper code books: these are digitised at ingest. The UKDA uses a mixed platform approach for infrastructure and has up to four different media storage types. There is guidance provided on the Web site concerning data deposit, data access and data management including preservation and the substantive Preservation Policy document is available[46]. The data management and preservation practice follows the workflow of OAIS.

A deposit form is online to assist researchers with the deposit process and they are requested to deposit data in a certain format to streamline processing: 70% material is acquired in this way. The UKDA together with ESRC is moving towards a life-cycle oriented approach where ESRC notifies UKDA within the first 3 months of award of grant to enable evaluation of the potential of long term preservation of the outputs. The award holder would then be contacted early on in the process before data has been collected, to advise on the preservation schedule and to assist in identifying and overcoming issues of consent, ethics, confidentiality, copyright, formats and metadata, so that these issues can be resolved before they become a problem. This can be achieved with ESRC funded data but the aim is to also take this approach with government bodies and the TNA. This has been very instructive and one example is the Department of Work and Pensions who are planning a new survey so UKDA have been discussing preservation planning with them at the start, to identify datasets worthy of preservation early in the lifecycle. This represents a very positive culture shift.

There is a Licence Agreement to deposit in the UKDA and compliance with rights clearance is taken on trust. Universities vary in their approach to rights: some are very strict and some are less so. These are of course issues for materials deposited in IRs in particular with primary data and its re-use. If you create a dataset based on another researcher's data, the right to do that may be assumed. The UKDA must consider primary rights in all materials and this is monitored

closely. The Licence assigns to the University of Essex the right to disseminate and publish the data and is relatively strict when compared to the more open world of Creative Commons. This has been recognised, but UKDA sometimes deals with highly sensitive data e.g. for police cases, and there is a Special Licence for use in particular instances.

Researchers are under obligation from both the ESRC and the AHRC to deposit their data with the UKDA under the terms of the grant contract, however some data are acquired from other sources and the UK Public Records Act underpins the preservation applied to government datasets. In this role, the UKDA acts as a facilitator both for the research councils, but also for the researcher. The UKDA has a pro-active and strategic collections policy and works with stakeholders to determine what is stored. Some material is rejected. The policy is fluid and dynamic and is supported by an Acquisitions Review Committee which considers demand in the short and medium term with a view to assisting the resource consequences of bringing the collection into the archive. The scale of ingest work and rights clearance can be underestimated by the researcher, and there must be justification for bringing a collection into the archive. This acquisition needs to be planned as a part of the research. If the data has been collected by another agency or research unit, then this information is also needed in order to investigate the possibilities for adding the data to the archive. There is scope for more dialogue with the ESRC in this context.

Decisions for new Programmes are made within the ESRC and this information needs to be communicated to the UKDA, in order to have the data resources available to support them. The UKDA works to the Research Resources Board but if a research theme is developed elsewhere there is a danger that UKDA are not kept informed. The highly structured nature of the research council organisations sometimes leads to inflexibility. The UKDA has five-year funding streams so strategic planning happens within this timeframe and can predict with confidence in order to have appropriate capacity and effort to support requirements. However occasional ad hoc projects which generate vast quantities of material which have not been forseen, would be problematic and additional funding support must be sought from the project in these cases.

Coherence of the collection is perceived to be a key issue and it is an important historical institutional issue that virtually every social science researcher in UK HE and also some outside, know and respect the UKDA. IRs have little social sciences research data: this is because there is a belief that most of the data creators and users use the "gold star" quality UKDA service. It is interesting to observe that the University of Essex does not have an IR. There is perceived to be little duplication of research data between an IR and UKDA and this is likely to be because of significant work at ingest. There is also a view that with enhanced discovery tools, the location of research data becomes irrelevant provided there is the technical infrastructure to guarantee access over the long term.

The high quality of the service is viewed as an incentive for use by researchers and the culture for deposit is well-established. For over two decades, ESRC led the thinking around open research investments and long-term preservation. It is well-embedded in the culture because quantitative social science and increasingly qualitative social science, cannot be done without data sharing at one level or another. Much social science requires access to data that no single researcher could produce themselves. The main source of data disseminated is not ESRC funded data but government data, such as large scale surveys which cost millions of pounds to produce. Much of social science is underpinned by the notion of data created by others unlike many other disciplines.

Barriers to collaboration are seen to be resources (curation is a cost-intensive process) and copyright. Universities are becoming increasingly concerned about any liability that may arise in connection with research completed by their staff. The data creator is having less say in what happens to their data and outputs. There is a perception that university contracts officers are developing unworkable conditions to protect themselves, even when they have signed the ESRC contracts in the first place, and this is a growing concern.

The UKDA also has an important outreach function in making people aware of what is in the archive and a training role in order to maximise people's use of the data. It has proved beneficial to get creators and users together led by demand for materials: the Census Registration Service goes out and explains how to get the data and how to use it by example.

There is also a need for better guidance for both the researcher and the user about legal obligations with respect to the data. There is a role for JISC here in promoting a better understanding of these issues, which would perhaps enable repositories to be better populated with materials and for use of those materials to be enhanced. This is linked to the new role of the institution as publisher.

In addition, the UKDA has a clear and explicit relationship with the University of Essex: UKDA is not a legal entity. An IR may not have the same depth of relationship with an institution so there is a much greater risk associated with deposit of items with rights issues e.g. Ordnance Survey data. These are areas which require more investigation and guidance. JISC has a role to play in this context.

There are also issues around metadata creation. There may be conflicts around the use of tags by a researcher and the application of structured metadata and terminology in a standard fashion. It may be necessary to supplement what a researcher proposes, and use terminology from a recognised thesaurus such as HASSET (Humanities and Social Science Electronic Thesaurus). This facilitates better mechanisms for deep access for domain experts.

5.2.3 British Atmospheric Data Centre

The British Atmospheric Data Centre (BADC) is a designated data centre funded by NERC and based at the Rutherford Appleton Laboratories (RAL) at STFC (was CCLRC). The BADC is relatively mature as a data centre: its predecessor started in 1985. The BADC has a role in access, dissemination, managed storage and long term preservation. However it is also a source of community-based value-added services and tools which sit on top of the data and which encourage researchers to collaborate e.g. workspace tools. The BADC receives strategic direction from the Data Management Advisory Group at NERC and from the National Centres for Atmospheric Science Group.

BADC focus is on atmospheric data and associated data, including ocean temperature at the surface and chemistry data related to the atmosphere. Data are from a range of sources such as weather balloons, weather stations, the Met Office, numerical models, climate model data, satellite data, weather radar data and measurements from planes and balloons. There is a significant volume of data from the Met Office. The BADC has no duty to keep this for the long term however the data are used widely so the BADC acts in facilitation mode and much of the effort is in this area. The BADC also act in facilitation mode for ESA and NASA data. Data from NERC Directed Mode Programmes and from NERC facilities, such as radar and spectroscopy at RAL and from other NERC projects, is held. In Directed Mode there might be 10-20 projects in a programme producing data and some is stored at BADC. There is scope for greater join-up between the BADC and the grant allocation process and this is being addressed within NERC. The funding of data management amongst multiple small projects in Response Mode is an issue and there is a perception that a different business model is required by the research councils to support their data centres.

Many data files and formats are collected though there are two preferred formats NetCDF and NASA AIMS. There are also Met Office proprietary format (PP) and GRIB, a meteorological format for model data. BADC find it difficult to make the distinction between raw and processed data. The ingest process varies between datasets. If the dataset is a 3rd party one, then liaison with the Met Office is required e.g. for weather radar data, and this is an example of the facilitation role. BADC has policies for dealing with different data types e.g. simulations.

Versioning of datasets is very complex for BADC data holdings. A dataset is generally a large group of files and each file may have versions. A dataset may be a specific instance of a group of versions of data files but for most datasets, there are limited changes. However some data changes in real time by the minute, some have official versions from the original source and some may have been processed. There is scope for more work on versioning of complex datasets. There are also granularity issues around use of identifiers: what exactly is being identified? Is it the whole dataset or a version or a subset or a data granule? BADC uses internal local identifiers and there is more work to be done in this area.

For NERC Thematic Programmes, BADC is requested to write a Data Management Plan for the whole programme, which involves a survey of the funded projects and this happens at the start of the programme with input from the Steering Committee. There is less input to smaller funded projects. Data from projects are sometimes received in a gradual submission process and sometimes in large chunks; usually there is a chasing process when a project is near to closure. This need to follow up on acquiring the data are very time-consuming and compliance is an issue. Academics involved in thematic programmes tend to be supportive because they can see the advantages and that use of common formats allow them to share and use other people's data. There is usually a need to "convert" newly funded researchers.

Within the Data Management Plan (DMP) is a Protocol which acts as an agreement between the Programme and the BADC i.e. a sharing arrangement. The agreement is with an individual and not an institution. The varied conditions of employment within institutions and the associated IPR issues are a cause of concern. An individual by signing the NERC grant award allows NERC the right to use and share the data but it remains their intellectual property. There is a separate agreement with the Met Office about how to share and store their data which is based on their standard sharing contract with appendices and conditions. There is scope for more guidance on IPR issues.

BADC have a set of Selection Criteria based on usability and usefulness:

General Data Selection Criteria

Usability of Dataset Evaluation Section

Quality of Data - Data should have sufficient error and quality information to judge how the data can be applied.

Usable Data Format - A standard format using standard conventions with appropriate read software available.

Conditions of Use - The conditions of use allow practical use of the data and are set to appropriate levels.

Reputable Author - The author or organisation producing the data are authoritative. The data are traceable to that author.

Documentation - Information is available detailing production, coverage, ownership, authors and other metadata.

Usefulness of Dataset Evaluation Section

Data Quality - Time/Space coverage is large with few gaps and high resolution. There are no known corrupt files or errors within the data.

Uniqueness of Data - There is no other data available as an alternative. The data needs a primary archive or it will disappear.

Current Usage - There is currently a large number of users accessing the dataset or likely to use the dataset in the future.

Potential Strategic Use - Current and planned research programmes are likely to use the data.

Usefulness of Parameters - Parameters are likely to be used outside of the projects that collected them.

BADC has embraced the OAIS Model and has re-organised according to the OAIS workflow. This is viewed as a good model. BADC is responsible for preserving data from NERC-funded programmes in perpetuity but this is under a contractual arrangement. A contingency plan is required to deal with passing the data to another site if required. There needs to be more work on the preservation role of 3rd party services and movement of data between them.

The BADC requests researchers to reformat their data into an acceptable format for ingest and this is seen as a barrier to deposit. There is scope for greater advocacy and awareness-raising work and a role for data centres such as BADC to market and promote their services more pro-actively. PIs are visited to make the case for deposit. The collaborative tools to enable data

sharing with other colleagues are an incentive and seen as useful in this context. BADC has some experience of trying to "sell" access to the data holdings, in particular the Met Office datasets and this has been effective in the past because people want this data. It is harder to persuade researchers to deposit their data. Workshops at conferences have also been held with some success. Some NERC service produce continuous data from aircraft facilities and this is found to be more useful to the community. Sporadic data from projects is less useful.

BADC is not aware of many researchers who are depositing datasets in an IR and there is a perception that very few IRs could manage the formats, and that BADC and IRs are mutually exclusive with regard to the data holdings. However there has been a long term need for a better way of managing the documents that accompany the data e.g. a project Final Report, and that these documents should be linked to the associated data. A document repository is seen as a possible solution to this issue and the CLADDIER Project has been investigating these types of issue. Interaction with the CLADDIER Project has also been informative with regard to the role of citations and the issues around populating these sorts of repositories. IRs are considered to be at a fairly primitive stage, however there is potential for IRs to hold data in the future. There are perceived risks, since most IRs are managed by librarians without deep subject knowledge, rather than science experts and there is a likelihood of technical problems with regard to formats and metadata management. It seems clear that policies will be required to determine which data are deposited in which data centre or IR, and to clearly set out the remit of an IR. There are also issues of trust: IRs are relatively new and their longevity is not proven.

5.2.4 European Molecular Biology Laboratory – European Bio-informatics Institute

The European Bio-informatics Institute (EBI) at Cambridge is a part of the European Molecular Biology Laboratory, which has its headquarters in Heidelberg. It is located on the Wellcome Trust Genome Campus which also houses the Wellcome Trust Sanger Institute and is one of the world's largest concentrations of genomics and bio-informatics expertise. EBI receives total funding of over 26 milllion euros from a range of funders including EMBL, the UK research Councils and the Wellcome Trust. The budget is projected to rise to 43 million euros in 2011. Users of EBI databases number around one million. The collection of data are a collaborative international process and there are many projects which include partners from HEIs both in the UK and beyond. Work is performed on collection databases of DNA sequences, protein sequences, microarray expression data, macro-molecular structures, protein interactions and biological pathways (Reactome). The oldest data resource is EMBL-Bank which dates back to 1985. The EBI opened in the mid 1990s and currently employs about 300 staff. There are integrative linkages and interactions between the databases. The EBI databases also have linkages to other data resources in areas such as biodiversity, medicine and chemistry. EBI has recently launched EBI-eye, a search tool which searches across all the EBI databases, Medline and Reactome.

This study will focus on the ArrayExpress datasets. Microarray experiments are used to generate large datasets which include gene expression, comparative genomic hybridization, genotyping and chromatin-immunoprecipitation experiments. ArrayExpress is an open repository for MIAME-compliant microarray data. MIAME is the Minimum Information About a Microarray Experiment that is needed to interpret the results of the experiment unambiguously and potentially to reproduce the experiment. The repository includes both raw and processed data. Submission of MIAME-compliant data to ArrayExpress has been adopted by most scientific journals as a condition for publishing a paper. Statistics from July 2006 show that ArrayExpress contains data from almost 50,000 individual arrays and this number tends to double every 12 months. Strategic direction for ArrayExpress development is provided by the Scientific Advisory Committee and senior scientists.

Figures 2 and 3 Slides from Graham Cameron, EBI.
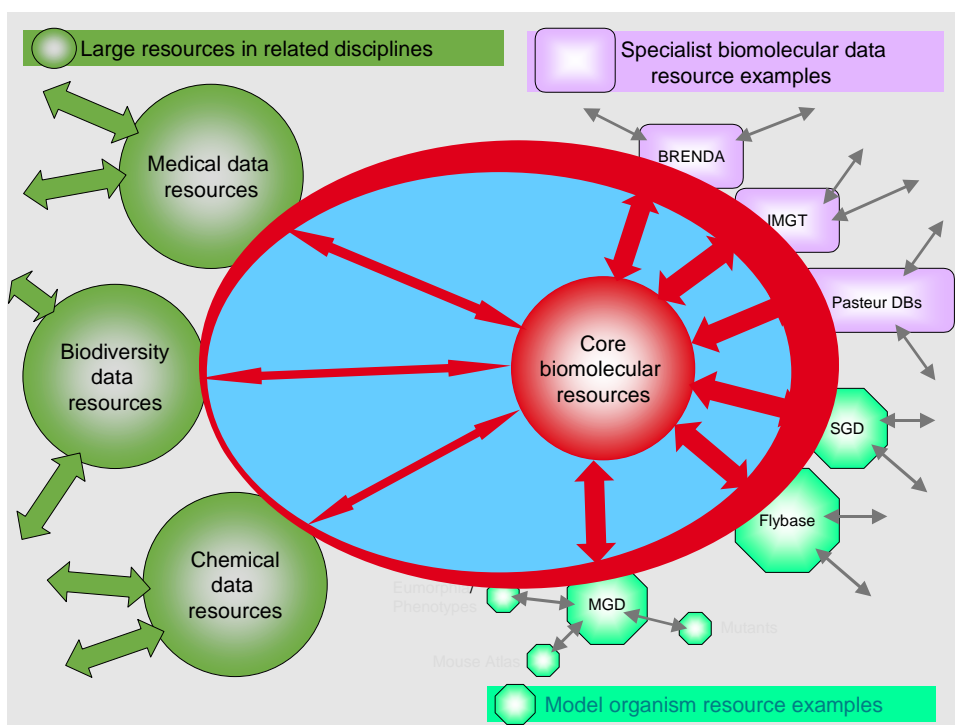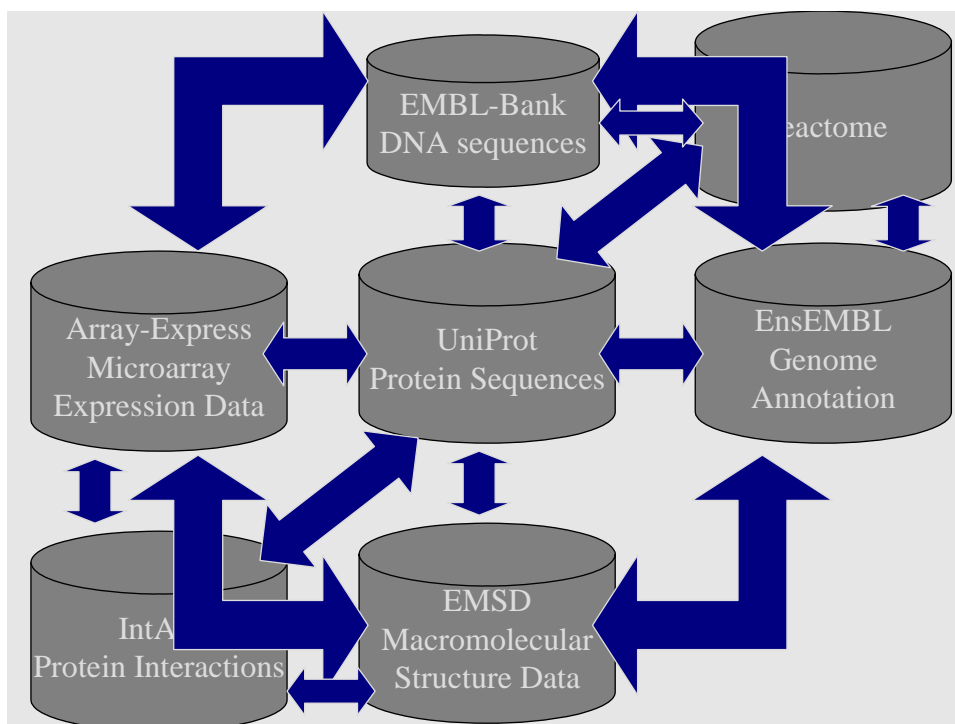
ArrayExpress has four major purposes: 1. to serve as a primary database for archiving data supporting publications or projects that generate public microarray data, 2. to provide easy access to gene expression data, 3. to facilitate the sharing of microarray data, array designs and experimental protocols and 4. to integrate microarray data with information in other relevant database such as UniProt and Ensembl. ArrayExpress allows storage of pre-publication data confidentially whilst allowing access to authorised users, such as journal editors and referees.

Upon publication of the paper, the microarray data are made public. Data can be annotated and re-annotated to update information about the genes represented on the microarrays. Data are submitted with the textual description of the abstract and validated annotations together with ontologies if available, and information about the platform, version of the software, all experimental protocols and data processing software. This provides a rich metadata set which is required for the computational processing across the data, and which adds value through data integration. Updates on datasets are handled within the data warehouse. The decision to create a new dataset or make an amendment is handled on an ad hoc basis and versioning is transparent. The Distributed Annotation System (DAS) has been developed at EBI and the Sanger Institute, founded on the principle of annotations being implemented on a distributed basis, and there is a standard model with protocols for 1st party (depositor), 2nd party (database group adding value) and 3rd party (release to the community so other scientists can add value), annotations of gene sequences.

A well-developed ontology has been developed within the microarray group. Researchers are encouraged to use the Gene Ontology (GO) but typically many don't which results in a need to re-annotate the dataset. Researchers tend to annotate the array at the start wheras annotation at analysis is of greater interest. This has resulted in a shift in workflow and a higher rate of change of annotation compared to 5 years ago. A new generic high-level data model with application-specific sub-models is being developed to deal with this shift. EBI are developing a workflow system (Taverna) to link all the systems together and this is a major development over the coming year. The aim is to work towards individual gene expressions. A curator-selected set of data and annotations relating to 150 species is held in the ArrayExpress Data Warehouse. The quality of data are high: a user scoring procedure based on set criteria in a recent study resulted in a 100% data met the required criteria. However this level of data curation is expensive.

MIAMExpress is the Web-based data submission tool for biologists; MAGE-TAB is a spreadsheet-based format for annotating and communicating microarray data in a MIAME-compliant fashion. MAGE-ML is an XML-based data exchange format developed by the Microarray Gene Expression Data Society (MGED), which is an international organisation that aims to facilitate the sharing of microarray data and annotations using accepted standards such as MIAME. Many microarray laboratories can export their data in MAGE-ML format from their internal databases or from their laboratory information management system (LIMS). Pipelines with array manufacturers have also been established facilitating automated submission of array data. Tab2MAGE is a standalone spreadsheet submission system for biologists who wish to use a spreadsheet-based submission process. The tools are a key aspect of the repository and the curators are also involved in the software development processs. ArrayExpress is an excellent example of streamlined data management, storage, curation and tools to facilitate high-throughput biology.

There are conditions for submission: each dataset has an accession number and data are kept private during the review period. Data takes about one month to "work up" and there are policies which give guidance. Frequently policies are developed with a consortium working on an initiative and included within the consortium agreement. Life Science Identifiers (LSIDs) are not used because they were unstable when ArrayExpress was set up. In the future if needed, datasets could be mapped to LSIDs. EBI recommend that authors cite array design identifiers in articles but most authors cite an experiment identifier. EBI holds data in trust but does not seek ownership of the data. Data are not released if IPR issues are unresolved. Data from other microarray databases such as Stanford, are also made available and there is a view that provided the data are managed to a common standard, there is no problem for users to hold their own data locally. The requirement for repositories to continually deal with new technologies has an impact and a fluid approach to managing data are required. Data integration is seen as the next major step.

EBI works on the principle of open data made available immediately, which is in conflict with certain confidentiality requirements. Not all data are peer-reviewed. If a paper is accepted by a journal then the data must be accepted too. There is a view that all data should not be treated as being correct. Some sequence data may not be associated with a paper and may not be peer-reviewed. Some data in the structures databases are incorrect and this is recognised by

the community. Whilst validation is carried out, it is not perfect. Re-sequencing can be carried out if required and citations will point to the correct sequence. The citation task is perceived to be a challenging one.

EBI do not make any formal statement about preservation however they hope to be able to retain every database entry in perpetuity (assuming that the data remains of relevance to the needs of the users). Clearly there is a funding implication associated with this aim.

Researchers who use ArrayExpress are usually well-informed and there is a social requirement for them from journals as they must deposit their data in order to publish. Data curators at EBI also participate in a pro-active training programme, which covers use of the tools and processes including annotation and re-annotation. Around forty tutorials or lectures are delivered per year either through invited speakers or courses run on a global basis. There is also an EBI Roadshow and EBI staff visit institutions such as MIT and SDSC and can tailor courses to suit requirements. All material is online and EBI is developing an eLearning platform. There are plans to ramp up training on-site with a dedicated training room in the new building.

5.2.5 University of London Computer Centre Digital Archives

The University of London Computer Centre (ULCC) provides digital archive and preservation services for 3rd parties and is involved in a number of associated R&D projects. ULCC also has a role within the University, advising on electronic records management and helping to get digital repository projects started. ULCC operates primarily as a service with 18 staff. Currently the largest in financial value and size of digital assets held is the National Digital Archive of Datasets (NDAD) service run for the National Archives. This is government data of which some is open access and available to the public and some is closed. The service operates almost entirely using inhouse software set up 10 years ago. The NDAD archives include datasets and born digital textual materials and scanned textual materials. The datasets are government records. Selection criteria are the importance of the data in demonstrating government activity and decision-making, rather than the quality of the data for researchers. This makes a case for preserving poor quality data because it supported government decision-making in the past. The archive includes one-off surveys e.g. contaminated land survey of Wales in the 1970s and the ongoing school census. It resembles the type of material within the UKDA and there has been duplication in the past but this was viewed as acceptable because of the different funding source, the UKDA had no responsibility to provide access indefinitely. However UKDA is now the place of deposit for public records and has a duty to preserve material in the longer term.

There are standard workflows and procedures for NDAD data however although these are well-documented, as new material is brought into the collection, some detail has to be "invented" and software written to deal with the new materials. The norm is to deal with a small number of large and complex items each year (30 datasets per annum on average). A single item might involve thousands of accompanying documents. Some items may include 1000 datasets whilst others may be very simple and fit on a floppy disc so there is a huge variation of scale. ULCC is exploring mapping workflow onto the OAIS model.

ULCC also provides a repository for the School of Advanced Studies SAS (an amalgam of research institutes), running on the DSpace platform and preserving the research outputs of the School. This covers recordings of public lectures as well as peer-reviewed publications. ULCC is also starting two new projects: one for the Museums, Libraries and Archives Council (MLA) preserving the outputs from the BIG Lottery Fund projects. These are about 400 community-based projects with resources largely on DVDs but with Websites, oral histories and some printed material. The main aim is preservation rather than access. The second project is for the Linnean Society and this is primarily digital images associated with specimens of insects and plants but also personal papers of Carl Linnaeus. Here public access is the key driver.

Content is varied as has been described and includes both raw and processed data. The workflow within the DSpace repository is standardised and the ingest process is largely implemented by academic staff acting as champions and ancillary support staff. They receive technical help from ULCC and have an additional person within the School of Advanced Studies who has an advocacy role. There is some direct deposit and some mediated deposit within the school. There are links with SHERPA LEAP which is a London-based alliance of repositories.

There are formal agreements for all services and some are very detailed with Service Level Agreements. As a result of the role in providing 3rd party services, ULCC endeavours to clarify IPR issues before the service is started and to ensure that ULCC is given the required licences to facilitate the preservation work. The organisations offering the contract need to understand their obligations so in the case of government data, licences were needed from HMSO. Some government bodies are not Crown Copyright e.g. Coal Authority, so specific licences are required from them and they may also hold material licensed by others.

For the institutional repository for the SAS, the department was made aware of the IPR issues and as a result, some of the material may not be provided on open access because of unresolved IPR issues e.g. recordings of seminars and public lectures. In this case, there are signed agreements giving permission to make the recordings but it is not clear what is permissible beyond that. Performing arts materials pose special problems for rights clearance.

Versioning has not been an issue because most material is not changing and in any case has been considered from an archival viewpoint i.e. occasionally an item has changed over a period of years and it may be considered as a different item which in hierarchical description forms part of a single entity with different manifestations. It is not clear whether this approach will work for all types of material: some government records people solve the versioning problem by not creating prior copies. This is a bigger issue for internal records management, however there is a policy which is followed within ULCC so the procedures are clear. ULCC has preservation responsibility for the length of the contract or in perpetuity, for all the material stored. In some cases the responsibility is for preservation only and not access. For the SAS repository, the institution has accepted responsibility in perpetuity.

There is a view that a collaborative approach is essential and acts as an enabler for implementation. Loss of control over materials is perceived as a barrier when trying to "sell" preservation services: people are frightened of losing intellectual control. ULCC have been tackling this by demonstrating to people that whatever repository system is in place, control is offered through the provision of a Web form allowing them to allocate permissions and track who has looked at their materials. People are re-assured knowing that they can carry out these tasks. Advising researchers that this is possible forms part of the advocacy role.

ULCC views IRs as in the early stages but can see opportunities for overlap between research council funded data centres and IRs. For SAS, for some of the materials, deposit in the AHDS is a potential option. There is a view that a repository should reflect the research outputs of the institution, however there is no issue about materials being deposited in another data centre but ULCC would prefer not to duplicate effort to describe these items and would rather be able to link to such objects. This is seen as a key requirement: a mechanism that allows all parties to deposit objects in a single place and be confident that the materials will be visible from all other appropriate places. Allied to this is a greater acceptance of the notion of virtual repositories with multiple ways to access materials. This is perceived as a cultural issue: for some people the institution is the focus whilst for others the subject is the focus and for interdisciplinary areas this is becoming increasingly important. The issue is for people to be able to accept that multiple ways of accessing digital content is acceptable.

5.3   Digital repositories

As yet, there are relatively few UK institutional repositories containing research data. A small number of JISC projects working in this area were interviewed, which represent a narrow range of disciplines.

5.3.1   eCrystals Federation / eBank Project

The eBank Project funded by JISC is now in Phase 3 and is scoping the viability of a distributed global federation of repositories for crystallographic data describing molecular structures (see Figure 4). The project is led by UKOLN at the University of Bath with partners at the University of Southampton, the Digital Curation Centre and the University of Manchester. The National Crystallography Service (NCS) is also based at Southampton, providing a facility for data collection, work up of the data and publication of the data to the chemistry community. There is

a commitment to archive the data and make it available at any time in the future. eCrystals is the primary repository and is currently supported by the NCS and an EPSRC grant with direct support from the institution under negotiation. The entire output of the crystallographers at the School of Chemistry is being archived in the IR from 1st January 2007, initially into a private escrow repository but then made available to the public within 3 years (if not before). This is now NCS policy and is published on the Web site.

Figure 4 eCrystals Global Federation Model
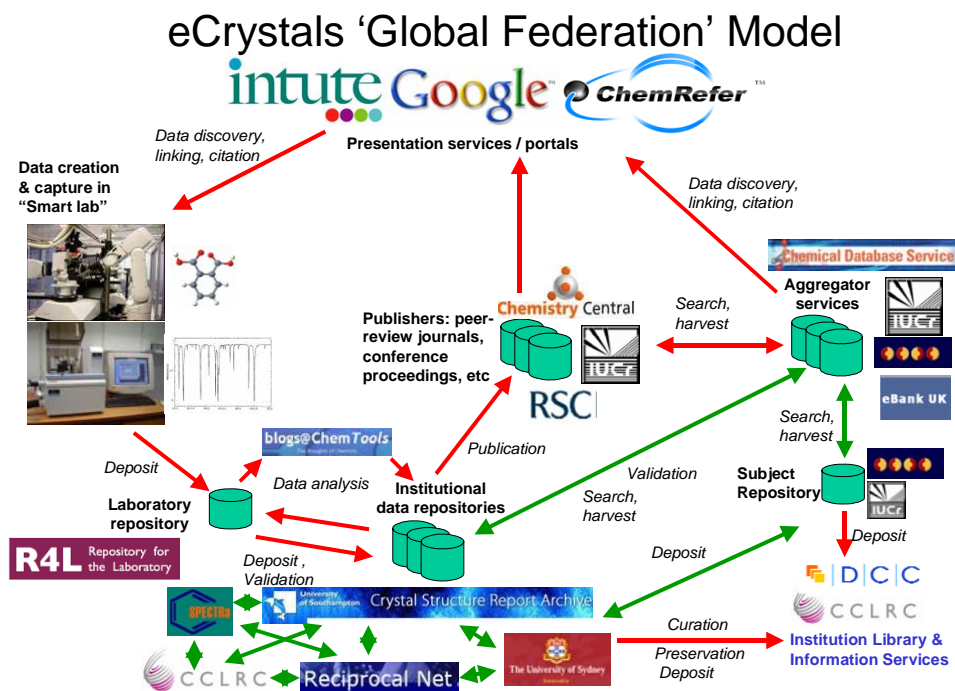
Crystallography data files are in the Crystallographic Information File (CIF) format which is mandatory. Most of the data are processed and may include ASCII text and JPEG images. Deposit occurs at the end of an expert workflow. Associated projects such as R4L are exploring collection of analytical chemistry data within a laboratory repository from a range of heterogeneous techniques. The deposit workflow has been described[47] and captured in the DigiRep Wiki. The CIF is usually submitted as an email attachment to the Cambridge Crystallography Service (CCDC) for validation prior to publication in an article. This is not a condition of grant but a requirement for publication in many crystallography journals. The CIF is deposited with the data centre and then a link made between CCDC and the journal article.

The scenario envisaged with an operational federation of institutional repositories would entail metadata describing the selected records and the CIFs deposited in the eCrystals repository (or any other IR), being harvested or copied by CCDC when the data was publicly available. Whilst this would result in some replication, CCDC do not hold all the data files associated with a crystal structure. In addition, in this model the institution would retain IPR and ownership of the data, and the process would change from a "push" process to a "pull" process with the onus on the harvesting agent such as CCDC. In addition the IR would also contain the journal article or eprint referring to that dataset. There are issues related to the journal policy and for the Amercan Chemical Society, if data or an eprint was published in an IR, this constitutes publication and there is a "prior publication" problem. Within a federated model, the data centre would harvest from a group of distributed institutional repositories on a regular basis.

At present there are "gentleman's agreements" and no formal agreements in place. The CCDC have said they are happy to accept data in this way but this is an informal verbal agreement. They do not yet have automated harvesting mechanisms. These changes are at an embryonic stage and Phase 3 is investigating such issues. At present there is no formal versioning

process. If a record is incorrect or needs updating then removal can be requested with a replacement or a new record is generated and a link made between the two. This process is under review with the aim of making it more formalised.

The eCrystals repository at Southampton uses the eprints.org software which has been modified to support data. Metadata describing the crystal structures is exposed as OAI-PMH for harvesting by the prototype aggregator service[48] developed and located at UKOLN. The aggregator is based on the e-Prints UK service which was developed using the Cheshire software platform. The aggregator service was embedded in a learning portal PSIgate which was part of the Resource Discovery Network/Intute and is used by students to discover digital resources. A pedagogical Evaluation Study[49] has been completed to gain some preliminary evidence about the potential application and benefits of exposing students to primary research data as part of their online learning course.

Each crystal has been assigned a persistent identifier, (a DOI issued by the DOI registry for datasets at the National Library for Science & Technology TIB, University of Hanover, Germany), and a domain identifier: the International Chemical Identifier or InChI. With this functionality, a dataset can be cited using a standard format *http://dx.doi.org/10 .1594/ecrystals.chem.soton.ac.uk/145* (eBank has a provisional data citation policy), and can be discovered using Google.[50] Datasets also have links to derived research publications to demonstrate proof-of-concept added-value linking capability. Underpinning the data repository, there is an eBank data model and metadata schema application profile for describing crystal structures, which have been published on the project Web site[51].

Semantic interoperability within the crystallography domain has also been investigated as part of the eBank work and the issues have been documented in a report[52]. Given the e-research tenet of enabling inter-disciplinary science, there are significant challenges associated with the best use of name authority files, keywords, cataloguing terms and descriptors. Technical challenges being explored within Phase 3 include comparing data outputs from different laboratory workflows, consideration of the metadata schema used to describe crystal data in the distributed repositories, and requirements for metadata normalisation services to facilitate harvesting by aggregator services.

There are some perceived cultural barriers to this new way of working. There is resistance amongst researchers to changing current working practice with views of *"Why should we change"*? There is a perception of this new process adding to daily workloads and questions such as "*What's in it for me*"? being voiced. Automatic deposit processes would help in this area. The CCDC provides a range of additional services such as sub-structure searching, and the development of cross-repository search and other tools would also help to demonstrate added value.

In general, the crystallography community has been fairly receptive to the IR proposals. The current publishing process is both a barrier and an enabler. Authors will do whatever the journal requires in order to publish, but they are worried that publication in an IR will jeopardise their peer-reviewed publications. However visionary publishers such as the International Union of Crystallography and the Royal Society of Chemistry, are happy to fit in with this new model and to alter the way they work or add additional processes to current methods. Data centres are pleased to see positive ways to enhance the capture of data into their databases. One issue in this context is the lack of standardisation of metadata describing the datasets for harvesting. The eBank application profile has a role to play here, however other crystal repositories may not wish to adopt this profile. eBank Phase 3 project work includes a significant advocacy role in talking to publishers and other stakeholders to explore future working partnerships in both technical and non-technical areas.

eBank Phase 3 is also exploring curation and preservation aspects, and in particular is examining different approaches to repository certification. Following eCrystals work, the University of Southampton is now working towards a data preservation policy and has a Working Group chaired by the University Librarian in close partnership with the Head of Information Systems. The Group is currently capturing data preservation requirements from colleagues across all institutional departments in a major review exercise. It is hoped to use the eCrystals repository as an exemplar of good practice.

5.3.2 SPECTRa Project

The JISC-funded SPECTRa Project (Submission, Preservation and Exposure of Chemistry Teaching and Research Data), is led by the University of Cambridge working with Imperial College and is implementing a repository based on the DSpace platform for computational chemistry data. SPECTRa is one of the proposed partners in the eCrystals Federation. Cambridge developed DSpace IR implementations following collaborative connections with MIT, however it is envisaged that there will be multiple IR platforms in future. It is intended to offer the SPECTRa repository to the chemistry department at the end of the project. It is currently acting as a holding repository and some content will be transferred to the IR for long-term retention; data may remain in the departmental repository. Imperial is specifically exploring issues with chemistry theses: if a chemistry thesis has data associated with it, do you put the data and thesis together or store separately with links? DSpace is not designed to make such semantic connections and they have to be assigned manually. Cambridge has given a commitment to long term preservation for the IR and SPECTRa is designed to collect a particular type of data for the foreseeable future ie around 5 years.

It is envisaged that the number of aggregator services will grow in future and there will be a greater need to develop unifying technology to link the contents. The user will use search engines to locate content without worrying about where the material is located, For this reason, departments must archive and present their outputs in a discoverable manner. Dependence on centralised disciplinary repositories may be lessening, with greater reliance on institutions publishing data and associated metadata for discovery. This is linked to a view that data may be presented with a clear licence, which allows its dissemination without permission in a way similar to that envisaged by the Science Commons. However there is a requirement for adequate quality assurance processes to be in place for effective data sharing.

The SPECTRa deposit process currently uses the integrated access management tools within DSpace, though it is thought that more sophisticated access management tools may be required in the future. At Cambridge, most crystal structures are deposited by the Crystallography Service Manager, in contrast to the arrangements at Southampton where there is a more distributed deposit process. The Cambridge approach is a service provider relationship with the chemist, ie the crystallographer hands over the crystal data CIF to the chemist when they are both agreed that a crystal is complete. The crystallographer deposits the CIF file into the archive. The workflow is documented on paper: there are no electronic lab books or processes in contrast to the Southampton arrangement, and most crystal determinations are completed by one individual. This illustrates the variety of laboratory practices associated with crystal data creation and processing, and which must be considered as part of the data curation procedure. SPECTRa has used the eBank application profile as the core profile and further extended it to suit their local requirements for computation and spectroscopy. The quality of metadata are seen as a key issue, in particular to enable discovery across federations of repositories and across disciplines. There are also requirements to facilitate annotation of crystal structure files with computational information.

SPECTRa uses the DSpace handle system to allocate identifiers and also uses InChIs. The project is doing more work in this area to make the resolution process more robust. Naming of crystal structures is also an issue. A controlled vocabulary is perceived as useful for precise information but there is a trend towards free-text indexing which is seen to be just as powerful (and "users hate putting in keywords"). A mixed model of formal and informal terms has ensued, however computational chemists in the project have a different view: a belief that within five years all journal articles will be self-describing. The SPECTRa repository stores mainly processed data since many of the the raw data files e.g. NMR files, are large. There have been concerns voiced around the ability of chemists to get peer-reviewed papers out of re-analysing crystallography data: it is thought that some crystallographers would be uncomfortable with this. Whilst there is a very strong argument for capturing both raw and processed data, which do you expose? The basic procedure might be to capture both and determine the policy for dissemination subsequently. However one must consider regeneration costs versus storage costs, assuming that the crystals are reproducible. It is estimated that 99% of all molecules ever made, do not have physical instances.

Imperial College is developing plans related to its impending de-federation from the University of London. If a student chooses to receive an Imperial College degree, they will be mandated to deposit their PhD in the IR. It is also hoped to develop a mandate for published papers too. Linking from either the thesis or the paper to the underlying data, is being explored but DSpace is currently not well suited to such semantic linking of complex objects.

5.3.3   GRADE Project

The GRADE Project led by EDINA at the University of Edinburgh, is a JISC funded testbed repository for geo-spatial data streaming from a number of pilot sites: including the University of Edinburgh, University of Strathclyde, Kingston Centre for GIS and the University of Southampton. EDINA has a strategic role as a geographic data centre for JISC through Digimap and for ESRC via the UKBorders service. EDINA also has developed a suite of shared services for geospatial data such as the Go Geo! portal and the geoXwalk service.

Geospatial data are not normally deposited at EDINA; geospatial data generated from research is essentially lost within institutions and departments and not available for wider sharing. This was the principle underpinning GRADE and the reason for developing the GRADE repository. GRADE partners have completed departmental audits and have identified over 1000 datasets currently without a long-term home. The question arises, should they have been deposited with NERC or ESRC funded data centres? However a sizeable amount of data doesn't arise from research council funded research and GRADE has been examining this type of data. There will be legacy data from an earlier GIS project at Plymouth University together with some other relevant datasets from researchers who do not have another place to deposit. Some users recognise that they need a place to deposit their data. Others are less willing to share their data and there is a need for advocacy to try to make researchers comfortable about sharing and describing their data at the departmental / research team level, before potentially being willing to publish and share data more widely for re-use. The repository will contain derived and processed data in various GIS file formats e.g. shape files and data will have to be in required formats to be accepted.

Access to the repository will be through a registration process and login, and users must be working within a UK H/FEI. There are no formal agreements in place, however there is a mechanism whereby the user must agree to the Terms & Conditions of the repository but also agree to the Terms & Conditions for deposit ie agree to deposit in certain file formats. Shapefiles are complex packages, and currently zipped files are uploaded by the author. Validation software is being developed to automatically check the metadata. The UK geospatial standard is based on ISO 19115, and compliance with this standard is desirable. The repository is based on the DSpace platform but there are issues associated with the ability to edit metadata fields to facilitate compliance. There is a perception that IRs are not sufficiently "geared up" to handle data and specifically not geospatial data. There is increasingly friction between community requirements for rich metadata and JISC requirements for a core metadata set based on maximising interoperability. The minimum set within the ISO standard can be mapped to Dublin Core but DSpace does not permit editing of metadata elements. This raises challenges about how best to expose simple discovery metadata but also expose richer community metadata into more sophisticated community search engines? There are also issues with the way DSpace handles versioning of datasets. This is being investigated in a "sister" project at the University of North Carolina.

A survey of data-sharing within the geospatial community shows that it is currently achieved through informal means such as email. There is some use of data centres but evidence suggested that not all data was deposited as often as it should be as a requirement of funding. The survey revealed that users used internet-based search tools rather than data centre interfaces to discover datasets. There appeared to be little duplication of data. There is evidence that in the short term there is a need for a UK repository, but the medium to long term strategy is to encourage institutions to manage the data themselves, however this may be in conflict with a community view. There is sharing within a department ie on a disciplinary peer group basis, but not much sharing within an institution. However, the challenge for the discipline is that many users may be in other sectors e.g. government. GRADE proposes that users go through a

checklist set of questions to determine where is the best place for deposit. Duplication is not seen as helpful as it may be difficult to determine which is the authoritative version.

Geospatial data are associated with a challenging legal environment. Ordnance Survey data and satellite data has complex licensing requirements which act as a serious barrier to data sharing. The Guardian newspaper is running a "Free our Data" Campaign[53] launched on 9th March 2006 and designed to raise awareness and gather support for more open access to geospatial data. There are also issues associated with reuse of data by 3rd parties, that may be licensed under different terms and conditions.

There are perceived barriers to data sharing within the geospatial community such as legal issues and also a lack of awareness of resource discovery facilities. Data sharing is often carried out informally by email or burning a CD to send to a colleague and these approaches do not scale. There are also questions around how institutions will respond to the European Inspire Directive[54] which is seeking to require public bodies to make their spatial information services understandable and accessible across government and across national boundaries ie join up. However, making geographic data freely available would destroy the business model of agencies such as the Ordnance Survey. There need to be incentives to promote the benefits of sharing.

The GIS community has developed on the basis of it being difficult to share data because of publishing restrictions relating to rights. There is a view that more advocacy and awareness is needed, perhaps carried out by data centres, to encourage researchers and stress the importance of good information and data management practice starting at the point of data collection. Institutions are perceived to be unaware of the issues surrounding good management of geospatial research data. There is also a view that there needs to be more interaction between research council funded data centres and JISC data services to facilitate sharing of best practice.

5.3.4   StORe Project

The StORe Project[55] (Source-to-Output Repositories) is a JISC-funded activity led by the University of Edinburgh with six other University partners, the UK Data Archive and UKOLN (in a consultancy role). StORe is investigating a range of curation and preservation issues relating to research data and in particular has carried out a substantive survey of attitudes and aspirations of research staff towards data repositories. StORe is also developing pilot software to provide bi-directional linking functionality between data sets (source) and the derived publications (output). The survey, which has included seven disciplinary reports, has revealed some interesting results.

The survey suggests that in general there is low awareness of data repositories, and whilst there is some support for disciplinary data repositories, there is little interest in institutional data repositories. The survey highlighted diverse disciplinary cultural and behavioural approaches to data handling and management. There were also marked intra-disciplinary differences e.g. in chemistry, where computational chemists and chemical informatics staff are keen to pursue data repository options, in contrast to others who do not see the relevance of repositories. Within archaeology, only the science-based researchers saw the relevance of repositories. They did not associate data repositories with support from library staff, but at the same time some acknowledged that they needed help with metadata creation, to supplement their disciplinary expertise. There was a working culture of self-reliance. Some researchers were generally confident in their abilities, whilst librarians saw them as "relatively unsophisticated". Contact with librarians and information professionals was "rare".

Academics were concerned about security issues relating to predatory access and exploitation of their data. They lacked confidence in the application of metadata standards and were sceptical about the value of providing access to data, in particular they had concerns about appropriate interpretation of the data, especially in the biomedical field. There were many concerns voiced about the quality of organisation of institutional repositories, which were regarded as bureaucratic and not very robust in terms of IPR management. There was a perception that "people feel exposed by them". Some researchers viewed data centres as more professionally organised and authoritative. Views about publisher services such as Science

Direct, were also more positive. Astronomy and the biosciences were viewed as being advanced in their data curation practice, and were considered to be good exemplars for more detailed case studies.

5.3.5   University of Edinburgh

The University of Edinburgh is developing the Edinburgh Research Archive (ERA) based on the DSpace platform, within the e-resources team in Information Services (IR). The repository service has the focus on "service" rather than on purely technical issues: they are looking at the issues of end-users and developing solutions to meet needs. Work has concentrated on advocacy areas such as marketing, liaison work and face-to-face sessions with staff, presentations at research committees, running seminars and producing promotional leaflets. ERA is a repository for research outputs. Currently this comprises e-theses and eprints, however the definition has deliberately been left loose. The aim is to move from a project basis to full integration within core services. Whilst the initial primary purpose of ERA has been for access and dissemination, the team are now starting to explore its role in long-term preservation.

There is a mediated deposit service and over 1050 objects have been deposited with more to be processed. Most material is born-digital and IS has re-allocated resources to work on ERA. Some PhD theses may contain data in appendices in flat spreadsheet files, which are normally converted from Microsoft Excel to PDF. However, not all research data would be included in a PhD. When the thesis is deposited with the Library, the data are not described in detail. There was a general lack of information associated with research data: where it is deposited, how it is managed and it was expected to be discipline-dependant. Some subject librarians were thought to have a view e.g. astronomy liaison librarian.

IRs were considered to be in their infancy and self-archiving was not perceived to have "gone main-stream" despite best efforts of individuals. There are early adopters but much inertia in the current system and the very conservative nature of academics, makes progress slow. It was not expected to change swiftly and this was thought to be a limitation on the adoption of new technology. The main driver for IRs was viewed as coming from the library community, however in considering relationships, academics who bring funding into the University are perceived to have more influence than the University Library, which is seen as a service-based unit. Librarians have embraced scholarly communications issues and are trying to get academics to change their behaviours, but the power balance is viewed as a barrier preventing change being pushed through. Advocacy is seen as the first step in effecting change; the next step is a formal mandate from a higher university authority and / or research councils. Relationships between IR staff and academics need to be enhanced to progress IR developments. There was a perceived gap in knowledge of what kinds of data were being produced across the spectrum of disciplines. This information would then inform researcher requirements for dissemination and data management. The area was considered to be emergent and required a detailed "roadmap".

5.4   Other key stakeholders

Some additional organisations are included as key stakeholders.

5.4.1   Learned society publisher: International Union of Crystallography (IUCr)

The IUCr produces a number of journals at its headquarters in Chester, which are published by Blackwell Munksgaard. The print editions of the journals are distributed by Blackwell Publishing; electronic editions of all IUCr journals are available via Crystallography Journals Online at http://journals.iucr.org/. The IUCr has published journals since 1948 and currently has eight titles: *Acta Crystallographica Sections A – F*, the *Journal of Applied Crystallography* and the *Journal of Synchrotron Radiation.* Data curation has been important from the outset, since crystal structure analysis publications contain detailed discussion of the results of well-defined experiments to investigate particular properties or types of matter involving experimental techniques such as X-ray diffraction. *Acta Crystallographica Section C Crystal Structure Communications* and *Section E Structure Reports Online* contain almost exclusively such

crystal structure reports, and this information forms a routine component of longer research articles giving further descriptions of the properties. These reports are data-rich containing structural diagrams and structural data including 3-d positional co-ordinates, atomic motions, molecular geometry and chemical bonding data (see Figure 5 below).

IUCr has enforced a policy of requiring authors to deposit structure factors with the journal as supplementary documents supporting the scientific arguments presented in an article. These structure factors represent a reduced data set, which capture the key aspects of the raw experimental data whose volume is too great to publish, and which would allow repeat analysis at a later date to enable scientists to recalculate a structural model. These primary data sets are freely available on the Web and may also be referenced by non-IUCr journals. The final derived data contained in a standard CIF (Crystallographic Information File) consisting of a six-dimensional structural model, constitutes the description of a crystal structure in the literature. IUCr requires that this structured data are deposited with the article.

Figure 5 Acta Cryst. Brian McMahon, IUCr
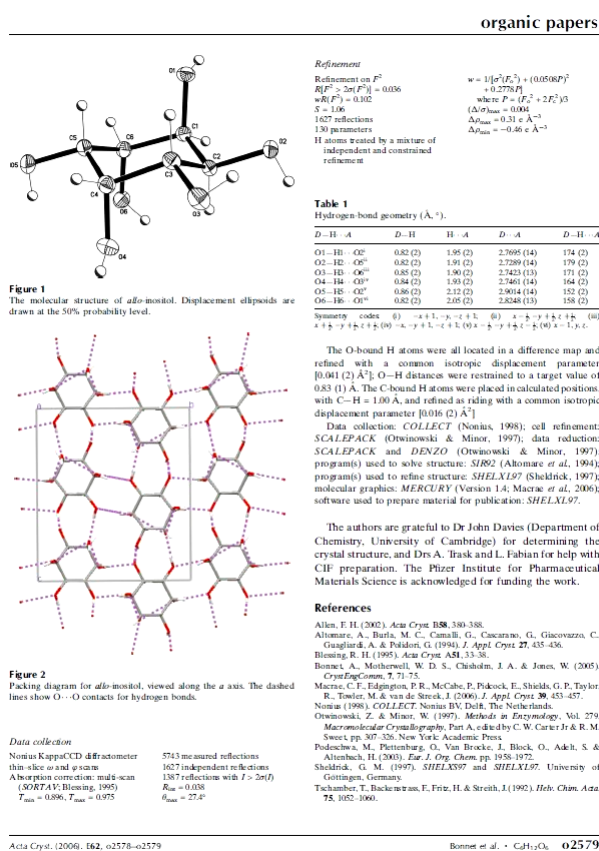
The importance of data to the publication process is illustrated by the spectrum of data publication potentially contained in IUCr journals (Source Brian McMahon, IUCr):

- Radical new knowledge of crystal/molecular structures (rare) i.e. Nobel-prizewinning contributions (more likely to be published in *Nature* or *Science*). Such work would be of major importance, closely peer-reviewed and the data subjected to the highest quality control standards.

- Radical new methods for solving / refining structures (rare) i.e. major innovations in methodology. Initial publications might have relatively low-quality data for novel structures since the technique has been newly-invented to solve a difficult problem, or high-quality data of existing reference structures.

- Major contribution to the understanding of chemical properties of one or a class of compounds (*Acta Cryst. Section B*) or significant new information about a compound or group of related compounds *(Acta Cryst. Section C)*. Whilst this substantial chemical / physical information could be published in chemistry or physics journals, it is likely that the crystal structures underlying the discussion would be important enough to be published individually in a crystallography journal such as *Sections C or E*. The full description of a new structure would be published in *Section E.*

All of the above data, which are all freely available on the Web, will also be curated by synoptic databases such as the Cambridge Structural Database at the Cambridge Crystallographic Data Centre (CCDC).

- A routine structure determination or a partial or incomplete structure determination, may be published in *Section E*, but more likely will not be written up and will remain in a lab notebook, and the data will never enter the public domain. An institutional repository is an excellent location to store such data and the eCrystals Federation/eBank Phase 3 model, represents a good solution to this data curation challenge.

Some other chemistry journals will accept supplementary / supporting information including structural data sets, and make them available on the Web, but others do not. The data may be in compliant CIF format or may not. Databases such as CCDC, may harvest data and metadata from these journals or may not, given the effort required to ingest non-compliant data. There is some voluntary deposit in such databases, but it is patchy, and there is a real risk of total loss of valuable research data.

Figure 6 checkCIF screenshot, Brian McMahon, IUCr.

The IUCr is particularly supportive of data publishing initiatives such as eBank / eCrystals, which use standard data exchange protocols (CIF, OAI-PMH, DOI etc.) and are adopting a federated approach to foster resilience, interoperability and common information management practice. The use of common quality assurance procedures is also important: checkCIF is becoming the defacto standard within the discipline, although it was originally constructed for use with the journal (see Figure 6 above). The value of digital repositories in this context, is for the publication of all the supporting data collected during an experiment, including the raw data images archived as part of the National Crystallography Service. IUCr plans to provide links between subsequently published articles and the corresponding eBank/eCrystals data records. It is acknowledged that rights to the ownership and dissemination of the data need careful consideration and handling. The application of checkCIF software provides a quality control mechanism for structures which have not (yet) been peer-reviewed. Similar linking principles apply to biological macromolecules in the Protein Data Bank (PDB), but community practice is different. Where relevant to crystallography, these structures are published in *Acta Cryst. Section F.* It is desirable that linking between data and articles should be based on the same quality assurance principles, and there are some technical issues relating to the comparative functionality and application of different identifier solutions in this context.

5.4.2 Digital Curation Centre

The Digital Curation Centre (DCC) was set up in 2004 with funding from JISC and EPSRC. The DCC aims to provide a range of services to promote and facilitate good practice in the stewardship and curation of research data, both now and over the longer term. It is currently in Phase 2 and the work programme covers creation of resources, service delivery, provision of expert advice, technical development and original research.

A number of cultural and social factors were cited as barriers to data sharing in a collaborative manner. These include perceived competition for intellectual assets between IRs and subject repositories. There was a belief that IRs and subject repositories should be working together since subject repositories give critical mass and have an important community proxy role. Data deposit is not routinely embedded in research workflow, and a major cultural change needs to occur to address this within the research community. It is often viewed as an "extra piece of work" and positive drivers such as data citation, where citation is a primary scholarly indicator of value, to enable (visible) re-use, are required. There was a view that the key drivers must change, but that this will be a slow process and will take many years.

IRs tend to be managed by "generalists" located in institutional libraries and information services, who do not usually have data skills, whereas subject repositories are staffed by domain experts with well-developed data handling skills. Data sets tend to be unique without multiple copies, are not self-describing and may require much interpretation, particular competencies and domain knowledge. Representation information, which may include software programs, is required to assist with interpretation. Closer partnerships between staff working with IRs and those working with subject repositories, is essential to help to build engagement with the designated community.

IRs have a fundamental argument for sustainability because of their position within an institution. They are frequently located in departments with a persistent service orientation i.e. libraries or information services, whose main role is the management of information. However, many (but not all) scholars, care primarily about their discipline or domain, and secondarily about their institution. This potential barrier needs to be addressed in order to make IRs sustainable as services. Scholars who have hybrid skills in domain science and data/information handling, will be a valuable resource in the future.

5.4.3 Research Information Network

The Research Information Network (RIN) was set up in 2005 "to lead and co-ordinate the provision of research information in the UK". Since that time it has commissioned various studies, organised events and seminars and acted as a facilitator to bring various research groups together. RIN is increasingly concerned with data stewardship issues and has published a study of Research Funder Policies and a set of draft Principles for the Stewardship of Digital

Data. The five high-level principles embrace Roles and Responsibilities, Standards and Quality Assurance, Access, Usage and Credit, Benefits and Cost-effectiveness, and Preservation and Sustainability. Both of these documents have been referenced earlier in this report.

Towards the end of this current piece of work, RIN published a report of researchers' use of academic libraries and their services[56], co-funded by CURL. This was a major survey carried out by Key Perspectives, and highlighted the urgent need for librarians and the research community to work together to clarify the roles and responsibilities of key players in managing data outputs, at national and institutional level. The Report showed that researchers' awareness of new developments in scholarly communications is low, particularly issues concerning open access to research outputs. The RIN has also set up a scholarly communications group which includes publishers, funders, librarians and data centre representatives, and which will develop a collaborative agenda of projects and information.

Future planned studies include work on research costs and income flows and on the publication and quality assurance of research data outputs.

6 Synthesis and Discussion

This section of the study presents a synthesis based on the collated findings from the interviews, the workshop presentations and subsequent discussion and analysis. It is arranged in nine sections: Co-ordination and Strategy, Policy and Planning, Practice, Technical Integration and Interoperability, Sustainability, Legal and Ethical Issues, Advocacy, Training and Skills, Roles, Rights, Responsibilities and Relationships.

6.1 Strategy and Co-ordination

The eScience Curation Report published in 2003 and quoted in Section 4, described a patchy picture with regard to data curation, with a lack of government-level strategy and co-ordination. The recent OSI e-Infrastructure Working Group Report highlighted the need for a national infrastructure which presupposes *"not only a high level of integration and co-ordination, but also, in key areas, intervention at the policy level".* Based on the relatively small sample investigated in this study, the landscape in 2007 has not changed substantially over the last four years. There is still a lack of co-ordination in the development of government-level strategy amongst UK research funding organisations and other bodies, with regard to curating and preserving the large volumes of data generated by the funded research programmes. The requirement for a more "joined-up" approach to strategy development can be viewed at different levels and more co-ordination is needed at **all** levels:

- Across sectors e.g. arts and humanities research data outputs and cultural heritage data
- Between funding organisations
- Within funding organisations
- Between institutions e.g. universities, research council institutes, national libraries, TNA
- Within institutions
- Between different data centres and data services
- Across related disciplines
- Within disciplines.

The data landscape is complex and heterogeneous. There is great variation in data curation practice across disciplines and strategy based on a mixed model approach is likely to be most effective; a "one size fits all" solution will not work in this complex data environment.

The development of a co-ordinated UK strategy for data curation and long term preservation will have a number of benefits. Such a strategy will:

- Inform current and future planning for funded research programmes

- Inform current and future planning for support infrastructure for data curation and preservation

- Develop the research workforce and ensure capacity-building for the future

- Move towards assuring the survival of our scientific heritage in the longer term

- Maximise the return on investment by government agencies in research activities

- Make the UK more competitive in the global economy resulting from research innovation.

The recent (March 2007) signing of an agreement between the JISC and the Research Councils, is a positive step forward which will support future more co-ordinated strategy development. In a global context, such a strategy would complement data initiatives emerging in the US and elsewhere, and facilitate partnership and collaboration with other key research funding organisations and federal governments.

In the UK, there are examples of collaboration and good practice. NERC is moving forwards with planning in their domain in partnership with AHRC in certain areas; the MRC and Wellcome Trust are addressing data curation at a strategic level; MRC have talked to AHRC, BBSRC and NERC about infrastructure developments and planned data support services. For NERC, a data and preservation strategy will form an element of a broader knowledge strategy, which in turn will inform a long-term (5-year) scientific strategy. However, strategy development to build a common infrastructure needs to happen across **all** funding organisations as a joint activity, which is fully co-ordinated and "joined-up". Such a strategy also needs the visible high-level support of appropriate senior representatives within the funding organisations.

As a preliminary piece of work, there is an urgent need to gather detailed information about the current state-of-play regarding data holdings, most usefully in the form of a gap analysis of current data collections. One element of a future strategy would then be to address identified gaps in current data holdings. A further mapping might identify gaps in data curation infrastructure and support services across disciplines, since in order to deliver a data curation and preservation strategy, there is a requirement for an adequate support infrastructure.

The Strategic e-Content Alliance (SEA) is a cross-sectoral group with representatives from various relevant sectors: there may be a role for SEA in facilitating or commissioning at least a part of this work. The Research Information Network is also active in this area and has commissioned a number of relevant studies. However the fact remains that there is much more to be achieved and a number of recommendations are proposed:

REC 1. JISC should commission a disciplinary DataSets Mapping and Gap Analysis, with associated curation and preservation support infrastructure.

REC 2. Research funding organisations should jointly develop a co-ordinated Data Curation and Preservation Strategy to address critical data issues over the longer term.

REC 3. The Strategic e-Content Alliance should consider adopting a facilitation role in promoting a cross-sectoral strategy for data curation and preservation.

The degree of strategic co-ordination within individual funding organisations, and the nature of partnership links between funders and their funded data support services, appears to be variable. Both ESRC and NERC have links with their funded data centres through named staff and through formal committee reporting lines. The NERC model of a Data Management Co-ordinator with a dedicated role, appears to work well. However in general, there is scope for stronger, more formal and more pro-active links between funders and their funded data support services. There is also scope for more integrated and enhanced planning within funding organisations between Research Programme Planning Committees, who make funding decisions on research projects and programmes, and with infrastructure support staff who have to provide operational services to curate and preserve the outputs of the funded research. Whilst this planning is more evident where very large programmes are proposed, for responsive mode grants, procedures appear to be more ad hoc. More formal connections will facilitate better forward planning for best use of resources, effort, capacity and financial support, in order to maximise return on investment, and will support more explicit audit trails for effective use of

public funds for research. The Wellcome Trust model, where the Policy and Advocacy Team sits within the Strategy Planning & Policy Unit in the centre of the organisation, ensures that good links between strategy and operations are maintained and that policies are embedded in practice and are monitored, evaluated and disseminated effectively.

Within higher education institutions, and in particular at research-led institutions, there appears to be very mixed practice in handling the data outputs of research activity. The StORe survey suggests that management of research data within academic institutions, is at best, rather ad hoc. It seems likely at this time, that most HEIs will not have any formal strategy in place for curating and preserving the data outputs from their research activities. Whilst many universities are now implementing institutional repositories for storing and disseminating the textual interpretations of the results of research ie eprints, there is currently no equivalent drive to manage primary data in a co-ordinated manner. Indeed, there are a number of major questions for institutions which need to be urgently addressed around roles and responsibility, future strategy, implementing good practice, developing staff skills and resource provision. We will return to the question of institutional responsibility in Section 6.9.

However there is also an immediate need for acceptance of the importance of these data issues to be addressed across institutions, and various bodies have a role in this context at different levels. Universities UK are well-placed to facilitate engagement and discussion at the highest levels: the stewardship of locally-created research data which represents the intellectual property of the institution and its employees, is of critical relevance to senior managers with research roles. SCONUL, CURL and UCISA are also well-positioned to engage and influence University Librarians and IT Directors and to facilitate collaborative strategy development. The CURL eResearch Taskforce is already working to raise awareness in this area. The DCC is also addressing related issues and with the DPE project, has recently published the DRAMBORA Digital Repository Audit Method Based on Risk Assessment[57], which may provide some pointers in approach. However, once again the full picture across institutions is very sketchy: what happens to research data created within UK institutions? What data are being effectively stored within institutions? JISC is very well-placed to facilitate an information gathering exercise within institutions to map what types of data are present, how they are managed and where they are deposited for long-term preservation. The University of Southampton has already initiated a review of data issues following eBank Project activity, and Southampton could act as a useful case study.

REC 4. JISC should develop a Data Audit Framework to enable all Universities and colleges to carry out an audit of departmental data collections, awareness, policies and practice for data curation and preservation.

Finally in this section, there is clearly much good practice, expertise and knowledge located within data centres and data services. However, perhaps because of the dual funding lines or because of cultural and community differences, there is little exchange of experience between data centre staff. This needs to be addressed through both formal and informal means, with regular face-to-face meetings and ongoing electronic dialogue. The Digital Curation Centre is well-positioned to bring these communities together, and this is a planned task within the Phase 2 Community Development Work Programme.

REC 5. The DCC should create a Data Networking Forum where directors/managers and staff from research council data centres, JISC data services and other bodies, can exchange experience and best practice.

6.2 Policy and Planning

There is a wide spectrum of policy implementation relating to research data across the research councils interviewed, ranging from no specific data policy statement, to publicly available, well-established and relatively mature policies, which are being reviewed in the light of changing political, economic, legal and technical drivers. The "arms-length" position of EPSRC is in strong contrast to those of the other research councils interviewed. Published policies include MRC, Wellcome (new 2007), ESRC (under review), NERC (under review and to be published in its new form in October 2007). BBSRC who were not interviewed, have published their data-

sharing policy in April. There was very strong support at the consultation workshop, for all research councils to publish a data management policy.

It seems beneficial that research councils should pro-actively liaise with their peers in policy development. It was noted by one respondent that there is a challenge to achieve a balance between prescriptiveness and enablingness between organisations, and the degree of granularity and level of principles for operation is challenging. However, funding organisations should make their policies clear and consistent and make them accessible and appropriate to their intended audience. These policies should encompass data management and curation, data publication and sharing, and data preservation. They should take into account the changing landscape of data centres and data archives, institutional and subject repositories, social networking software such as wikis and blogs, and other Web data publication opportunities.

Once implemented, policies should be "enforced" where possible and this aspect is explored further here and in Section 6.3. Data deposit compliance requirements for researchers are currently very weak. Whilst the traditional forms of publication such as scientific peer-reviewed papers may contain results data, in certain disciplines such as omics and crystallography, it is a requirement of publication that data are submitted to an appropriate data centre prior to publication. The growth of data-rich publications, whether they are in journal form such as *Acta Crystallographica* or *Molecular Systems Biology,* or as a database such as the *Nature Signalling Gateway*,[58] is likely to increase. Datasets are a valid and peer-reviewed research output in some disciplines, and future research assessment exercises will need to devise mechanisms for the incorporation of data outputs into the process, which in turn will encourage data deposit. The requirement to submit data for validation prior to publication, acts as a helpful incentive for researchers to deposit their data in a managed archive, whether it is in a research council data centre or in an institutional repository or in a subject repository (or whether it is in more than one of these locations). The provision of collaborative tools and workspace may also be an incentive for data deposit.

A further very interesting development is the potential to share data using social software and this is already beginning to occur in certain communities such as chemistry[59]. In parallel, more open approaches to the peer review of scholarly papers are being explored by some journals[60] and the various mechanisms for assuring quality may also apply to published datasets. Work is needed to scope current practice in all these aspects, assess the potential and evaluate associated data curation and preservation issues. Of course there are also cultural aspects to consider, and both funding organisations and institutions need to reflect on such trends when reviewing and formulating future policy for their communities.

REC 6. Each research funding organisation should openly publish, implement and enforce, a Data Management, Preservation and Sharing Policy.

REC 7. All relevant stakeholders should identify and promote incentives to encourage the routine deposit of research data by researchers in an appropriate open access data repository.

REC 8. JISC should commission a scoping study to investigate current practice, assess future potential and evaluate the curation and preservation issues associated with sharing research data through social software forums.

Funder data policies should contain a set of core principles providing guidance to researchers on requirements for data management and curation, data publication and sharing, and long-term preservation. The new Wellcome Trust Policy is an example of a simple and clear document with helpful supporting Q&A. It makes a requirement for applicants in specific cases to provide a Data Management and Sharing Plan and states that "*these …. plans will be reviewed as an integral part of the funding decision"*. Guidance is given on the contents, which should include reference to data quality and standards, use of public data repositories, intellectual property, protection of research participants and long-term preservation and sustainability. The guidance on use of public data repositories states "*the Trust will expect researchers to deposit their data into recognised public data repositories where possible and a number already exist for many types of fundamental biological data.*" During this study, there

was polarisation of views as to what constitutes an appropriate place for data deposit and we will return to this point later. Some details of the assessment process by Trust reviewers are given; the diverse nature of research data in this domain is noted and each plan will be considered on a case-by-case basis. Funding will be dependant on acceptance of an appropriate plan. MRC similarly advocates use of Data Management Plans.

NERC has a requirement for its funded data centres such as BADC, to work with researchers to produce a Data Management Plan, in the case of Thematic Programmes in Directed Mode. A typical BADC Plan will contain information on the types of data expected to be created, a list of the appropriate data centres for deposit, information on file formats, metadata standards, file naming conventions, the data submission process, data ingest and checking procedures, some detail on the archive structure and a data submission schedule. The Plan will also typically include information relating to access to the data as a Data Protocol sharing agreement. The plan should include full cost information, Web portal developments and data retention/embargo periods. The joint production of such a plan is a valuable exercise both in terms of the curated data outputs, but also in terms of the awareness-raising, and learning processes that the researcher must follow, in order to develop the Plan. Whilst there will need to be local review and formulation on the exact contents, the overarching principles and core content are sound, and could act as a basis for a template or pro-forma for wider implementation by all funders. In addition, project final reports should include structured results describing what datasets have been deposited, what standards and procedures have been followed and what goals have been achieved. A further incentive to data deposit compliance, may be to retain final grant payments until formal acceptance of the final report, with the required level of detail of data curation, management and preservation practice. For example, published/curated datasets could be identified by the use of identifiers/URLs in a standard data citation format. However, an incentive-based approach may be more effective than a punitive position in the longer term.

REC 9. Each funded research project, should submit a structured Data Management Plan for peer-review as an integral part of the application for funding.

Institutions need to have a data management, curation, sharing and preservation policy. The Southampton approach of forming a Working Group to address these issues is a good first step. Such a Group should include representatives from all relevant areas of the institution including the appropriate senior manager, academic staff, researchers and students, the Library, IT services and legal expertise. The institutional policy should recommend that researchers deposit their data in an appropriate public data repository, (which may not necessarily be only in an institutional repository), where these exist. This policy-setting activity should follow initiation of the institutional data audit described in Recommendation 4. Responsibility for co-ordinating and implementing policy development will depend on local practice, but should reside with a member of the institutional Executive or senior management team.

REC 10. Each higher education institution should implement an institutional Data Management, Preservation and Sharing Policy, which recommends data deposit in an appropriate open access data repository and/or data centre where these exist.

6.3 Practice

In previous presentations about data curation practice, it has been helpful to refer to the various stages of the research cycle and one interpretation of the cycle is given below in Figure 7.

(e)-Research Life Cycle view of Data Curation?

(New) knowledge extraction: data mining, modelling, analysis, synthesis

Data processing

Formulate hypothesis / ideas, test, experiment, observe: data creation, collection & capture

Data processing

Adding value: Data linking, annotation, visualisation, simulation

Data processing

e-Infrastructure

Open access

Collaboration

Data management storage & validation: description, deposit, self-archiving, preservation, certification

Data processing

Data processing

Scholarly communications: data disclosure, publication, citation, discovery, re-use
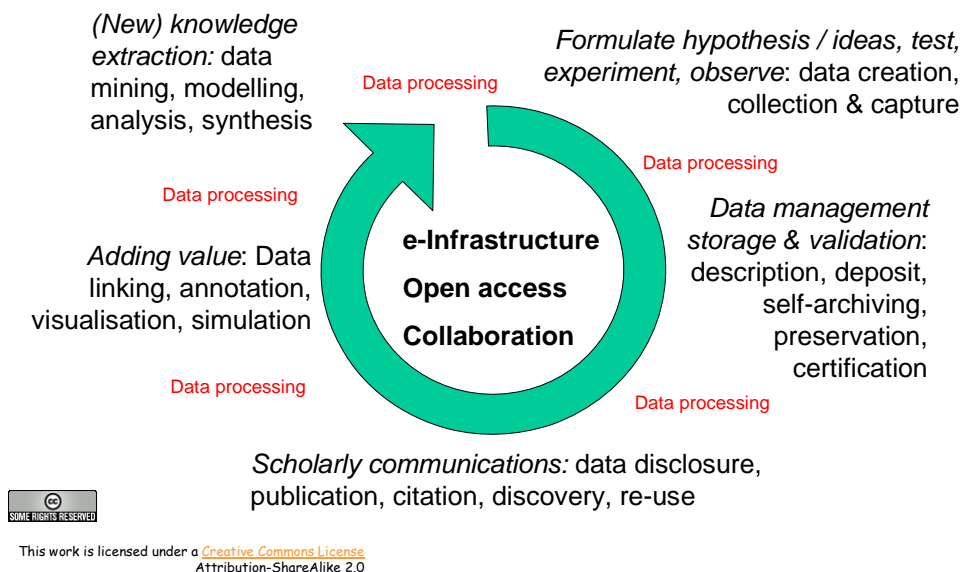
Figure 7 e-Research Life Cycle and data curation, Liz Lyon.

How much do we know about individual researchers' practice regarding data creation and capture, data management, curation, preservation, data publication and re-use? The answer is not enough, and recent surveys (StORe Project, CURL survey, RIN studies), strongly suggest that (with some specific exceptions), most communities of researchers are at best ill-informed of good practice, unaware of the issues and do not seek advice and guidance. The DCC has organised a range of Information Days, workshops and other professional events to engage with the research community (and is continuing to so in Phase 2 work), however, there is much work still to be done. The DCC SCARP Project will engage with particular disciplines to learn more about data curation practice "at the coalface". The EPSRC-funded Knowledge Information Management Project[61] has been working with the engineering community to advise on best practice for data curation and long term preservation.

The interviews illustrated varied practice across different domains, and to inform future practice, it will be useful to have a number of in-depth case studies of good practice and of areas where there are real and perceived challenges. These case studies are best carried out in partnership with data centres. The "omics" area, perhaps focussing on ArrayExpress, is an obvious candidate. The management of clinical trials data would also be an interesting area to study. The new Diamond facility could provide an excellent longitudinal study, starting now with the first scientists on the beam lines, and continuing for several years as data volumes accumulate. Population based longitudinal cohort studies of epidemiological data such as BioBank, would be valuable, since these raise a number of critical issues around consent, confidentiality, use of human subjects and IPR.

REC 11. JISC should extend the work of the DCC SCARP Project, to increase the range of in-depth Disciplinary Data Case Studies on data management, curation, sharing and preservation.

Data are increasingly being captured automatically at source in the field, and by instruments in the laboratory as part of integrated laboratory information management systems (LIMS). However there are issues around the use of proprietary software and varied metadata schema. Whilst in principle, this is clearly a desirable way forward, in many cases the lack of standards and lack of co-ordination of manufacturers, (for obvious competitive reasons), makes subsequent data integration and interoperation more challenging. EBI provided a nice example

of co-ordinated working, where microarray data are collected to an agreed community standard via data pipelines from the array instrumentation. Remote telescopes are another example where huge volumes of data are captured automatically at source to defined standards. In addition, the metadata describing datasets may be manually or automatically enhanced with additional descriptors, and semantic tags, at a later stage in the workflow.

REC 12. JISC should commission a scoping study to evaluate the processes and issues around data and metadata capture at source from a range of instrumentation and laboratory equipment, as part of the end-to-end research workflow.

The Repositories Research Team at UKOLN has worked with Eduserv to develop a metadata schema application profile for scholarly works. Further work is needed to assess whether such a generic model can be practically applied to data sets, which are frequently of a highly complex nature, stored as dynamic databases with many rows and columns, and constantly changing data values. Application profiles may be more appropriately developed for sub-disciplines such as the published eBank application profile, which describes molecular structures in crystallography.

REC 13. JISC should commission a study of the applicability of generic data models and metadata schema application profiles for scientific data.

The BADC has a set of data selection criteria based on the concepts of "usability and usefulness". Other data centres also have such criteria, which may be publicly available on the Web. It would be beneficial for these criteria to be more widely shared, with a view to testing their applicability in other domains and developing some common policy and practice. This sort of information is also of value to institutional repository managers and the Data Networking Forum might be an appropriate arena for this exchange.

A number of points were made around the sustainability and scalability of deposit mechanisms for researchers. These included a perceived need to investigate mechanisms for the batch deposit of multiple and different datasets into the same repository, and also for the deposit of a single dataset into one or more different repositories or data centres. In this latter context, the effectiveness of single deposit and multiple linking should be considered. There were issues around deposit of particular types of data content such as theses, where the main item is a textual document, but which might be supported by significant quantities of data. There has also been work to develop a repository deposit API for scholarly works (eprints) and this is being progressed in the new SWORD project. Wider co-ordination is being pursued through the work of the JISC Common Repository Interfaces Working Group (CRIG). Can this approach be extended to data sets? A number of co-operative high-level workflow models were described during the interviews (e.g. AHDS and University of Belfast), based around partnerships between the researcher/institution and the data centre or service. More work is needed to explore which models work most effectively, embracing both the institution and the data centre in a collaborative arrangement.

REC 14. JISC working through the Cross Repository Interfaces Working Group (CRIG) and domain partners, should investigate the feasibility of repository deposit APIs for disciplinary data.

REC 15. There is a need to identify and promote scaleable and sustainable operational models for data deposit, which are based on co-operative partnerships with researchers and common standards.

What do we mean by data publication? This study has observed various approaches to data publication and citation. Peter Buneman of the DCC has a view from the database perspective; the CLADDIER Project has published a Briefing Paper[62], others will consider publication of data within more traditional journals. In each case, the data set must be uniquely identifiable in a data citation. More work is needed both on understanding the diversity of data publication, but also in examining the viability of a standard data citation format. The Research Information Network, CURL and JISC have published a call for Expressions of Interest for an investigation in this broad area. The JISC Scholarly Communications Group could also usefully begin to address these issues.

REC 16. The JISC Scholarly Communications Group should collaborate with the RIN Communications Group, to review data publishing principles and good practice.

The lack of willingness to share data in certain communities of scientists was raised many times: for social scientists it is accepted practice, whereas for population scientists, it is rare. The Wellcome Trust is exploring a range of approaches and incentives to facilitate better data-sharing. These include providing flexible levels of access dependant on the type of user, implementing a "club"/closed group model, and use of embargo periods. It is clear that any scheme needs to have in-built flexibility and be a constituent part of a Data Management Plan.

REC 17. JISC should commission work to investigate the effectiveness and applicability of different mechanisms for managing access and data-sharing across disciplines.

The debate is continuing around the appropriateness of institutional repositories for long-term preservation of both textual materials and data. The working relationship between the BADC and the CLADDIER Project has led to shared learning and mutual benefits, helping all parties to understand the evolving relationship between institutional repositories and data centres. The PRESERV and SHERPA DP2 projects are also exploring this relationship. In the latter the AHDS will be testing a model of using a FEDORA repository as an access layer, and the iRODS model providing a preservation layer. There also needs to be more work examining the impact of assigning data preservation responsibility to 3rd party services.

REC 18. More work is needed to identify integrated information architectures, which link institutional repository and data centre software platforms.

One further area of uncertainty, in terms of having adequate evidence to inform service development, is around the degree of re-use and re-purposing of data sets. How much data are used again in recombination or re-analysis? What features of a dataset and its associated metadata facilitate re-use? Simple discovery of its existence? A comprehensive set of metadata describing its properties? Some sort of quality rating or peer-review imprimatur? There is a perception that at present, most data are not re-purposed (with notable exceptions such as astronomical survey data which is routinely mined to extract further information). Clearly a researcher should acknowledge the source of their data (if they are not the creator), provide provenance information to assure its quality, and if appropriate, add annotations and tags to add value to the data holdings. Whilst in certain domains such as genomics, these processes are well-established, in other disciplines good practice is less commonplace.

In addition we do not have adequate evidence describing the requirements of researchers regarding the sort of tools they need to re-purpose and add value to data sets. There are a number of visualisation tools in use, and the National Text Mining Centre is developing tools for manipulating and analysing textual corpora, but what tools and services are required for data manipulation, annotation, re-interpretation, transformation and knowledge extraction?

REC 19. All relevant stakeholders should commission a study to evaluate the re-purposing of data-sets, to identify the significant properties which facilitate re-use, and to develop and promote good practice guidelines and effective quality assurance mechanisms.

REC 20. JISC should initiate a survey to gather user requirements from practising researchers to inform the development of value-added tools and services to interpret, transform and re-use data held in archives and repositories.

6.4 Technical Integration and Interoperability

The adoption of common standards by any community provides a robust foundation for successful data integration, sharing and interoperability. The data curation procedure at EBI for the management of microarray data, is greatly facilitated by the adoption of a suite of community standards centred around the MIAME standard (Minimum Information About a Microarray Experiment), with co-ordinated development by the international MGED Society. The relative value of "good-enough" versus "completely comprehensive" descriptions was raised in this context. In addition, there is growing awareness of the value of ontologies within the genomics community, focussing on application of the Gene Ontology, however there is still

resistance to use and curators at EBI do much re-annotation of datasets following submission. EBI are working on a new high-level data model to deal with this unwelcome trend. What is the picture in other disciplines and sub-disciplines? Are there published guidelines for levels of metadata descriptions, classification and taxonomies, factual name authorities and other semantic details? Once again, in-depth case study exemplars of good practice would be valuable. Work on developing name authorities for people, within the library and archive communities, should be referenced. Application profiles for data sets were briefly discussed in the previous section. The JISC is developing the IE Metadata Schema Registry as a place of publication for application profiles. What provision is made for application profiles for data collections? What are the funding and maintenance arrangements for such registries?

REC 21. JISC should work with domain partners to identify and promote the mechanisms which have been successful in achieving intra-disciplinary consensus on community data standards.

REC 22. An assessment of the effectiveness of registries and other infrastructural services for the development and adoption of community data standards, is needed.

A range of identifier schemes is in use to describe data sets. Taking one domain as an example, in crystallography, the eBank / eCrystals repository uses DOIs, the SPECTRa project uses CNRI Handles, and both use the domain-based InChI. It is understood that the new Diamond facility is planning to use an "inhouse" standard RB (Rutherford Beam) number as an identifier but this is not unique, and currently STFC (was CCLRC) has no plans to use either DOIs, handles or InChIs to describe their crystal structures (N.B. information taken from pre-publication eBank Phase 3 interview). Taking another example, EBI assigns a unique accession number to each dataset and recommends that authors use "array design identifiers" to describe microarray data-sets, however in practice most authors cite an experiment identifier. The more familiar Life Science Identifiers (LSIDs) are not used at all, because they were unstable when ArrayExpress was developed. Aside from the varied practice in assigning identifiers, different identifiers display different resolution functionality, which has implications for service development.

The management of versions of datasets is clearly a complex and very challenging issue for many data centres. Does the underlying hierarchical archival model scale to this extent? At what level of granularity should datasets be versioned? What is the role of annotation in describing different versions? Should all versions be made public? The most recent version? The "best" one? Who is responsible for making changes to submitted datasets? How are version changes made known? How are other (linked or dependent) services made aware of version changes?

The Distributed Annotation System (DAS) used at EBI, founded on the principle of distributed annotations with a three-level model, may have wider applicability elsewhere. Annotations are an integral aspect of many data publication processes and identifying some common standards and good practice guidance would be valuable.

REC 23. JISC should commission development work to investigate the application of identifiers to datasets and produce guidelines for good practice in data citation.

REC 24. There is a need for technical work to determine models and best practice for version control of complex datasets.

REC 25. JISC should work with other stakeholders to investigate different annotation models and standards for datasets, and to develop guidelines for good practice.

The ability to link objects, concepts, data and services, creates the matrix of network inter-relationships, which underpins the Web environment. We have seen in pioneering projects such as eBank, that the linking of primary data to textual interpretations of that data, is a very powerful and valuable feature, and this is now being explored by a number of other JISC-funded repository projects such as SPECTRa and CLADDIER. This study has highlighted a further potential outcome of the Web-based linking of data to text. There is some evidence that such virtual links may facilitate real connections between physical services i.e. between data centres and institutional repositories in Libraries. Taking this thinking a step further, this virtual linking

may also facilitate partnerships between people i.e. between the data community and the Library, information and publisher communities. Some supporting evidence has been observed during the course of the eBank project, though it was presented as a requirement for an interdisciplinary team with a mix of computing, library and domain skills. It also appears to be an (as yet intangible) outcome of the CLADDIER Project, with institutional repository staff working in partnership with BADC.

The ability to link related digital objects provides the exciting potential to connect empirical data elements with related derived structures and interpretations, across disciplinary boundaries. For example, crystal structures contained in the eCrystals institutional data repository, may be linked to related protein structures stored in the Protein Data Bank, or with related small molecules described in PubChem or to scholarly papers in IUCr journals or in Wellcome Trust sponsored sources such as UKPMC. For this matrix of links to be fully implemented and of real scientific value, the linking mechanism(s) need to be robust and reliable, reciprocal, scaleable and ideally automated, with some autonomic (self-healing) features. Interdisciplinary data sources such as PDB, eCrystals and PubChem, need to adopt a core set of common standards including the assignment of identifiers and the use of ontologies. Published data dictionaries and semantic mappings will assist in building these relationships between services. The OAI Object Re-Use and Exchange (ORE) Project[63] is working to develop technical solutions in this general area.

REC 26. JISC should fund repository technical development projects which build on OAI-ORE work and create robust, bi-directional interdisciplinary links between data objects and derived resources.

Discovery mechanisms also need to operate across disciplines. The NERC DataGrid is aiming to provide seamless discovery of datasets across NERC-funded data centres. This concept needs to be extended to operate across a much wider canvas. Search and discovery of data sets needs to be implemented across institutional repositories, data centres and data archives, blogs, wikis, peer-reviewed publications, databases and any other published data. The Intute Search Project is developing cross-search functionality for eprints as a part of the Integrated Information Environment, and the application profile for scholarly works is expected to provide the foundation for this search function. The development and adoption of a similar common standard schema would greatly assist delivery of a cross search function for data holdings, however this may be much more difficult to achieve, given the inherent complexity and heterogeneity of data sets. This raises a challenging area of conflicting requirements. For the integrated and interoperating environment envisaged, a relatively simple core set of metadata elements would underpin discovery. However, domain experts, require a much more detailed set of descriptors working at a level of granularity that far exceeds the Dublin Core standard. There is also the dual requirement for human discovery and for machine-to-machine discovery. In the future, one can assume that with data sets in particular, there will be increasing requirements for machine-to-machine services such as data mining, visualisation and transformation, all of which are dependent on prior data discovery.

Finally, data holdings in repositories and data centres need to expose their metadata to Google and other search engines to facilitate wider discovery. The contents of many institutional repositories are visible in Google Scholar; data holdings could be similarly exposed.

REC 27. JISC should fund technical development projects seeking to enhance data discovery services, which operate across the entire data and information environment.

6.5 Legal and Ethical Issues

Rights issues have been raised in a number of the interviews and discussions during this study. Particular domains were highlighted where the establishment and clearance of intellectual property rights is difficult and this is acting as a barrier to making data sets publicly available. These include performing arts data and geospatial data sets, where licences create restrictions on the use and re-purposing of datasets, and permissions must be acquired and correct attribution made in such cases. The GRADE Project has recently published a paper outlining a licensing strategy for the sharing and re-use of geospatial data in the academic sector[64]. It

explores the legal position regarding the deposit of geospatial data in a repository and the extraction of source data for various purposes.

Further complex issues arise in the case of population data, patient data and other epidemiological materials, where there are specific requirements for confidentiality, consent and pseudonymisation to enable access and use of particular data collections. These types of issue are commonplace in managing MRC, Wellcome Trust and ESRC funded datasets. The UK Data Archive makes use of a Special Licence for such cases.

For the researcher, the rights area is something of a "minefield" and many are quite understandably cautious about releasing their data for wider access and use, because of concerns over rights issues. The JISC funds the JISCLegal Service, which provides a range of advice and guidance to the UK academic community. It seems however, that there is a real need for enhanced advice and support targeted at the research community, in particular around the deposit and re-purposing of datasets as part of the scholarly communications process. There is also scope for more clarity about the position of institutions as publishers of data and around their liability in this context. This requirement for clarity of law and how to interpret it was voiced during the study, however it was acknowledged that this is a challenging area requiring expert intervention.

The JISC is currently funding a study by Naomi Korn, Charles Oppenheim and Charles Duncan, which is examining IPR and licensing issues in derived data from a UK perspective. JISC is also exploring the development of model licences for use in disclosing research outputs to the wider community. This follows on from the JISC Licence to Publish for journal article publication. These licences will be based on the Creative Commons licences and Science Commons work. NERC is working on a "licence to publish," and the UKDA and AHDS also use model licences for deposit, which are available on the Web. More co-ordination in this area would clearly be advantageous.

REC 28. JISCLegal should provide enhanced advice and guidance to the research community on all aspects of IPR and other rights issues relating to data sets.

REC 29. Work by JISC and the research councils, on developing model licences, should be co-ordinated so that a minimum set of standard licences are adopted more widely.

6.6 Sustainability

Whilst there are a number of well-established data centres providing a wide range of infrastructural curation and preservation services to particular domains, the funding for these centres has accrued incrementally. The EBI as one of the most fully-developed of these centres, has a very significant budget contributed from a number of funding sources, to carry out the complex curation processes required for the omics data. Many highly expert curators are employed, the reference databases are vast and themselves form part of the new e-research infrastructure. In contrast, in other areas such as engineering, there are no equivalent data centres to provide supporting services and tools to that community. Clearly in the past, the various funding bodies have adopted different strategic positions regarding infrastructure provision, with some accepting responsibility for such services, whilst others view the provision of infrastructure as an institutional responsibility. More recently digital repositories have emerged as a component of the institutional e-infrastructure and whilst currently there are few repositories containing research data, this picture may well change. Looking across the funding landscape, whilst there may be valid reasons for these contrasting strategic positions regarding infrastructure services, this divergent set of views is not constructive and is certainly not in the best interests of UK research.

The increasing volumes of data generated from eScience applications, from new large-scale facilities such as Diamond and from the myriad of "small science" programmes, suggest that this rather diverse approach to data curation and preservation should not continue. It is time for a financial review of the situation: we need to understand the full costs of curating data so that we can plan and budget for the required infrastructure more effectively. We need to construct new economic models to provide a robust foundation to funding plans, models which link research strategy and programme development, to operational support and infrastructure provision.

All research funding bodies should now recognise the emerging need for new infrastructure to curate and preserve the UK scientific heritage, and should provide appropriate funding. The e-infrastructure should include a comprehensive network of data centres and repositories with expert staff trained as professional data curators and data scientists, adequate data storage facilities, tools, value-added services and a long-term strategy to ensure their continued development and maintenance over time. Data centres and repositories should have clear and accountable funding lines, with 5-10 year planning horizons.

REC 30. The JISC should work in partnership with the research funding bodies and jointly commission a cost-benefit study of data curation and preservation infrastructure.

REC 31. The JISC should commission work to construct new economic models for preservation and data sharing infrastructure, to develop sustainable solutions.

6.7 Advocacy

A number of recent surveys such as StORe Project results, demonstrate that awareness and understanding of data curation and preservation good practice is generally poor, although there is variability across disciplines. Given this very low baseline, what are the awareness, promotion and advocacy requirements? How should advocacy be delivered most effectively? Researchers need to be made aware of their curatorial responsibility. To achieve this goal, the community needs to be sufficiently well-informed to be able to devise adequate Data Management Plans for newly funded research activity. During many interviews, it was stressed that advocacy needs to be targeted and tailored to specific disciplines and sub-disciplines. Advocacy messages need to be clear and consistent, and ideally will be harmonised across funding bodies. Surveys suggest researchers particularly need advice concerning technical standards for data curation. The DCC has an important role to play in this context.

This study revealed a range of outreach approaches for the promotion of data curation and preservation good practice, and a mix of methods was found to be most effective. These include talking to stakeholders and researchers at conferences, organising roadshows, running workshops at conferences, visiting PIs at the start of projects to make the case for data deposit, and identifying academic staff as advocacy champions within departments. Data centres and institutional repository staff need to go out and pro-actively promote why they exist, and why they should be used. It may be useful to focus on the postgraduate community in the first instance. An example of effective outreach was demonstrated by EBI.

The view was voiced that outreach and promotion of data curation and preservation, is a shared responsibility, and a role for funding organisations and data centres alike. Clearly, political incentives will help to promote data-sharing, and the data deposit requirement prior to article publication, may be the most effective incentive for data sharing.

REC 32. The DCC should promote co-ordinated advocacy programmes, targeted at specific disciplines, and which address technical aspects of data collection and deposit.

6.8 Training and Skills

The awareness and skill levels of researchers regarding data management, is variable. Community practice in data sharing is also polarised. Evidence suggests that in addition to advocacy programmes, there is a requirement for training materials and community skills development. Much expertise is concentrated in data centres and we need to consider how best to encourage and formalise a flow of expertise from these data centres to institutions, where staff are being appointed to manage and develop repositories. Various approaches have been employed to raise community skill levels, including the development of "How to" guides, training tutorials, courses, and use of e-learning materials. The EBI is a good example demonstrating a pro-active approach to training, with the result that the genomics community is relatively well-informed regarding data curation practice. The DCC will play a key role in up-skilling the community through the provision of the Data Curation Manual, tools, services, workshops and other supporting materials. In addition a number of EC-funded projects such as Digital Preservation Europe (DPE) and PLANETS, are also developing substantive training materials

and programmes. However, there remains a critical and growing requirement to develop workforce capacity in data curation skills and competencies.

REC 33. The DCC should collaborate with other parties to deliver co-ordinated training programmes and supporting materials, targeted at researchers in specific disciplines, to build workforce capacity within the sector.

The role of curators has been referenced at several points within this study. The professional curators at EBI have valuable skills in data curation, accompanied by in depth subject knowledge. One interpretation of curators who collect, describe and connect data, is the idea of the community proxy role, (referenced in the NSF *Long-lived Data Collections Report*), where curators try to understand the ontology of the domain and define standards, working in partnership with the practising scientists. Frequently the work of the data curator is not acknowledged in publications and this needs to be addressed. The data scientist or data curator, plays a key role in the scholarly publication process and deserves to be acknowledged and rewarded appropriately. There is also scope for greater convergence of subject expertise with library and archival skills, as part of professional development and training.

It was observed during the study, that the "net generation" is more open to data sharing. Indeed, it is likely that over time, "native data scientists" (a term used by Liz Lyon in a Keynote to the DCC Conference 2006), will emerge through the acquisition of relevant skills learnt through the standard educational curriculum. However, for now, there is a dearth of skilled practitioners, and data curators play an important role in maintaining and enhancing the data collections that represent the foundations of our scientific heritage.

REC 34. A study is needed to examine the role and career development of data scientists, and the associated supply of specialist data curation skills to the research community.

REC 35. JISC should fund a study to assess the value and potential of extending data handling, curation and preservation skills within the undergraduate and postgraduate curriculum.

Key chronological dependencies within the set of Recommendations are also noted. The Disciplinary DataSets Mapping and Gap Analysis (Rec 1) will inform the development of a co-ordinated Data Curation and Preservation Strategy (Rec 2) by research funding organisations. The Data Audit Framework (Rec 4) will ideally be in place to form a common basis for the development of institutional Data Management, Preservation and Sharing Policies (Rec 10).

6.9 Roles, rights, responsibilities and relationships

One key element of the brief for this study was to investigate roles and responsibilities of data centres and institutions creating data repositories and to examine the relationships between parties. In discussion with Mark Thorley, NERC, it emerged that "roles, rights and responsibilities" are the themes upon which the revised NERC data policy will be based (due to be published in October at the NERC Data Management Conference to be organised by DCC/UKOLN). The Table below was informed by the slides presented by Mark at the consultation workshop, however it has been reviewed and extended as a result of the findings from this study and now contains a number of new elements. These include: a) a column for "relationships", since co-ordination and working in partnership is seen to be critical to the successful development of e-infrastructure services for data, and b) a line for "institution", since scientists and researchers are usually employees of an institution, which itself has roles, rights, responsibilities and relationships. Note that entities may have multiple roles e.g. an institution may also have a role as a publisher.

Table 1 Summary Table of Roles, Rights, Responsibilities and Relationships (with acknowledgement to Mark Thorley, NERC).

Role	Rights	Responsibilities	Relationships
Scientist: creation and use of data	Of first use. To be acknowledged. To expect IPR to be honoured. To receive data training and advice.	Manage data for life of project. Meet standards for good practice. Comply with funder / institutional data policies and respect IPR of others. Work up data for use by others.	With institution as employee. With subject community With data centre. With funder of work.
Institution: curation of and access to data	To be offered a copy of data.	Set internal data management policy. Manage data in the short term. Meet standards for good practice. Provide training and advice to support scientists. Promote the repository service.	With scientist as employer. With data centre through expert staff.
Data centre: curation of and access to data	To be offered a copy of data. To select data of long-term value.	Manage data for the long-term. Meet standards for good practice. Provide training for deposit. Promote the repository service. Protect rights of data contributors. Provide tools for re-use of data.	With scientist as "client" With user communities. With institution through expert staff. With funder of service.
User: use of 3rd party data	To re-use data (non-exclusive licence). To access quality metadata to inform usability.	Abide by licence conditions. Acknowledge data creators / curators. Manage derived data effectively.	With data centre as supplier. With institution as supplier.
Funder: set/react to public policy drivers	To implement data policies. To require those they fund to meet policy obligations.	Consider wider public-policy perspective & stakeholder needs. Participate in strategy co-ordination. Develop policies with stakeholders. Participate in policy co-ordination, joint planning & fund service delivery. Monitor and enforce data policies. Resource post-project long-term data management. Act as advocate for data curation & fund expert advisory service(s). Support workforce capacity development of data curators.	With scientist as funder. With institution. With data centre as funder. With other funders. With other stakeholders as policy-maker and funder of services.
Publisher: maintain integrity of the scientific record	To expect data are available to support publication. To request pre-publication data deposit in long-term repository.	Engage stakeholders in development of publication standards. Link to data to support publication standards. Monitor & enforce public. standards.	With scientist as creator, author and reader. With data centres and institutions as suppliers.

The polarisation of views regarding the role of institutional repositories for data was marked. The principle of an academic researcher or scientist depositing their data "somewhere" in a managed repository is a good starting point, since at present even this does not universally occur. Scientists and researchers need guidelines to help them to deposit their data in the appropriate location(s). EBI is styled as a custodian of data and data centres embrace a stewardship role in their data curation activities. Institutions also have a stewardship role for data created by their employees, which in general is yet to be recognised, and institutional repositories may provide a viable working solution[65]. The Edina/SHERPA Prospero project is developing the concept of "TheDepot," which is a repository for institutions that don't have an IR ("Put it in the Depot" slogan). The Depot[66], which is based on ePrints.org software, will be launched in June 2007.

It is tempting to propose that institutional repositories have a primary role as a short-term, easily accessible store, whilst data centres and data archives have a more long-term role in preservation. Institutional repositories are relatively new structures and their ability to survive as robust, trustworthy archives is unknown, so this simplistic construct may be a little premature. What perhaps is of greater importance is the reliability, robustness and scope of connections and linkages between the content, the workflows and mechanisms for data migration between repositories, and the maintenance of versions of the data if it is in multiple locations. Federation models for repositories are developing around organisations, disciplines, geographical regions and software platforms, and a mixed economy seems likely for at least the immediate future.

It was interesting to learn that data centres also act as facilitators for research councils and for the researcher (in locating and securing data from other sources such as the Met Office, in the case of BADC). Data centre managers and directors also welcomed feedback from users and funders on the data stored, and this two-way dialogue is to be encouraged. For advocacy and training initiatives, the DCC has the potential to act as a "catalyst"

Whilst particular named individuals in key roles will have responsibility for data management within their own organisation, the importance of co-ordination with peer groups and peer organisations is of critical importance. Institutions, funders, data centres and disciplinary communities all have a responsibility to work together to derive shared and agreed policies and good practice, to ensure the continuing effective provision of data curation and preservation services, and a more open approach to data sharing.

7 Modelling Data Flow

Exemplars of the various elements of the emerging data infrastructure from this study, are mapped to the JISC Information Environment in Figure 8.

Two new high-level data flow models derived from the Summary Table and study findings are proposed, which represent contrasting examples of data curation good practice.

The Domain Data Deposit Model in Figure 9, is based on a strong integrated community foundation with well-established common standards, policies and practice. It is supported by clear funder policy, advocacy and investment. The data centre provides training to the community and proactively promotes its services. Disciplinary examples are "omics" and the social sciences.

Figure 10 shows a Federation Data Deposit Model, where groups of repositories have joined together in a federation, which is based on some agreed level of commonality documented in some form of partner agreement, but where there is a broader practice base. In the diagram, the federation has an institutional foundation, but it might equally be based on software platform, format type or regional geographic boundaries. Policy, advocacy and training are provided primarily within the federation. The federation is supported by strong institutional buy-in, and there may be little or no investment by research funding organisations. There will be some common technical standards, but the maturity of these may be variable. An example is the eCrystals Federation.

Figure 8. Mapping data infrastructure components to the JISC Information Environment.
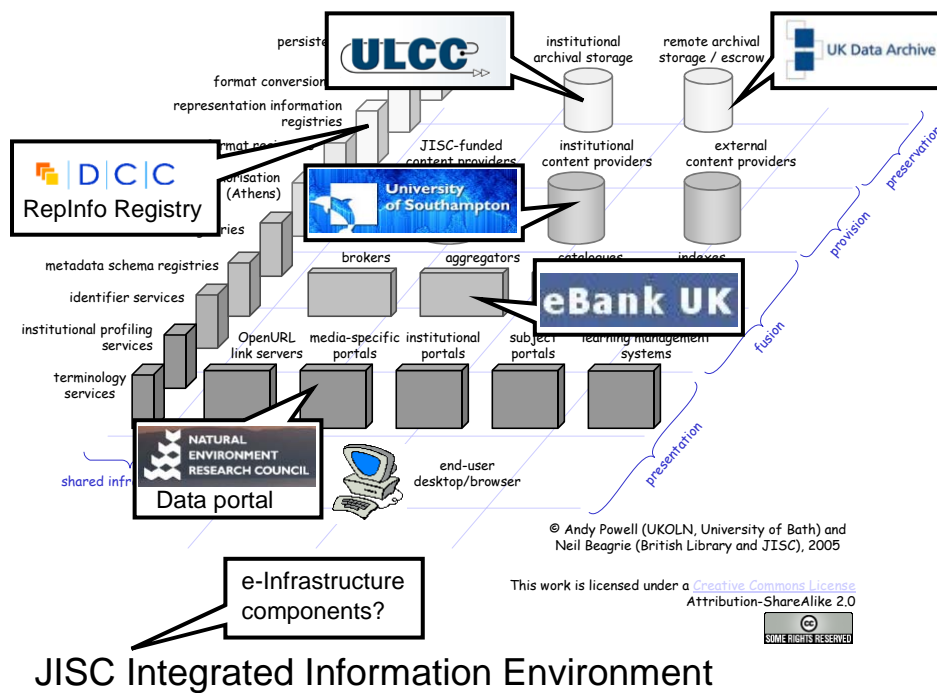

JISC Integrated Information Environment

Figure 9 Domain Data Deposit Model.

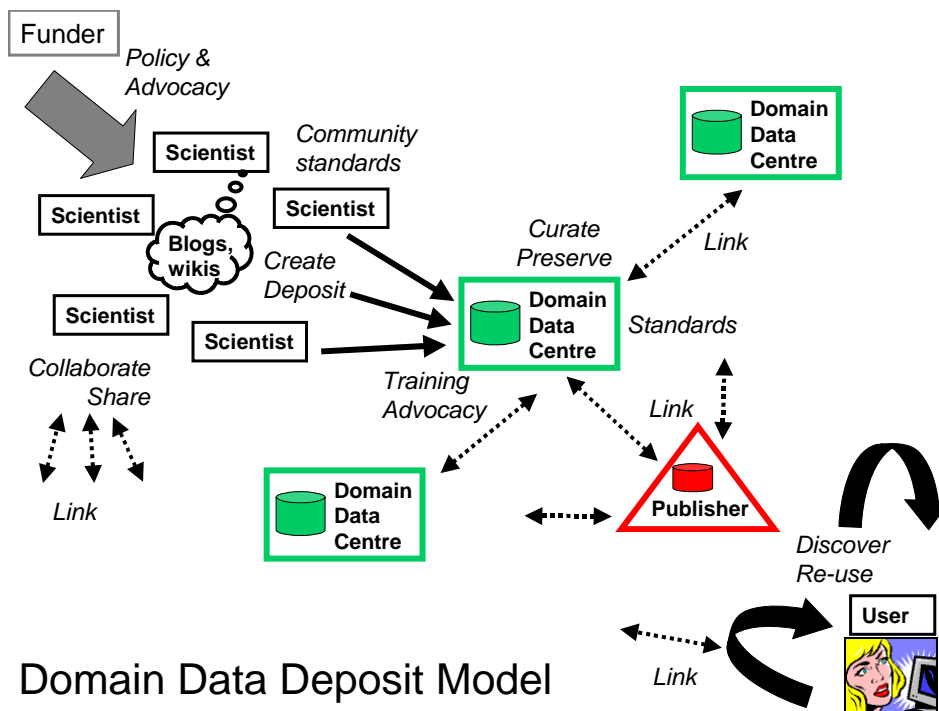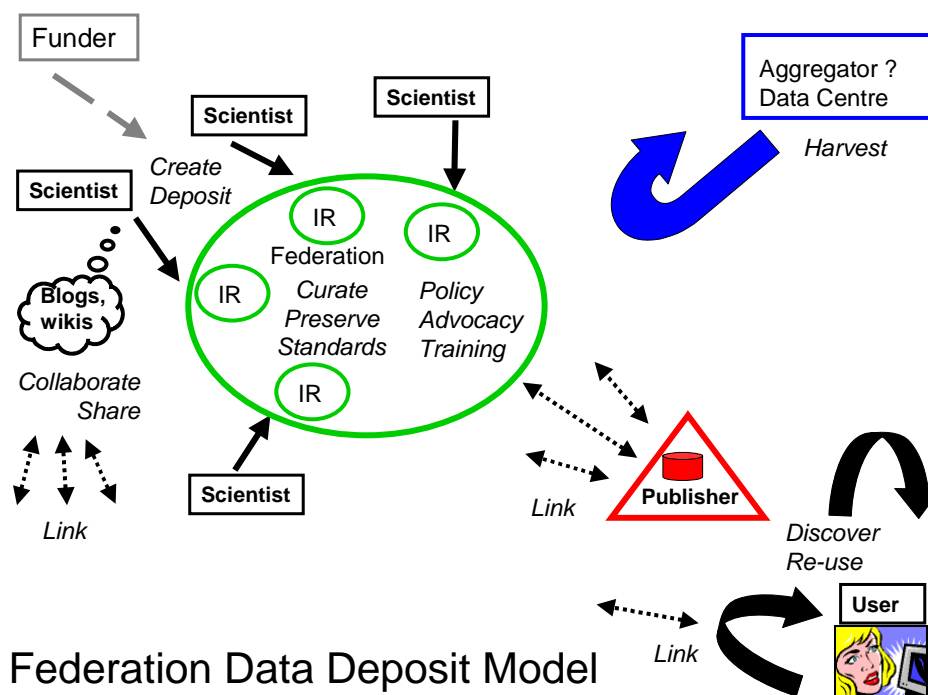Domain Data Deposit Model

Figure 10 Federation Data Deposit Model

Federation Data Deposit Model

8 Conclusions

The findings of this study demonstrate that whilst there is good practice in data curation developing in some domains at both the strategic and operational levels, there is much work still outstanding and urgent. There is a real need for tangible leadership and cross-domain strategic co-ordination, ostensibly at the highest levels of research funding organisations, to put in place the infrastructure and services to effectively manage the burgeoning data deluge. This co-ordination activity will be greatly informed by a detailed picture of current data holdings. Institutions have a particular role in implementing effective data management systems for research data outputs, and the evidence gathered suggested that a systematic audit of practice in this area is needed. There is much scope for enhanced planning approaches across the piece, and the development of best practice guidance for the working scientists, to help them fulfil their responsibilities as data creators and authors.

The landscape showed an emerging data infrastructure, which is both fragmented and incomplete. Consensus on community data standards is patchy and as yet, within the domain constraints of this study, there appears to be limited common foundations. There are a number of technical issues where further work is needed, including addressing the use of persistent identifiers for data, the versioning of complex datasets and deriving models for the annotation of datasets. There is also scope for some standardisation of licences to manage IPR arising from the creation and re-use of research data outputs. Whilst these points largely address the technical sustainability of the e-infrastructure, there is a major requirement for a better understanding of the economic implications of providing such an infrastructure over the longer term.

Moving on to consider the human aspects of data curation activities, many researchers appear to be unaware of the range of issues associated with data management best practice and there is a growing requirement for co-ordinated advocacy, training and skills programmes to equip the

research community with the appropriate competencies to foster the Science Commons envisaged for the future.

From the outcomes of the findings, a full set of Recommendations are presented together with a Summary Table of Roles, Rights, Responsibilities and Relationships, which attempts to capture the essence of the Report. The Summary Table is illustrated by two high-level Data Models.

9 Appendix: Interview pro-forma, Interviewees and workshop participants

UKOLN Data Consultancy Semi-Structured Questionnaire

Interviewer: Liz Lyon

Interviewee:

Role/position:

Department:

Institution:

Date:

1. *Can you give more detail on the specific duties associated with your role/position?*

2. *What is the status / type / platform / content and maturity of your digital repository/ies (DR)?*

3. *What is the primary purpose of your DR? (Tick as many as relevant)*

   a) *Enhancing access to intellectual assets*

   b) *Facilitating dissemination of research record*

   c) *Long-term preservation*

   d) *Managed digital storage*

   e) *Other*

4. *Who is responsible for the DR:*

   a) *Strategic decision-making, policy development, collaboration?*

   b) *Technical implementation and development?*

   c) *User support?*

   d) *Funding?*

   e) *Other?*

5. *Does your DR contain research data? If so what disciplines? Data files? File formats? Raw and/or processed data?*

6. *Describe the deposit process? Who? How? When? Workflows?*

7. *Is research data deposited in data centres? Which departments / disciplines? Raw / processed? Is it also in the IR?*

8. *Describe the data centre deposit process? Who? How? When? Workflows?*

9. *What is the relationship between the data centre and the institutional/departmental repository?*

10. *Are there formal agreements in place? What are they?*

11. *What is the position regarding:*

   a) *IPR*

b) *Versioning*

c) *Preservation responsibility*

12. *What are the barriers and enablers to a collaborative approach?*

13. *Do you have any comments / views on adoption, take-up, embedding and cultural change icw DR?*

14. *How could the relationships between data centres and IRs be enhanced?*

15. *If you could identify three key issues to be addressed what would they be?*

16. *Any other comments?*

Workshop and Interview Participants

Organisation	Name
AHRC ICT Programme	David Robey
AHDS	Sheila Anderson
BADC	Sam Peplar
STFC (WAS CCLRC)	Brian Matthews, Matthew Wild, Stephen Crothers, Xiaobo Yang, Esther Conway
DCC	Chris Rusbridge, Manjula Patel, Bridget Robinson
eBank / University of Southampton	Simon Coles
EBI	Graham Cameron, Helen Parkinson
EDINA	Peter Burnhill
EPSRC	Sue Smart
ESRC	Sian Bourne
GRADE	Dave Medyckyj-Scott, Rebecca Seymour
IUCr	Brian McMahon
MRC	Allan Sudlow
NERC	Mark Thorley
RIN	Michael Jubb, Stephane Goldstein
SPECTRa / University of Cambridge	Peter Morgan, Peter Murray-Rust, Henry Rzepa, Alan Tonge, Jim Downing
STORE	Graham Pryor
UKDA	Kevin Schurer, Matthew Woollard
UKOLN	Rachel Heery, Traugott Koch

ULCC	Kevin Ashley
University of Edinburgh	Theo Andrew
Wellcome Trust	Robert Terry

10 References

[1] Research Information Network *Stewardship of digital research data: draft for consultation.* (Research Information Network, April 2007) http://www.rin.ac.uk/data-principles

[2] Hey, T. and Trefethen, A. "The data deluge: an e-science perspective." In *Grid computing: making the global infrastructure a reality*, ed. F. Berman, G. Fox and T. Hey (Chichester: Wiley, 2003), 809-24 http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/research/esci/datadeluge.pdf

[3] Lord, P. and Macdonald, A. *e-Science curation report* (Joint Information Systems Committee, 2003) http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

[4] HM Treasury, Department of Trade and Industry, and Department for Education and Skills, *Science and innovation investment framework 2004-2014* (HMSO, 2004) http://www.hm-treasury.gov.uk/spending_review/spend_sr04/associated_documents/spending_sr04_science.cfm

[5] OSI e-Infrastructure Working Group, *Developing the UK's e-infrastructure for science and innovation* (Office of Science and Innovation, 2007) http://www.nesc.ac.uk/documents/OSI/index.html

[6] NST Cyberinfrastructure Council, *NSF's Cyberinfrastructure vision for 21st century discovery*, v 7.1 (National Science Foundation, July 20, 2006) http://www.nsf.gov/od/oci/ci-v7.pdf

[7] e-Research Coordinating Committee, *An e-Research Strategic Framework: interim report* (Australian Government, Department of Education, Science and Training, September 30, 2005) http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/e_research_consult/interim_report

[8] Agencies join force to share data *Nature* 446 (March 22, 2007) 354 http://www.nature.com/nature/journal/v446/n7134/full/446354b.html

[9] *Towards 2020 Science* (Microsoft Research, 2004) http://research.microsoft.com/towards2020science/

[10] 2020 Computing special issue, *Nature* 440 (March 23, 2006) 409-419 http://www.nature.com/nature/journal/v440/n7083/index.html

[11] Ball, P. "The common good." news@nature.com (August 20, 2004) doi:10.1038/news040816-14 http://www.nature.com/news/2004/040816/pf/040816-14_pf.html

[12] Carlson, S. "Lost in a sea of science data." *The Chronicle of Higher Education* (June 23, 2006) http://chronicle.com/free/v52/i42/42a03501.htm

[13] Lyon, L. Keynote address: Reflections on open scholarship: process, product and people. 2nd International Digital Curation Conference, Glasgow, UK, November 21-22, 2006 http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/presentations.html#2006-11-dcc-conference

[14] OpenWetWare http://openwetware.org/wiki/Main_Page

[15] eBank Project http://www.ukoln.ac.uk/projects/ebank-uk/

[16] Lyon, L. "eBank UK: Building the links between research data, scholarly communication and learning." *Ariadne* 36 (July 2003) http://www.ariadne.ac.uk/issue36/lyon/

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

[18] *Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communiqué* (Organisation for Economic Co-operation and Development, January 30, 2004) http://www.oecd.org/document/15/0,2340,en_21571361_21590465_25998799_1_1_1_1,00.html

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

[20] *Research and the scholarly communications process: Towards strategic goals for public policy* (Research Information Network, March 2007) http://www.rin.ac.uk/sc-statement

[21] *Research funders' policies for the management of information outputs* (Research Information Network, January 2007) http://www.rin.ac.uk/files/Funders'%20Policy%20&%20Practice%20-%20Final%20Report.pdf

[22] ROARMAP Service http://www.eprints.org/openaccess/policysignup/

[23] Petition for guaranteed public access to publicly-funded research results http://www.ec-petition.eu/

[24] *Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee on Scientific Information in the Digital Age: Access, Dissemination and Preservation* (Commission of the European Communities, February 14, 2007) http://ec.europa.eu/information_society/activities/digital_libraries/doc/scientific_information/communication_en.pdf

[25] Open access - complying with the Wellcome Trust grant requirements http://www.wellcome.ac.uk/assets/wtx033844.pdf

[26] Science & Technology Facilities Council http://www.stfc.ac.uk/

[27] Heery, R. and Powell, A. *Digital Repositories Roadmap: looking forward* (UKOLN, University of Bath; Eduserv Foundation, April 2006) http://www.ukoln.ac.uk/repositories/publications/roadmap-200604/

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

[29] Robinson, J. Repository Ecology, JISC Conference 2007, Birmingham, March 13, 2007 http://www.jisc.ac.uk/conference2007

[30] ROARMAP Service http://www.eprints.org/openaccess/policysignup/

[31] NSF's Cyberinfrastructure Vision for 21st Century Discovery http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

[32] National Science Board, *Long-lived digital data collections: enabling research and education in the 21st century* (National Science Foundation, May 23, 2005) http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf#search=%22long%20lived%20digital%20nsb%22

[33] EPSRC Statement on Access to Research Outputs http://www.epsrc.ac.uk/AboutEPSRC/ROAccess.htm

[34] MRC Policy on Data-sharing and Preservation
http://www.mrc.ac.uk/PolicyGuidance/EthicsAndGovernance/DataSharing/PolicyonDataSharingandPreservation/index.htm

[35] Avon Longitudinal Study of Parents and Children (ALSPAC)
http://www.alspac.bristol.ac.uk/welcome/index.shtml

[36] MRC National Survey of Health & Development (NSHD) http://www.nshd.mrc.ac.uk/

[37] Dukes, P. Making the most of our data, RIN Workshop, London, December 5, 2006
http://www.rin.ac.uk/files/Peter%20Dukes%20v2.pdf

[38] Lord, P., et al. *Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models* (UK e-Science Technical Report UKeS-2006-02, August 2005)
http://www.nesc.ac.uk/technical_papers/UKeS-2006-02.pdf

[39] NERC Data Centres http://www.nerc.ac.uk/research/sites/data/

[40] NERC Data Policy Handbook, v 2.2 (December 2002)
http://www.nerc.ac.uk/research/sites/data/policy.asp

[41] Noor, M.A.F., Zimmerman, K.J., and Teeter, K.C., "Data sharing: how much doesn't get submitted to GenBank?" *PLoS Biology* 4(7) (July 2006) 1113-14
http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pbio.0040228

[42] Wellcome Trust Policy on Data Management and Sharing
http://www.wellcome.ac.uk/doc_wtx035043.html

[43] AHDS http://ahds.ac.uk/index.htm

[44] iRODS http://irods.sdsc.edu/index.php/Main_Page

[45] Stormont Papers Project http://stormontpapers.ahds.ac.uk/index.html

[46] UK Data Archive Preservation Policy http://www.data-archive.ac.uk/news/publications/UKDAPreservationPolicy0905.pdf

[47] R4L deposit workflow
http://www.ukoln.ac.uk/repositories/digirep/index/All_the_Scenarios_and_Use_Cases_Submitted#R4L

[48] eBank aggregator service http://ebank.ukoln.ac.uk/

[49] Grainne Conole, *External evaluation of the eBank Project* (December 2006)
http://www.ukoln.ac.uk/projects/ebank-uk/evaluation-report-dec-2006/evaluation-report-december-2006.pdf

[50] Coles, S.J., Day, N.E., Murray-Rust, P., Rzepa, H.S., Zhang, Y., "Enhancement of the chemical semantic web through the use of InChI identifiers." *Organic & Biomolecular Chemistry*, 3 (2005) 1832-34 doi:10.1039/b502828k

[51] eBank Metadata schema application profile http://www.ukoln.ac.uk/projects/ebank-uk/schemas/

[52] Koch, T. *Terminology and subject access issues* (UKOLN, University of Bath, August 2006)
http://www.ukoln.ac.uk/projects/ebank-uk/dissemination/termino-public.html

[53] The Guardian newspaper's "Free our Data" Campaign, e.g.: Arthur, C., and Cross, M. "Give us back our crown jewels." *The Guardian*, March 9, 2006 http://technology.guardian.co.uk/weekly/story/0,,1726229,00.html

[54] Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007, establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) *Official Journal of the European Union*, 50, L 108 (April 25, 2007) 1-14 http://www.ec-gis.org/inspire/

[55] StORe Project http://jiscstore.jot.com/WikiHome

[56] *Researchers' Use of Academic Libraries and their Services report: a report commissioned by the Research Information Network and the Consortium of Research Libraries* (April 2007) http://www.rin.ac.uk/researchers-use-libraries

[57] DRAMBORA Digital Repository Audit Method Based On Risk Assessment http://www.repositoryaudit.eu/

[58] UCSD-Nature Signaling Gateway http://www.signaling-gateway.org/

[59] Coles, S. Capturing and sharing research data, JISC Conference 2007, Birmingham, March 13, 2007 http://www.jisc.ac.uk/conference2007

[60] PLoS ONE http://www.plosone.org/

[61] KIM Project http://www-edc.eng.cam.ac.uk/kim/

[62] CLADDIER briefing on Data Publication, May 2007 http://claddier.badc.ac.uk/trac/attachment/wiki/wp10-mtg/CLADDIER-datapub-briefing-20070514.pdf

[63] OAI Object Re-Use and Exchange Project http://www.openarchives.org/ore/

[64] Waelde, C. and McGinley, M. *Designing a licensing strategy for sharing and re-use of geospatial data in the academic sector* (GRADE Project, March 2007) http://edina.ac.uk/projects/grade/gradeDigitalRightsIssues.pdf

[65] Lynch, C. A. "Institutional repositories: essential infrastructure for scholarship in the digital age." *ARL Bimonthly Report* 226 (February 2003) http://www.arl.org/resources/pubs/br/br226/br226ir.shtml

[66] The Depot. http://depot.edina.ac.uk/