

Report of a workshop on research and development priorities to support research data curation.

Sponsored by the Andrew J Mellon Foundation and the Joint Information Systems Committee (JISC).

Held in Washington DC, 14 December 2007

Report by Neil Jacobs (JISC)

Contents

Workshop objectives.....	2
Case study: Bench science (Chemistry) – led by Liz Lyon, Simon Coles and Peter Murray-Rust	2
Case study: Humanities (Languages, Perseus Project) – led by Greg Crane.....	8
Infrastructure models for sustainability	10
Advocacy and dissemination	11
Institutional strategy, policy and planning	11
Legal and ethical.....	12
Human, cultural and partnership issues.....	12
Managing and funding research and development programmes.....	13
Conclusions and specific suggestions	15

Workshop objectives

The workshop was intended to:

1. Identify and document a small number of case studies to illustrate key issues that arise in planning and implementing infrastructure programmes related to the curation of research data
2. Enable and capture a discussion of these and other relevant issues among an invited group of experts
3. Develop recommendations, based on analysis of the case studies and experts' experience, on how to align the discipline-specific and infrastructural demands of programmes concerned with research data

Case study: Bench science (Chemistry) – led by Liz Lyon, Simon Coles and Peter Murray-Rust

Bench science produces heterogenous, often proprietary, data, and probably constitutes a larger problem space in terms of sheer volume, as well as complexity, than 'big science'. Processed data is typically on laptops or similar. Bench scientists don't really think about data curation, they don't value data highly but see the publication as the valuable output. Incentives that might encourage a change of view might include the potential for better data management in face of the increasing data deluge, data validation, data peer review, reuse/mining, reanalysis, and mandates. Having said that, departments are increasingly aware of the potential value of IPR in (eg) crystal data.

A departmental approach has been taken in some projects, with a departmental repository. Institutions are rather guarded about committing to curating several hundred varied databases / repositories from their departments. There is a working group at Southampton University in the UK looking at the implications that would flow from such a commitment. However, it is worth emphasising the potential importance of the academic department as a locus for advocacy and good practice in terms of data curation.

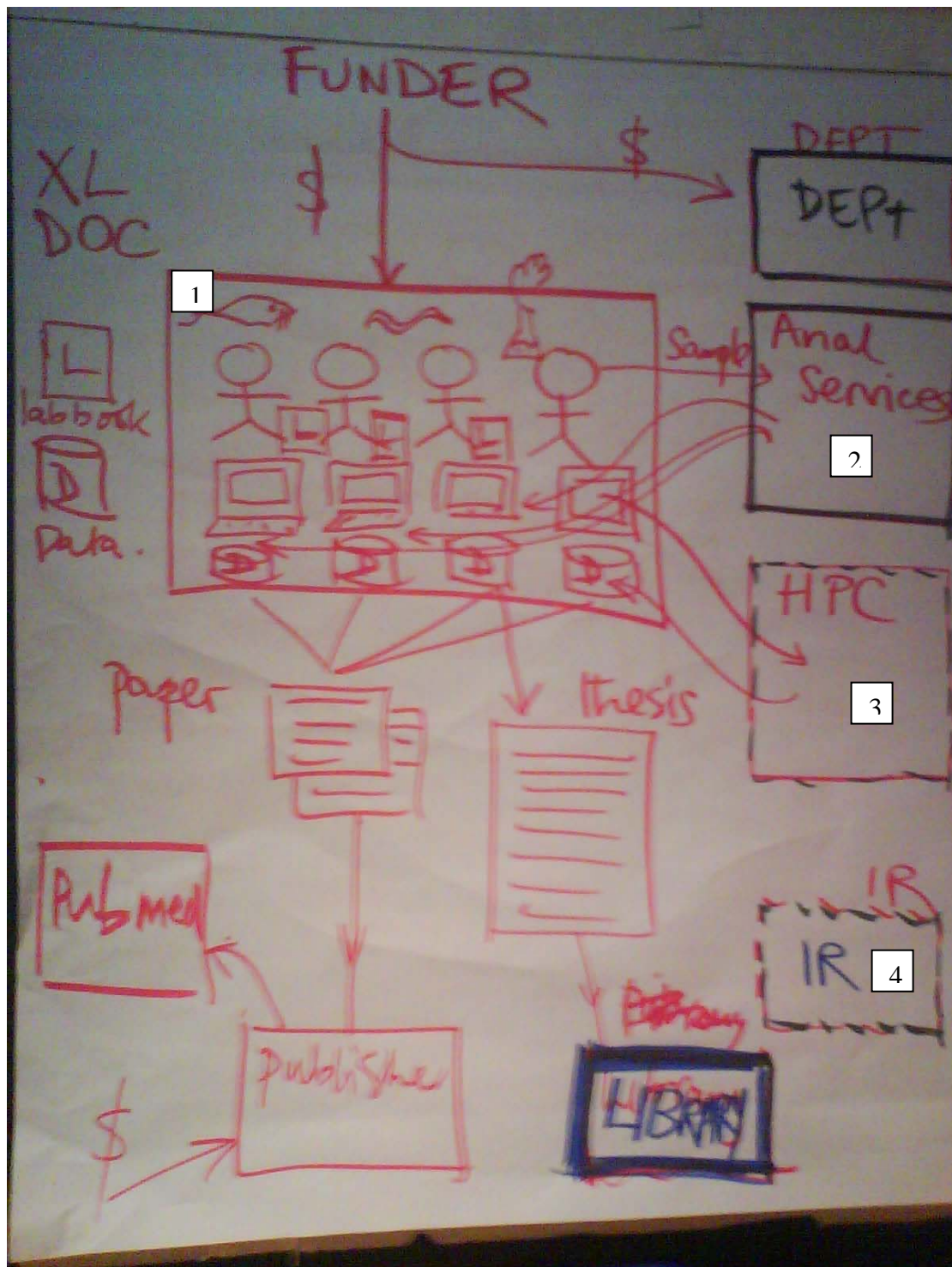
The instruments and other apparatus (cameras, sensors, etc) by which data is created often generate files in proprietary formats and are usually not designed as, or to be compatible with, curation environments. It has proved difficult for engage manufacturers in any moves to open standards, for obvious commercial reasons. Express demand from senior scientists would be the key driver for such moves, but this demand has not been evident so far.

Key developments that would aid progress toward better data curation include:

- tools and vocabularies that are sustained over time.
- lightweight laptop repositories that can be federated

A schematic was presented to describe the current situation (figure 1):

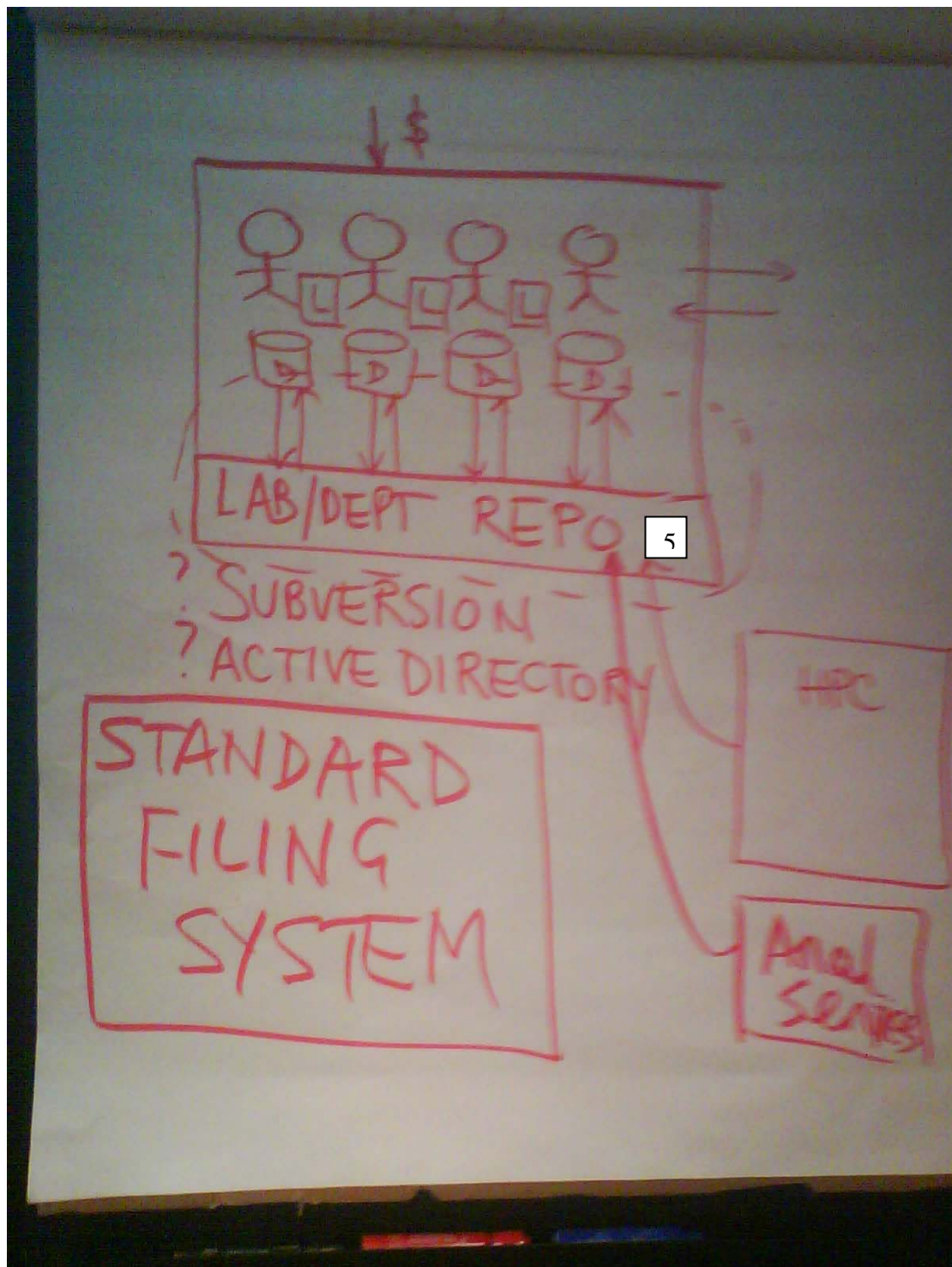
Figure 1: Bench science as practiced now



In this diagram, the departmental laboratory (1) is populated by researchers working with their own relatively independent machines (in which the data is stored), interacting externally with analysis services (2) and compute services (3). The role of the institutional repository (4) is not entirely clear.

This was then contrasted with a potential alternative laboratory set up (figure 2).

Figure 2: Bench science, revised infrastructure



The key intervention here is the introduction of a laboratory or departmental repository, giving a standard shared filing system, and which serves as the interface to analysis and compute services. Data is still held on the researchers' own machines, but it is replicated in the laboratory repository. Thus, from the researcher's point of view, the intervention is practically invisible (it is described simply as a 'backup'). This is important because

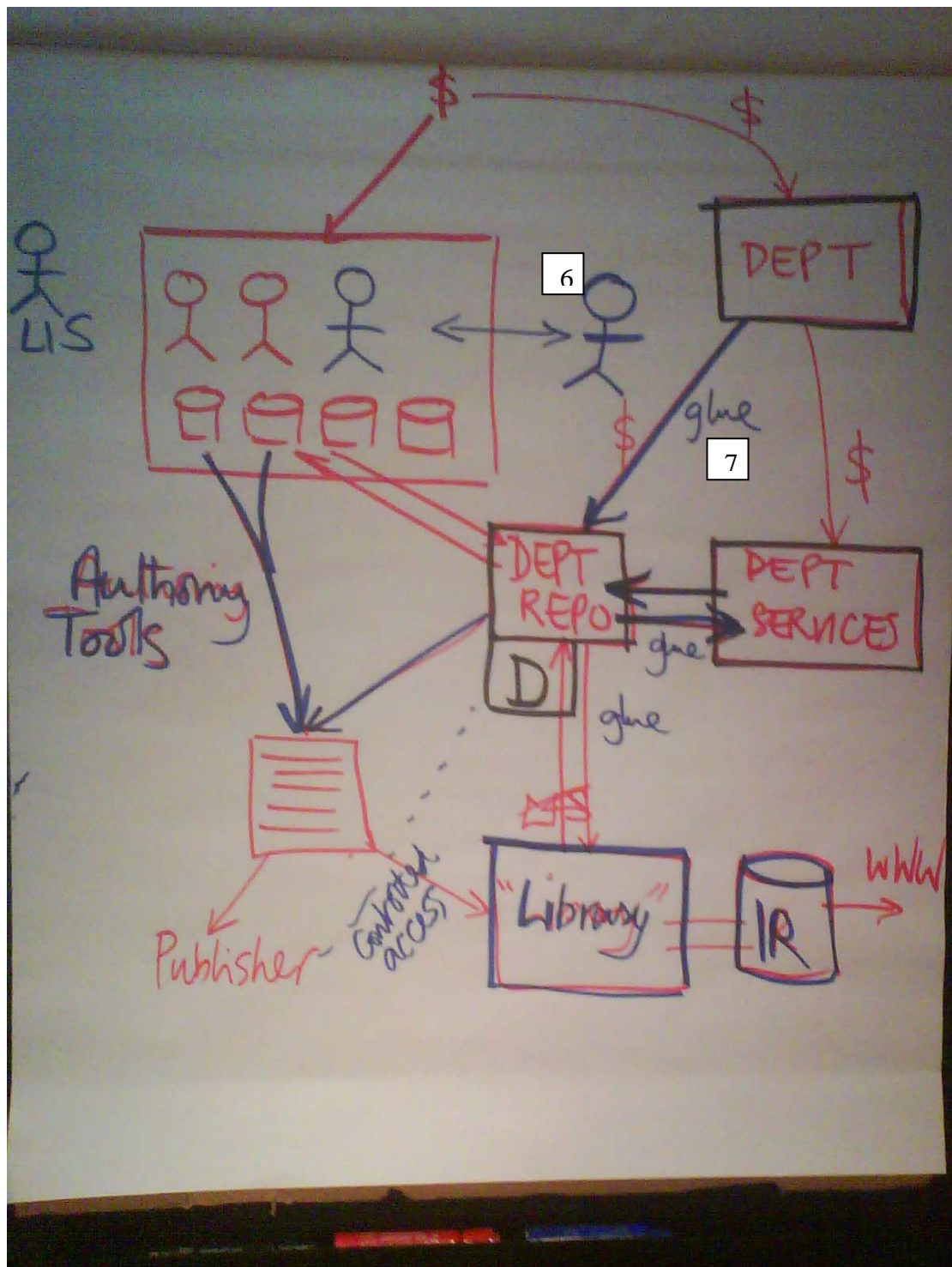
chemists (for example) are generally unaware of infrastructure issues and do not recognise their importance. They are also likely to resist 'outsiders' with little legitimacy 'telling them how to do their work'. Whereas chemists do not see the value in an improved business process, they do see value in data archiving and more exposure for their work. Their concerns can be summarised as:

- potential for data loss from current infrastructure (especially loss of one's own data, which is not uncommon)
- lack of continued access where, for example, data is held on the laptops of research students or post-doctoral researchers.
- potential loss of particular versions of papers or theses
- lack of sharing data, leading to inadvertent replication / unnecessary work being done
- poor availability and management of the tool chain (workflow) necessary to undertake research
- poor authoring environment, restricting the type and quality of reporting possible from the research

The key issue for this intervention is to review the workflow and establish when is the 'golden moment' at which data should be captured.

Putting the revised laboratory infrastructure into a broader context, gives figure 3.

Figure 3: Bench science, revised infrastructure, wider context



The new elements in this infrastructure are in blue, and represent an intervention by library / information service staff. There are two key aspects to note:

- the presence of an embedded support work, who would have information specialist skills and domain expertise, and would manage the departmental repository;

- The importance of 'glueware', provided and maintained by the local library or information service – the local element is key for the glueware to be acceptable and appropriate to local conditions.

It is of course likely that these departmental repositories should be federated in various ways to add value for researchers using them.

Case study: Humanities (Languages, Perseus Project) – led by Greg Crane.

The goals of Perseus are to produce more knowledge and enhance access to it. In doing so, it expands the research agenda. However, to achieve that requires an infrastructure that is language-neutral.

It has been hoped that institutional repositories would play a key role in underpinning Perseus, but “no movement has had more promise or has delivered less on that promise than institutional repositories”.

The key project bottleneck relates to the question, who will preserve the environment (both the infrastructure and the service model)? The attributes of Perseus are generic to modern research, and stress:

- Data rather than documents, so that key technologies here include the semantic web and FRBR¹, Canonical Text Services (enabling structured annotations), and named entity authority.
- Recombinant data, so the key challenges here are how to preserve mashups and other entities that are/were produced ‘on the fly’.
- Dynamic data, which is continually updated, and richly interlinked such that assertions are fully referenced / evidenced
- Books, readers and libraries talk to each other.

The implication from the above is that the curation task is hard. Researchers are working with probabilistic, multi-database, self-adaptive, messy, occasionally transient systems, including annotation services, and will need to know their state at any point in time, in order to be able to show the provenance of inferences that were made on that evidence base, and the degree of confidence that might be placed in the person or system that provided the evidence or interpretation on which the inference was based. It is not yet clear what can be preserved from these interactions. Having said that, Google already preserves Perseus to some extent.

The Perseus project provides a number of lessons that would be valuable for a funder to consider. Firstly, the initial investment in structured mark-up did not reap immediate rewards but has since been the basis for Perseus’s success. Secondly, Perseus does work well for non-Indo-European languages, showing that there are service patterns (such as ‘named entity service’) that are common. It may be that the statistical techniques used to infer a level of trust in those adding annotations are also common across domains.

It is possible, then, for a third party to curate a set of centralised services that would contribute to environments like Perseus. However, these would need to be defined carefully to ensure that they really were valuable across a range of disciplines. For example, specific morphological services (on language form) are variants of a general class of morphological service with a fairly

¹ Functional Requirements of Bibliographic Records

common structure, elements of which might be curated by a third party. Specific services might then be built using those as components. The basic data (images, metadata, etc) might also be curated elsewhere. Questions that then arise are (i) what would motivate environments like Perseus to delegate the curation of certain services or service components to a third party, and (ii) what organisation might be interested in being that third party? This is discussed further under 'Infrastructure models for sustainability' below.

Infrastructure models for sustainability

Data, software, services, tools and vocabularies need to be sustained over time, and increasingly with versioning, roll-back and other facilities. There are a range of business models for infrastructure, some of which were discussed at the workshop, in particular in relation to the Perseus case study.

One model might be a subscription model (cf JSTOR, Cambridge Crystallographic Data Centre), where the third party charges a subscription as its business model. The costs would be spread over a large number of users, and so subscriptions would be small. However, the disadvantage of this model is that open access to the data/services is a key success factor in the growing success of the disciplinary model exemplified by Perseus. Other disciplines, with different funding arrangements, might be able to support subscription services, although the evidence from MyExperiment (see Carole Goble's presentation at the IDC08 conference) would suggest that the open access paradigm may be widely appropriate for data services.

A second model would be for institutional libraries or information services to curate the services and make them freely available, perhaps as a result of endowment funding to ensure long-term sustainability. The immediate challenge is that libraries (and the institutional repositories they manage) typically do not have the necessary domain-specific curation skills. They may be able to curate, for example, images and metadata, leaving Perseus (or others such as Google) to build innovative services over this content.

Of course, Google is amassing a considerable collection of content itself, and is building fundamental tools such as geographic and quotation tagging over this content, and this perhaps constitutes a third model, working in parallel with purely academic initiatives.

A fourth model sees disciplinary hubs emerging as collaborations of those academic departments with an interest in a particular kind of data, in partnership with their institutional libraries. Again, an endowment funding model might work in some countries. These hubs could offer a critical mass of expertise, both from the domain and library worlds, and would be able to implement the 'glueware' noted in the 'bench science' case study above. The 'ecrystals' federation may approach this model in time. It should be noted, however, that some academic departments are increasingly aware of the commercial potential of the data they produce, so that an open access model may meet resistance in this model too. Appropriate licensing may enable data to be freely shared for non-commercial purposes.

A fifth model (or variant of the fourth perhaps) would focus more on the sites wherein much science is done and managed – the laptop. This would see lightweight federations of laptop repositories, perhaps backed up using the fourth model as a curation infrastructure.

At the technical level, a service-oriented approach provides for a separation of data (and its curation) from the services that can be built over it. This could – in principle – enable cross-subsidy between the two if, for example, revenue streams could more easily be found for one than the other. Some archives have taken this approach. However, the Web2.0 ‘perpetual beta’ model raises huge challenges for curators, in that systems are rarely stable enough to preserve in the traditional sense. It also raises challenges in terms of combining Web 2.0 services with enterprise architectures. In some sense, the platform over which the Web2.0 applications work needs to be solid, and we have perhaps yet to articulate this challenge such that it can be met effectively.

It is likely that a mixed economy will prevail, but this is different to the unco-ordinated and undeveloped picture that exists at the moment.

Advocacy and dissemination

If most researchers have to be convinced of the need for substantial investment in new infrastructure, then there is a considerable challenge ahead. Researchers are interested in research, not infrastructure, and will not easily accept ‘outsiders’ intervening. The next generation, often user-developers in the Web2.0 style, are already building their own data and workflow services, without giving too much consideration to curation. Therefore, it may be a better use of resources to try to embed curation into these services and workflows, and the instruments and sensors from which the data comes. However, this raises a different advocacy challenge, as the suppliers of many instruments and sensor systems have built proprietary formats for the data, which are hard to share and curate. In turn, these suppliers are likely only to heed demand from the (senior) researchers and managers who buy their equipment, rather than from the library community. The lesson here, then, is that advocacy for data curation needs to reflect what researchers already know they want (archiving and exposure), but may also need evidence to persuade senior managers to become its ambassadors. Such evidence could be provided by short, targeted reports.

Institutional strategy, policy and planning

There is a risk that organisational structures will not adapt to support the ambitions of researchers and research funders. The vision of interdisciplinary / trans-sector data sharing and curation may require some substantial changes to organisational arrangements, and it is not clear that this possibility is currently being considered. For example, several of the infrastructure models noted above require curation experts embedded within departments or federated disciplinary centres, and it is not clear that the leadership, management support, funding priorities, careers structures or skills needed for this approach are sufficiently well recognised at present.

Legal and ethical

As research becomes more multi-disciplinary, an increasing proportion will include data relating to human subjects, and the complexity that this introduces into data sharing and curation will rise, especially given significant variations in the ways in which existing guidelines are interpreted locally. Some approaches have been suggested and tried to reduce this complexity.

Human, cultural and partnership issues

An increasing proportion of researchers are user-developers, taking 'Web 2.0' for granted, using 'perpetual beta' services to create their scholarly record, and implicitly challenging curators to keep up with them.

In terms of incentives for researchers and departments to share data, moves toward more formal (peer reviewed) data publication could offer one means whereby good data curation practice may be recognised, rewarded and sustained. However, the reward structures for those who build tools and infrastructure are not noticeably improving, and this is a serious block on innovation. Methods need to be found to reward the development of tools, services and infrastructure, perhaps by ensuring they can be effectively cited.

There are now exemplars in terms of researchers, developers and curators working together to provide lasting services for managing and sharing research data, but this is a fragmented proto-culture. Many more opportunities are needed for cross-community conversations, across domains, disciplines, sectors, nations and so on. Such opportunities as do exist need to be better structured to ensure that lessons for strategic planners and funders are captured and can be taken forward.

It is worth noting, though, that a degree of realism will need to be maintained when considering the extent to which the infrastructure for data curation can be coordinated across these divides. Even the definition of data varies vastly between disciplines, and sector requirements diverge, so that while learning from each other is possible, deep alignment of infrastructure and tools is unrealistic and undesirable, and attempts in this direction will be resented and resisted by academic researchers. However, a good deal more social science research is needed to elaborate on these practices in specific contexts, to discover the extent to which more generic infrastructure (such as persistent identifier services, representation information registries) will align with researcher practice. The nascent Australian National Data Service has begun work in this area.

There are broad, existing initiatives in which partnerships can grow and proto-cultures be nurtured. These are broader than that represented at the workshop from which this report arose, for example the CODATA organisation.

Managing and funding research and development programmes

The workshop called for some general changes in programme planning and management, such as that reporting requirements be reviewed and made appropriate to the task. There were also voices in favour of more open calls for proposals (a 'responsive' or 'reactive' model), to enable ideas more easily to bubble up from the community, and for programmes to encourage more 'quick wins'. It is worth noting that it can be difficult to reconcile such approaches with the imperative for funders to take a strategic approach. There were also arguments in favour of fewer, longer, larger infrastructure projects.

Beyond this, some problems require a bottom-up approach and others require top down.

Bottom-up approach

In terms of research and development to create the infrastructure, it is important to put in place opportunities for disciplines and projects to learn from and influence each other. It is already clear that unexpected synergies are increasingly common across the disciplines, as more and more of them begin to see rich, annotated data and related services as being their key infrastructure. These synergies are difficult to spot in the abstract, and often only come to light as a result of practical work.

Research and development projects need to consider the implicit and explicit context around data. Often, the problems in sharing data between laboratories, disciplines, etc arise because methodologies and assumptions vary laboratory to laboratory. Breakthroughs happen when agreements are made on codifying hitherto implicit context, thereby enabling sharing. Lessons here include to work with social science expertise in eliciting implicit aspects of practice, and to work at the departmental level – academic departments have perhaps been undervalued as a unit of resource in this effort.

The recently announced NSF Datanet programme will be a useful case study in jump-starting data sharing and curation communities. Much may be learned and this learning should be shared between funders.

Top-down approach

The four or five infrastructure models described above tend to suggest a need for distributed, disciplinary or (federated) departmental repositories, around which a critical mass of curation and domain expertise can be collected, and that (for example) can show what tools can be trusted, enabling community review and feedback. Such approaches would relate to the reward structures for the people who build them.

However, it is worth noting that inter- and multi- disciplinary research will become much more the norm in the near future, as the affordances of the technology enable researchers to address previously unanswerable questions. In this world, a disciplinary approach will risk creating new silos. These new, large-scale interdisciplinary research programmes have a role to

play in promoting conversations between developers and practitioners across disciplines, as noted above, but risk the resulting infrastructure being developed without strategic direction.

There are a number of ways in which the problem space may be divided up to enable a strategic approach to be taken. For example, there may be generalised 'patterns' for curating and sharing scholarly data, perhaps for

- time-series data
- geo-coded data
- hierarchical structures
- three-dimensional volumetric data

Such families of data may offer a way of generalising, to enable tools and infrastructure to be built that might address most – but not all – of the problem space.

In such a context as this, David Giaretta (during the IDC08 conference) noted that the following may constitute common services or service domains across much of the data curation problem space:

- Persistent Identifier system, to ensure uniqueness, consistency and permanence of reference
- Registries of Representation Information – to allow material to be rendered or otherwise processed appropriately, and understood, especially where the users may be unfamiliar with the context of production
- Tools & services for creating and preserving the preservation artefacts
- Provenance, digital rights management, access control, repository information management (structure, semantics, software, etc)
- Trustability: Audit and Certification
- Cost modelling, with appropriate domain parameterisation
- Virtualisation to assist use in disciplinary software
- Some aspects of access tools

To these might be added named entity and terminology services (see Perseus case study, above), but the challenge is how to generalise these services as patterns or infrastructure so that they are easy to install, maintain and serve to/by specific communities. Building 'out' from existing research services that already have sufficient demand may be a successful approach. In addition to this range of services, an inventory of scholarly services plus standardised (but customisable) research workflows that call on them may 'prime the pump' or leverage existing successful practice.

The scale of the challenge is daunting, and there are some implications of national, project-based, unpredictable funding streams that are unhelpful in meeting it. Projects take time to establish, multidisciplinary teams take time to establish a 'pidgin' language by which to communicate, user buy-in is hard to gain for short-term projects, etc. Further collaboration between national funders could enable a more strategic and longer term R&D approach to be pursued, although this can risk alienating national funders from their local constituencies.

Specific issues concerning software development

There is also a need to focus on the software development process, in an attempt to get tools and infrastructure more widely adopted. Challenges include the 'not invented here' syndrome, and software products developed by amateurs who may not either harden the software for widespread use, and productise it to enable it easily to be useable and acceptable in a wide range of local contexts. Funding mechanisms at the moment rarely allow for these business processes. One approach might be to fund national centres to host the software developers, and cycle the projects in and out.

At the other end of the development process, it is important to get pre-alpha versions of new tools in front of users early, to get their input before the tool gets too fixed. There is a tendency for developers to worry about presenting early versions to users, and this needs to be challenged.

In designing programmes that include software development, it is important to be clear what the aims are, for example, research and development, proof of concept, or community adoption / roll-out?

Conclusions and specific suggestions

1. A key point at which potential value is lost is at initial data capture. Laboratory equipment, medical equipment, sensors and arrays, cameras, scanning equipment, etc all capture data from the physical world and, for the most part, convert it into digital information in closed, proprietary formats. However, concerted effort will be required to get the attention of commercial suppliers on this issue. A survey might be commissioned of the best opportunities for integrating this kind of equipment into an infrastructure based on open standards, protocols and formats, and the heads of major research funders (as representatives of the research community) might then collectively use this evidence to contact relevant equipment suppliers to discuss this issue.
2. A study could usefully be done into how curator expertise might be acceptably embedded into disciplinary practice. The potential role for academic departments was emphasised at several points in the discussion, and these organisations / cultures are probably the primary context for such embedding. The biological sciences have made considerable progress defining skill sets and career structures for such embedded curators, and this now needs urgently to be rolled out. Such curators are likely to be domain specialists with information skills, rather than vice versa.
3. Arrangements need to be put in place across stakeholders to enable the recommendations from (2) to be implemented rapidly. Lessons should be learned from existing models, such as that used by the National Endowment for the Humanities.
4. The role of information specialists may evolve, perhaps becoming more research focused, but this is not yet clear and there is an urgent need to articulate the relationship between domain and information expertise

in support of data curation in various contexts, or under various infrastructure models or patterns.

5. Work could usefully be done systematically to bring together relevant guidance and practice relating to legal and ethical issues in data covering human subjects, across disciplines, research funders, jurisdictions and campuses. This work should suggest ways to reduce the potential complexity in dealing with such issues.
6. Structured opportunities should be provided to enable data curators to interact across existing divides, and for any lessons from such interactions that are relevant to strategic planners and funders to be captured and directed appropriately.
7. CODATA should be exploited as an international forum in which to coordinate work across countries, such as the conversations identified in (6).
8. Work is required to develop and then test, validate and embed infrastructure services (such as persistent identifier services, access management services) in real researcher workflows.
9. Considerable benefits could flow if funders and planners could work internationally to ensure a strategic approach is taken; the scale of the challenge is beyond any national resource. Mechanisms need to be developed to make this international coordination easier.

Annex 1

List of attendees

George	Alter	Associate Director	ICPSR
Brett	Bobley	Director, Digital Humanities Initiative	NEH
Christine	Borgman	Professor & Presidential Chair	Dept of Information Studies
Sayeed	Choudhury	Associate Dean Manager, UK National Crystallography Service	Johns Hopkins University
Simon	Coles		University of Southampton
Greg	Crane	Professor of Classics	Tufts University
Sigrun	Eckelmann	Program Director	Deutsche Forschungsgemeinschaft
Neil	Fraistat	Professor of English & Director, MITH	University of Maryland
Chris	Greer	Program Director	NSF (sent apologies)
Robert	Hanisch	Senior Scientist	Space Telescope Science Institute
Neil	Jacobs	Programme Manager	JISC
David	Kirsch	Assoc. Professor	University of Maryland
Carl	Lagoze	Senior Research Associate	Cornell Information Science
Clifford	Lynch	Executive Director	CNI
Liz	Lyon	Director	UKOLN
David	Millman	Sr. Director, Information Technology	Columbia University
	Murray	Unilever Centre for Molecular Sciences	
Peter	Rust	Informatics Head of Research and Development Department	University of Cambridge Goettingen State and University Library (SUB)
Heike	Neuroth	Associate Professor / Director, Library Research Center	University of Illinois at Urbana- Champaign - GSLIS
Carole	Palmer		National Endowment for the Humanities
Jason	Rhody	Senior Program Officer	NIBIB
Belinda	Seto	Deputy Director Director and Chief Architect, ARCHER Project	Monash University
Andrew	Treloar		
Marjan	Gerritsen	Programme Manager ICT & Research Program Officer for Scholarly Communications	SURF Foundation
Don	Waters	Communication and Information	Andrew W Mellon Foundation
Astrid	Wissenburg	Director	Economic and Social Research Council

Annex 2

Summaries of attendees' interests in the topic of the workshop.

George Alter ICPSR, University of Michigan

The Inter-university Consortium for Political and Social Research (ICPSR) was established in 1962 to archive and disseminate data to the social sciences and to provide training in research methods and statistics. We serve more than 500 member institutions, and users can download more than 6,000 studies from all of the social sciences. ICPSR has been very active in developing standards for curation, preservation, and sharing metadata, such as the Data Documentation Initiative (DDI). We are currently facing new challenges such as the preservation and dissemination of confidential data, providing access to data for on-line analysis, and preserving and disseminating video, audio, and other types of digital objects. We are also discussing ways to provide easier access to complex databases, such as life histories, and to create communities of users who can share their expertise about large and complex studies.

Brett Bobley and Jason Rhody NEH:

At the National Endowment for the Humanities (NEH), we fund research in all disciplines of the humanities (e.g. language, both modern and classical; linguistics; literature; history; jurisprudence; philosophy; archaeology; comparative religion; ethics; the history, criticism and theory of the arts; and those aspects of social sciences which have humanistic content and employ humanistic methods). One of the challenges in humanities research is overcoming the traditional approach of the "solitary scholar" doing his research in isolation. We feel that for humanities research involving infrastructure, humanities researchers need to embrace a different approach, perhaps one more akin with the sciences. That is, working in a team-based environment; being open about methods, results, and lessons learned; and sharing data and tools in an open fashion. This of course brings up the larger issue of HOW to share and WHAT to share. The NEH has been in discussions over the past few months with the US National Science Foundation and the US Institute of Museum and Library Services about ways in which we can facilitate the proper curation and sharing of humanities tools, methods, and data. We hope to bring other funders like Mellon and JISC into this conversation as well. We look forward to learning more about how other organizations are facing these issues.

Christine L. Borgman, UCLA

We have been conducting research on data management in the Center for Embedded Networked Sensing, a National Science Foundation Science and Technology Center, since 2002 [1-9]. The set of research questions from our interview study [7] directly address the concerns of this workshop:

- Data characteristics
 - What data are being generated?

- To whom are these data?
- To whom are these data useful?
- Data sharing
 - When will scientists share data?
 - With whom will they share data?
 - What are the criteria for sharing?
 - Who can authorize sharing?
- Data policy
 - What are fair policies for providing access?
 - What controls, usage constraints, or other limitations are needed?
 - What publication models are appropriate?
- Data architecture
 - What research design tools are needed?
 - What data acquisition tools are needed?
 - What data analysis tools are needed?
 - What data publishing tools are needed?
 - What data models do scientists need to generate or use the data?

The full range of questions is listed here to reflect the need to examine data sharing within a larger context of what data are being generated, to whom are they useful, and what policies and architecture are required to support sharing. We found that data in this research center take many forms, ranging from sensor readings to soil and water samples, and exist in many states of processing and verification. Often the useful data are derived from samples that are destroyed in the analysis process (and thus not available for sharing, per se). The data are stored on the computers of the individuals who collected them, on paper, on shelves, in refrigerators, and occasionally in shared repositories. While our science and engineering researchers are supportive, in principle, of sharing their data, very few of these data actually are reused by anyone other than the investigators who collected them. The reasons vary widely. Participating faculty students and staff have not determined (and often not discussed) who are the “owners” of a given dataset and who has authority to release them. The conditions under which researchers are willing to release data vary by the state of the data, the requestor, and timing, in all possible combinations. Our scientists and engineers have the best of intentions with regard to data sharing, but their practices remain driven by the long-standing ethic that the publication is the end product of scholarship, and the data are an interim product for local use.

The incentives for researchers to share their data derive mostly from principles of open science: scholarship must be made available for public critique and assessment. However, scholars compete as well as collaborate, and the academic reward system is a much stronger driver of scholarly practices than are technology or public policy. In my new book [10], I explore in depth the disincentives to share scholarly information in the sciences, social sciences, and humanities. These disincentives are divided into four categories: (1) rewards for publication rather than for data management; (2) the amount of effort required in documenting data for use by others; (3) concerns for priority, including the rights to control the results or sources until the publication of research; and (4) intellectual property issues, both the control and ownership of one’s own data as well as access to data controlled or owned by others. All four of these obstacles must be addressed in programs for sharing and curating data. “Carrots,” in the form of strengthening the incentives and weakening the disincentives, are more likely to be effective than “sticks” in the form

of strict requirements to deposit data. Deposit requirements can be met too easily by providing unreliable and undocumented datasets that are of little use to anyone – and expensive to curate.

1. Mayernik, M.S., Wallis, J.C., and Borgman, C.L. (2007). *Adding Context to Content: The CENS Deployment Center*, American Society for Information Science & Technology. Milwaukee, WI. 44 Information Today.
2. Borgman, C.L., Wallis, J.C., and Enyedy, N. (2007). *Little Science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries*. International Journal on Digital Libraries. <http://www.springerlink.com/content/f7580437800m367m/> Visited: 29 September 2007
3. Borgman, C.L., Wallis, J.C., Mayernik, M., and Pepe, A. (2007). *Drowning in Data: Digital Library Architecture to Support Scientists' Use of Embedded Sensor Networks*, JCDL '07: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries. Vancouver, BC Association for Computing Machinery. 269 - 277.
4. Pepe, A., Borgman, C.L., Wallis, J.C., and Mayernik, M.S. (2007). *Knitting a fabric of sensor data and literature*, Information Processing in Sensor Networks. Cambridge, MA Association for Computing Machinery/IEEE.
5. Wallis, J.C., Borgman, C.L., Mayernik, M., Pepe, A., Ramanathan, N., and Hansen, M. (2007). *Know Thy Sensor: Trust, Data Quality, and Data Integrity in Scientific Digital Libraries*. In L. Kovacs, N. Fuhr, and C. Meghini, Editors, *11th European Conference on Digital Libraries*. Budapest, Hungary. LNCS 4675 Berlin: Springer. 380-391.
6. Wallis, J.C., Milojevic, S., Borgman, C.L., and Sandoval, W.A. (2006). *The special case of scientific data sharing with education*. In A. Grove, Editor, *American Society for Information Science & Technology*. Austin, TX. 43 Information Today.
7. Borgman, C.L., Wallis, J.C., and Enyedy, N. (2006). *Building digital libraries for scientific data: An exploratory study of data practices in habitat ecology*. In J. Gonzalo, C. Thanos, M.F. Verdejo, and R.C. Carrasco, Editors, *10th European Conference on Digital Libraries*. Alicante, Spain. LNCS 4172 Berlin: Springer. 170-183.
8. Borgman, C.L., Wallis, J.C., Pepe, A., and Mayernik, M.S. (2006). *Personal Digital Libraries and Scientific Data*, American Society for Information Science and Technology. Austin, TX.
9. Shankar, K. (2003). Scientific data archiving: the state of the art in information, data, and metadata management. <http://cens.ucla.edu/Education/index.html> Visited: 19 January 2005
10. Borgman, C.L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Simon Coles
University of Southampton

There is an immediate requirement for raising awareness in the research community so that scientists realise the value and benefits in archival of digital research data at source. This not only involves providing the necessary tools, but also a solid infrastructure in which they trust. At the same time it is imperative that funding enables exploration of new models for institutional preservation and the role of centralised facilities, such as the library and information services, in the new information environment we find ourselves in. In most areas of science the primary

driver for adopting this approach is the publication system and it is necessary to engage communities and learned societies to develop and move this agenda forward.

Dr. Sigrun Eckelmann, DFG - DFG Views on access to primary research data

The DFG (German Research Foundation) is the central self-governing organisation responsible for promoting research in Germany. According to its statutes, the DFG serves all branches of science and the humanities. The DFG supports and coordinates research projects in all scientific disciplines, in particular in the areas of basic and applied research. The DFG also funds and initiates measures to promote scientific libraries and equips computer centres with computing hardware. This is the assignment of the group DFG/LIS.

In summer 2006 DFG/LIS published its funding priorities through 2015 identifying 17 main fields of activities. One of these measures deals with primary data:

„Access to primary research data should be enabled through the development of automated indexing and retrieval techniques, as well as of referencing and availability models for primary research data, in systematic cooperation between information service and data centres. Discipline-specific and object-type-specific repository structures for primary research data should be created. For primary data in the humanities, a nationally coordinated infrastructure will be necessary to ensure international visibility. Studies, workgroups and workshops should accompany these measures.

Access to primary research data. Funding of systematic collaborations between information service centres (primarily libraries) and scientific data centres or research-data-producing projects should focus on two aspects. On the one hand, the development of discipline-specific (e.g. economics) or object-type-specific (e.g. chemical formulas) indexing and retrieval techniques should be promoted to optimise access to pertinent data. On the other hand, models for referencing and availability of primary research data should be developed. Each pilot project providing access to primary research data should be a collaborative effort between one information service centre and one or more scientific data centres.

Complementary workgroups and workshops help to develop technologies, standards, processes etc. They should be planned and organised in collaboration with the German Initiative for Network Information (DINI) and with the Knowledge Exchange partners. Periodic studies will evaluate strategic funding objectives, technologies, standards and usage scenarios.

Development of repository structures for primary research data. Like institutional and discipline-specific document repositories, which store and provide mostly publications (i.e. secondary literature), repository structures for primary research data also merit their own line of action. Funding under this line of action should go to object- or project-specific approaches that pilot the development and networking of repositories for specific types of primary research data.“²

Beyond these strategic discussions DFG/LIS is already funding two projects. From 2003 to 2006 DFG/LIS has funded the project „Publication and Citation of primary research data“, a project by Max Planck Institute for Climate in Hamburg in

cooperation with other institutions, later it has been followed by the Central Technical Library Hannover.

The essential results of this project are: several institutions (agents) agreed on standards and structures for gathering and storing the reviewed data as well as for the metadata (DOI). The metadata are joined to the catalogue of the library. By this research data are reviewed, stored, can be identified via a DOI as persistent identifier, and can be cited like a normal publication.

This structure has been developed by researchers and librarians in order to build an organisational structure to assure storage and availability of research data. Furtheron it is aimed to motivate the researcher in publishing their research data. So, this is a bottom up approach developed for the geosciences (climate, sea, atmosphere). It is a structure beyond commercial interests.

Experience from the last months shows that this structure seems to be well accepted by reseachers also from other sciences. Life sciences, however, may ask for other structures, humanities as well.

Another project (University Trier, ZPID) funded by the DFG/LIS deals with primary data of psychology.

My personal view is that in the future we will have to deal with several discipline-specific structures/organisations like the above mentioned. The challenge will be to find an 'umbrella' for them also on the interntional level. Success of potential future organisations and structures above all depends on the motivation researchers to publish their data and on the international integration.

Bob Hanisch

Senior Scientist

Space Telescope Science Institute

- Data and data expertise is physically distributed. Any significant digital data management and curation facility must be inherently distributed, yet logically connected. Distributed data resources means distributed human resources, so management and coordination is complicated.
Participants in the project may not work for the formal management team directly, and may have split responsibilities at their home organizations. Initial enthusiasm and cooperation can, in time, give way to complacency and resistance as the problems get harder or the work turns from prototyping to operations.
 - o Data come from diverse sources, with diverse and inconsistent levels of metadata. Data quality itself varies, and may or may not be documented in the metadata. Data and metadata curation is a substantial effort, and must be planned for and undertaken early in the data lifecycle so that expertise is not lost. Incentives need to be identified for scientists to provide high quality data and metadata.
 - o Robust systems for distributed data/metadata management rely upon

standards, and these standards must have buy-in from their scientific communities. The standards process is both national and international, and takes time, coordination, cooperation, and compromise. Striking the right balance between the patience to get it just right and the urge to just do it and move on is difficult. Revising standards too frequently risks alienating data service providers, who cannot afford to make too-frequent changes to their software systems. Waiting too long to release a standard causes providers to lose interest or make up their own solutions.

o Systems for managing research data need to be designed to meet the needs and expectations of the end-user research community. That community, however, is mostly interested in their here-and-now research activities. It is hard to get them to define requirements. They will tell you what's wrong with what you've developed, but not tell you what they want because they don't know until they see it. Engaging the research community in a meaningful way is hard, at least until you have something of substance to show them.

Neil Jacobs JISC

As research becomes more data-centric, the need for a UK academic data infrastructure is growing. A recent UK national report noted that the requirements for a national e-infrastructure are broad, but crucially that the “requirements presuppose not only a high level of integration and coordination, but also, in key areas, intervention at the policy level.” [OSI eInfrastructure report]. One of JISC’s key strategic objectives for 2007 – 2009 is promoting the development, uptake and effective use of ICT to support research...” Our vision is to support the wide spectrum of research from mono to inter-disciplinary and from solo to large-scale research across all disciplines.

There are a number of challenges to tackle, if a vision of an academic research data infrastructure is to be agreed and realised. Researchers require an infrastructure to support the entire research lifecycle. There are a number of barriers to sharing data (including: confusion regarding intellectual property; poor metadata; unclear provenance/context; mistrust). However, there are a number of drivers: the increased emphasis on collaboration and the developing semantic web/grid technologies (for example, enabling discovery, annotation) and ease of use by researchers, for example. There are also significant differences across the disciplines, which need to be considered. Institutions are looking for sustainable solutions. Current concerns include: storage capacity; developing curation skills in the workforce; data management planning. Institutions vary widely in terms of data management policies and a clearer picture is required to inform future work.

At the national and international level, there is a general consensus that data curation is important, if we are to encourage use, reuse and repurposing. However, there is a current lack of analysis of the value proposition at institutional level and in a number of cases, at the subject and discipline level. JISC needs to help institutions fully articulate a business case for effective data management and to plan activities as a

priority. The recent report by Dr Liz Lyon, *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, concluded:

“...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 [...] to put in place the infrastructure and services to effectively manage the burgeoning data deluge.”

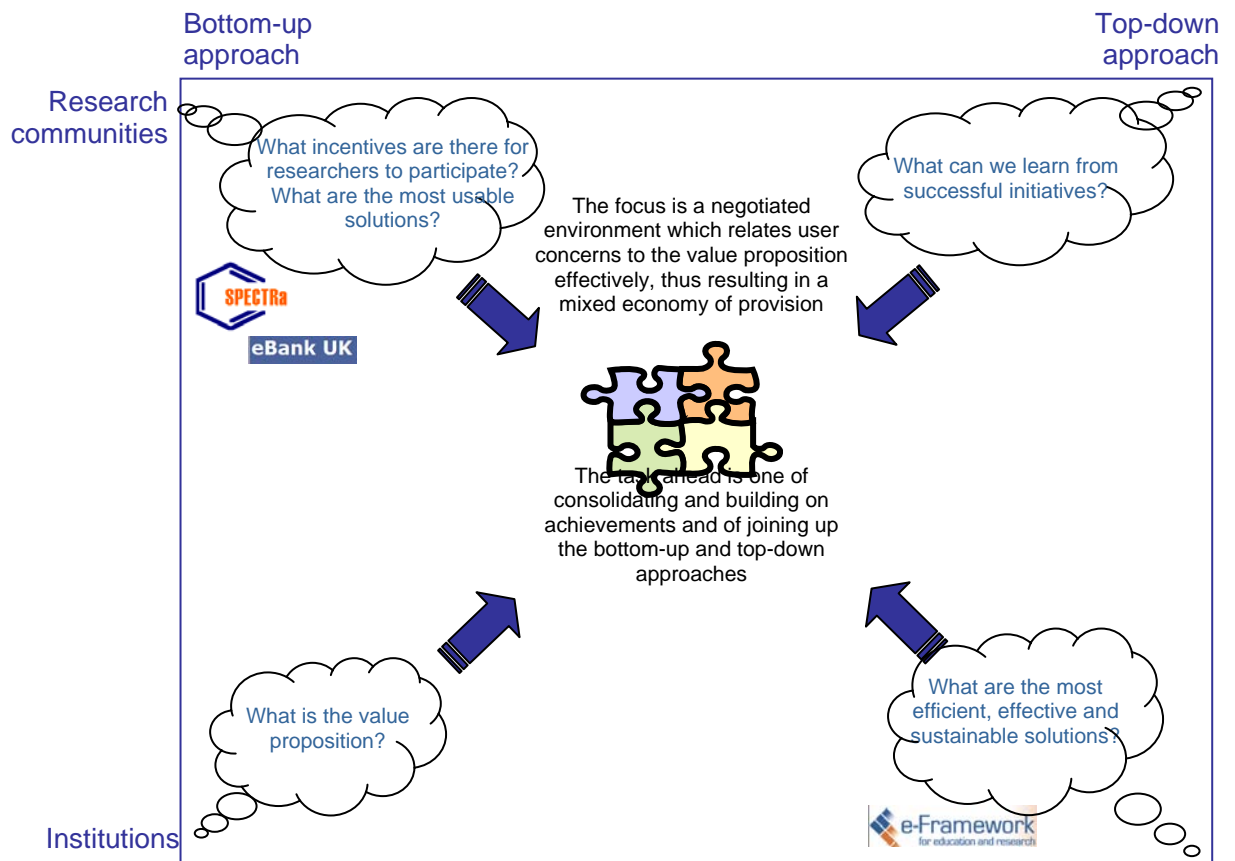
The curation and reuse of research data covers such a range of issues, that a number of organisations and groups are exploring this space. A number of reports, either completed or planned, have resulted in long lists of recommendations, many of which fall within JISC’s remit. There has not always been good co-ordination in the commissioning of many of these reports and it is important to work towards greater co-ordination in planning future activities. There is a risk that without concrete joined-up planning, the UK will fall behind other countries (e.g. US, Australia) in developing and sustaining a national data infrastructure.

JISC has commissioned a range of activities and projects relating to research data, which provide a platform on which to build a coordinated programme. Some of the work to date has taken a “bottom-up” approach, for example, the eBank³ and SPECTRA⁴ projects within the crystallography community; whereas other work has taken more of a “top-down” approach such as the eFramework programme⁵. A programme of work is being planned by JISC to elaborate and make progress toward in terms of consolidating and building upon the existing work. The task ahead is one of consolidating and building on achievements and of joining up the bottom-up and top-down approaches, the diagram below illustrates the current framework for this task.

³ eBank: <http://www.ukoln.ac.uk/projects/ebank-uk/>

⁴ SPECTRA: <http://www.lib.cam.ac.uk/spectra/>

⁵ E-Framework for Education and Research: <http://www.e-framework.org/>



David Millman

I am interested in both local motivation and global provenance issues. First, while local organizational structures routinely support both research and infrastructure, they often do so through entirely independent internal governance and funding processes. Specifically, "infrastructure research" itself often falls into neither area of support. Efforts such as NSF's Cyberinfrastructure programs will make good headway to cross these lines, but to the extent their immediate audiences are researchers in particular domain areas, institution-wide attention may require supplemental strategies.

Second, authenticity in provenance depends on a fabric of trust that is consistent across repositories. What policy, identity management, or other frameworks are necessary to insure this? Are repositories being designed with these interdependencies in mind? What lessons from virtual organization infrastructure research or from federated access management experiences would be useful to other institutions, research collaboratives and funding agencies?

Peter Murray Rust
University of Cambridge

The primary challenge is persuading scientists that their research data is both costly and valuable. The primary outcome of most funded research is papers which are often prepared with great effort, but data comes a poor second. To be fair, it's more difficult, particularly after the event. I am particularly interested in data collected in smallish projects and here we could benefit enormously from domain-specific authoring tools and a tradition of archive-as-created. This is common in software projects where unit-testing and common repositories such as subversion (SVN) work well as a complete record and preservation of the program. A similar approach could work for many common types of data particularly when captured by instruments (as much is). There needs to be communal will and this could come through learned societies and international unions (as in bioscience and crystallography). A tandem approach could include stronger requirements from institutions that graduate students archive data with their theses. It's good practice and will lead to higher exposure for the institution. Funders also have an important part to play - there is great emphasis on "Open Access" publication financed by funders but much less on Open Data.

Dr Andrew Treloar,
MACS PCP

I am currently the Technical Architect for the Australian Research Repositories Online to the World (ARROW - <http://arrow.edu.au>) project, and the Director and Chief Architect for the Australian Research Enabling Environment (ARCHER - <http://archer.edu.au/>) project. ARROW is expanding out from a document-centric 'traditional' institutional repository focus to start including large (multi-GB) dataset objects. ARCHER has been focussing on support for data-centric collaboration, as well as instrument integration, and has been working with ARROW on smoothing the migration of objects from a collaboration space to a publication space. This transition is the subject of an article I co-wrote for DLib (The Data Curation Continuum: managing data objects in institutional repositories - <http://www.dlib.org/dlib/september07/treloar/09treloar.html>). I am also leading a project, hosted at Monash University, to establish the Australian National Data Service (ANDS). ANDS is described in [_Towards the Australian Data Commons_](http://www.pfc.org.au/twiki/pub/Main/Data/TowardstheAustralianDataCommons.pdf), available online at <http://www.pfc.org.au/twiki/pub/Main/Data/TowardstheAustralianDataCommons.pdf>.

Research Data Infrastructure Challenges

In my view the main challenges in providing research data infrastructure are both technical and social. In the technical sphere, we need to make doing the right thing easy. That is, we need to do a much better job of making it as easy as possible for the creators of data to deposit data with a long-term future into well-curated institutionally-supported repositories. This should include automatic capture of the required metadata (descriptive, provenance, preservation) at the appropriate stages in the process. In the social realm, we need to make doing the right thing attractive. This means ensuring that the discipline, institution and funding reward structures and processes encourage data management and curation behaviours that will benefit not just the individual researchers, but also their colleagues and scholarship as a whole.

Research Data Infrastructure R&D Challenges

In order to address the challenges identified above, we need to undertake R&D. The main ***R&D*** challenges are trying to work out what things are discipline-specific, and what can be generalised. This is particularly important in a constrained funding environment where we can't afford to build lots of bespoke solutions. This leads to a related R&D challenge - at what point does a generalised solution become effectively useless for any specific discipline or researcher? Of course, we also have the problem of how we can deal effectively with sustainability in programmes made up of transient projects, but this is more of a funding/governance issue rather than R&D. The Open Source community has much to teach us

--

Sharing and curating research data⁶

Challenges in defining effective research and development programmes that will both result in broadly usable infrastructure and address subdiscipline-specific requirements.

Neil Jacobs (JISC) and Don Waters (Mellon Foundation)

Introduction

The knowledge economy is an important basis for wealth creation and improved quality of life in industrialised countries, and it depends on the exploitation of scientific and scholarly research. Investment in science and scholarship is rising, and the technological infrastructure required to manage its outputs – especially research data – is growing larger and more complex. However, it is not clear how best to design the research and development that is required to scope and build this infrastructure into the future.

Background

There have been a number of project-level initiatives that have sought to design and pilot infrastructure to enable research data in the sciences, social sciences, and humanities to be effectively captured, shared and curated. Two overarching approaches have been apparent:

1. **Top-down:** In the technical arena, this approach is manifest in such developments as the eFramework⁷, which has sought to define a modular, service-oriented development methodology. In the legal environment, it is evident in such developments as Creative Commons, whereby a license regime comprising a very limited set of variables offers a range of licensing options. Systems developed from the top down in these ways are supposed to enable the development of compatible applications that meet locally derived requirements, but theory often diverges considerably from practice.
2. **Bottom-up:** This approach focuses on addressing infrastructure needs from the perspective of practitioners in specific fields or subfields of study. Projects have been undertaken in the areas of crystallography, environmental science, climate research, astronomy, medieval studies, literary studies, archaeology, and others in the UK, Australia, and the US with funding from both government and private sources (JISC, NSF, Mellon, and so on). However, it is often not easy to ensure that elements developed from the bottom up have broad applicability and use beyond the parameters of the particular projects for which they were conceived.

⁶ ‘Research data’ here means anything analysed by academics in pursuit of science, research or scholarship.

⁷ EFramework for education and research: <http://www.e-framework.org/>

Funders in various countries are now considering how to design programmes for infrastructure research and development that use these two approaches in a coordinated way such that they maximize the advantages and limit the disadvantages. An optimal design would effectively:

1. Deploy a generalized and scalable infrastructure that facilitates application development in a variety of fields or subfields of scholarly activity;
2. Identify good practice requirements within specific fields or subfields;
3. Provide socio-technical solutions (technical, legal, etc) that both conform to the infrastructure and meet field or subfield requirements in a testable way;
4. Deliver versions of those solutions that can be used to assess whether they meet requirements in other fields or subfields;
5. Offer a framework for systematic feedback in which development of the broader infrastructure could be informed by local needs and applications and local needs and applications could benefit from investment in the broader infrastructure.

How can R&D programmes be designed to achieve these goals? There is little documented evidence. This international, invitation-only workshop is intended to help pool expertise in this area, bringing together both practising academics and programme planners (such as Mellon and JISC), to answer this question. The workshop will draw from case studies, each of which will be documented from both an academic and a programme planner perspective.

Intended outcomes

The workshop is intended to:

1. Identify and document a small number of case studies to illustrate key issues that arise in planning and implementing infrastructure programmes related to the curation of research data
2. Enable and capture a discussion of these and other relevant issues among an invited group of experts
3. Develop recommendations, based on analysis of the case studies and experts' experience, on how to align the discipline-specific and infrastructural demands of programmes concerned with research data

Benefits

If successful, this workshop and any follow-up work will provide an evidence base and examples of effective practice on which programme planners can draw to design better R&D programmes relating to the curation of research data.