

Report on RIN Workshop: Getting the most out of data, making the most of research

Tuesday 5 December, Royal Institute of Public Health, London

An Attendee's View from John MacColl, Head of Edinburgh University's Digital Library

John Wood, the Chief Executive of the Council for the Central Laboratory of the Research Councils (CCLRC) was the first name on the programme for this meeting, but due to illness he was replaced by a colleague, Juan Bicarregui, who gave a presentation entitled *The purpose and place of research data in the digital age*. He discussed data management at the CCLRC (which is about to merge with Particle Physics and Astronomy Research Council to become the 'Science and Technology Facilities Council'), how it supports the research lifecycle, and then addressed policy issues. He illustrated the vulnerability of data by beginning with images of Mount Etna erupting – a one-off event, so that if the data is lost it cannot be recreated. The scale of eScience was also illustrated. The ATLAS Project at CERN now has 2,000 scientists from 150 universities in 30 countries. It runs equipment which generates masses of data per second. CCLRC itself hosts a large number of repositories, and has data storage planned out to 2010, by which time it will be storing 20 Pb.

He remarked that we collect a lot of data today, not knowing why it might be useful later (e.g. Antarctic environmental data was collected long before scientists began looking for an ozone hole). The research lifecycle has a body of knowledge at its heart, which supports the Government's research agenda, as well as academic research. Researchers act through depositing in and access to that body of knowledge – working within a 'Virtual Research Environment', whether formalised or not. The researcher should not have to worry about the information infrastructure ('the Grid dimension') or the curation dimension. Support services are required to do that.

The UK Research Councils' initiative on access to research outputs is expressed in statements issued in June 2005 and 2006. There were four principles initiated in 2005:

- Ideas and knowledge derived from publicly-funded research are made available and accessible for public use, interrogation, and scrutiny, as widely, rapidly, and effectively as practicable
- Effective mechanisms are in place to ensure that published research outputs are subject to rigorous quality assurance, through peer review
- The models and mechanisms for publication and access to research results are both efficient and cost-effective in the use of public funds
- The outputs from current and future research can be preserved and remain accessible not only for the next few years but for future generations

In 2006 the statement (though not uniform across the Research Councils) was that deposit should be mandatory, and this has been followed up by a declaration that it should include data. The OECD Declaration (2004) produced 10 principles, the first of which is openness, and states the need for 'balancing the interests of open access to data to increase the quality and efficiency of research and innovation with the need for restriction of access in some instances to protect social, scientific and economic interests'.

The OSI e-Infrastructure Steering Group recently came out with five key findings:

- the need to promote data throughout its lifecycle
- the need to reduce the cycle time
- the need for greater use of simulation-based research
- the need to reuse data
- the need for standardisation

There is a need for persistent unique identifiers to enable global cross-referencing; also for full metadata; for data agents (data has to be 'intelligent' as to its use); for the ability for data to work with simulations; and for archiving (in a 'data forge'). Proper curation of data throughout the research lifecycle should eventually mean that electronic laboratory notebooks will be traceable back from the output. There are many sources of data which could be repurposed and used to support research, even though we might not yet think of them in that way (e.g. data from supermarkets). The e-infrastructure will also support public engagement. Standards are required in the areas of classification, open access, data citation, e-learning, software integration, and propagating best practice. There is a lot of data generated also by small-scale research projects (a 'long tail') which has no less need of management - but this will require a major culture change in the working practices of scientists and researchers.

Chris Rusbridge, Director of the Digital Curation Centre, then spoke on *Data creation and curation in research*. He addressed the question of data citations. Data is increasingly important as evidence. The basis of science is the verifiability of previous work by experiments. It is particularly important to capture unrepeatable observations and experiments, and there are also legal obligations motivating this activity. To curate scientific data, one needs to think about creation of the resource, its development, its acquisition, metadata and preservation. Observations made need to be fixed. The data which is important is not necessarily that which is observed, however. Evidence suggests that data which has been derived, refined, combined and processed is more highly valued. He gave an example of how two minutes of comparative data work at the Virtual Observatory at Johns Hopkins was sufficient recently to allow the declaration of the discovery of a new brown dwarf - an activity which would have taken much longer in pre-digital data days.

Data is meaningless without its context. It is important to capture metadata automatically from the workflow as much as possible. We need to record the 'computational lineage' of data as it is captured, processed by various research groups and eventually passed to policy makers who make decisions on the basis of it. On the question of access and re-use, there are ethical and rights control restrictions on re-use (especially in social sciences). We need new expressions of these rights for the long-term. We also need collaboration tools to allow annotation, discussion and review. With publication, an interesting question is 'when is data published?' (i.e. not just when an output appears in print). The use of citation to publish the underlying data is key here. Commonly the date alone is used (e.g. 'Retrieved on 8 Jan 2006'). But when making a citation one needs to be able to allow the reader to revisit the cited data *as it was on the date of citation*. Peter Buneman (DCC Director of Research) is working on a way to reference and access 'archived' past states of a changing dataset. This is not important for original observation, and not too important for incremental datasets, but it is very important for revisable datasets (e.g. genomics or geographical boundary data).

Chris Rusbridge then picked up the point which Juan Bicarregui had made about the different cultures which exist between small and big science: 'Small Science contains two to three times more data than Big Science, but it is much more at risk'. It may be controlled by the PhD students, a Research Assistant, Principal Investigator or Administrator. It is likely to be stored on local hard drives or at best shared network drives, and policy-led deletion gets rid of too much. The individuals concerned generally do not keep documentation and metadata adequately. As an example of an approach which addressed these deficiencies, he mentioned

the *eCrystals* project at Southampton, where a partnership with the institutional repository is archiving the data. Cambridge also ingests datasets into an institutional repository, but mixes them (e.g. crystallographic datasets with archaeological datasets), and is lacking in skilled curation. Only a tiny number of UK libraries claim to include datasets in their institutional repositories. The California Digital Library does so, but again from a document perspective only.

Citing some models which were broader than the institutional, Chris Rusbridge said that the British Library is taking a serious and robust approach, though one which is broadly oriented towards cultural heritage. The National Archives provides an archive for government datasets. OCLC has an archive which is demand-driven, but purely on a market response model. Portico does archiving in the specific domain of e-journals, and Iron Mountain does the same in the records management domain. The British Atmospheric Data Centre believes that institutional repositories are the wrong model, and that domain scientists are required to do curation, but Chris Rusbridge believes that we need institutions to step forward and take up a role here based on their fundamental sustainability. There needs to be a new collaboration between disciplines and institutions. He quoted researcher behaviour survey findings from Project *StORe*, suggesting that institutions were not yet creating environments in which researchers could take a responsible approach to data curation.

There followed a number of parallel sessions providing perspectives from different research areas. I attended the presentation by David Shotton of the Image Bioinformatics Research Group at the University of Oxford. Biomedical research data has a bottom-up data flow, in which there is a very large research community, highly distributed research activities, and the data is heterogeneous and largely unstructured. It is an open world – very different from physics and chemistry. In medical research, the one enormous difference from other domains in respect of data management is that of patient confidentiality. The *BIRN* project in the US (Biomedical Information Research Network) is establishing a cyberinfrastructure for confidentiality. He also cited the *CHERRI* project at Edinburgh, which has defined licence conditions for data reuse. The *eDIAMOND* Digital Mammography eScience Project found that patient confidentiality was much more difficult than the technology. All of this makes data sharing more problematic in medicine than in most domains.

Looking at scientific communication more philosophically, Dr Shotton argued that a scientific paper is an exercise in rhetoric - it does not simply report scientific observations. The goal is to convince the reader. They are also published for other reasons: to secure the next grant, to do well in the RAE, etc. Sadly, the bulk of our raw research observations, which could provide a lot of support in this task of persuasion, are never published. Historically there were reasons for this based on lack of print space – but these no longer hold. Supplementary information is not widely published however, and when it is it is usually poorly structured and not discoverable by search engines. Robert Muetzelfeldt (Honorary Research Fellow at Edinburgh) stated recently that there is a problem with the word 'data': for some people it covers everything from numeric tabular data to structured symbolically-represented information, while for others it means just the numeric kind. Scientific communication is therefore hampered by the lack of a shared understanding of terminology and concepts.

Metadata too brings enormous challenges. It can be structured in a variety of ways. Ontologies are formal explicit specifications of a shared conceptualisation, and therefore hugely important tools in e-scientific communication. The most successful in biology has been the *Gene Ontology*, first developed to annotate model organism databases and so provide interoperability between them. David Shotton and Robert Stevens (Manchester) have set up the *Ontogenesis Network* to take ontology development forward. But even with good ontologies, creating metadata is time-consuming and difficult. Retro-fitting is only very exceptionally justifiable. We need to make it *advantageous* for researchers to create metadata;

we must also capture as much of it as possible automatically. Picking up once again the need to document the research lifecycle, Dr Shotton stated that commercial e-laboratory notebook systems are widely used in the pharmaceutical industry, where drug licensing regulations require them. But they are expensive. An alternative is the *FlyData Project* – a ‘cheap and cheerful’ system, very recently funded by the BBSRC. He gave an example of fruit fly research where the images collected were the ‘endgame’, following a long and complex workflow. It includes the literature databases, but includes many stages and steps, including the manufacture of probes and experiments from which observations are selected and published. Each stage currently is expressed through uninteroperable recording methods (Excel spreadsheets, etc).

The challenges in optimising scientific communication systems are formidable. Typical bioinformatics ‘in silico’ research is labour-intensive. Searches are frequently repeated (because databases change, or because searching can be quicker than navigating). This has the advantage of inserting expert human intervention at many stages, but the disadvantage of being costly. Automating such tasks would be possible if the data were uniformly tagged, but service providers don’t want to have to restructure all their data into XML. One tool which can help is *BioMOBY* which provides a single user interface to a variety of underlying bioinformatics services. *my Grid* at Manchester is based on an open architecture. *Tavern* (developed at EBI) allows views of intermediate results. Data visualisation, metadata and provenance all remain big challenges in this area however. *ihop* is a text-mining tool which mines PubMed Central for all instances of a particular gene and its association with another gene (by finding both names in the same sentence).

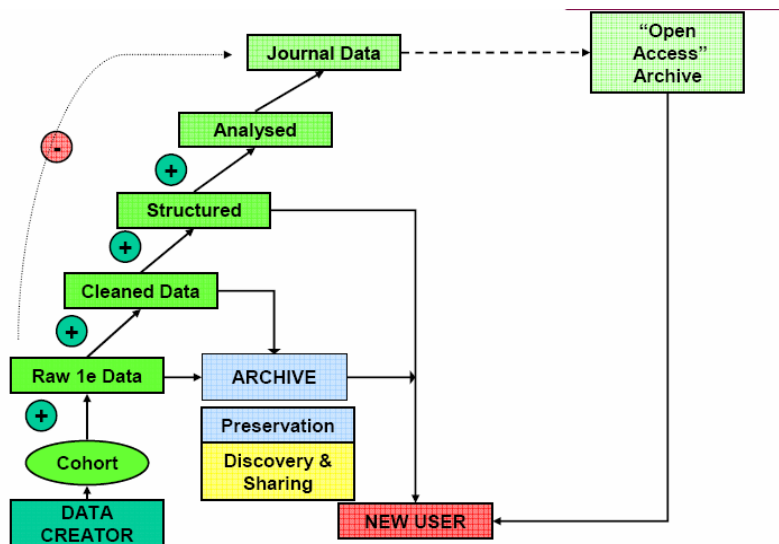
In much biomedical research it is not appropriate to submit the data to large-scale databases. Shotton is promoting a different approach – ‘publication@source’. The data should be published into personal or institutional repositories, with its required metadata. To do this, he proposes the ‘Data Web’. The data is simply published ‘wherever’, along with its metadata. The metadata is then harvested into a central ontology-based cross-searchable data web registry (expressed in RDF). It has a rigorous semantic underpinning, and takes users back to the original sources. The advantages are that it will be distributed, decentralised, evolvable, scalable and open. The Data Web provides support for open access data, going deeper than Google because it includes the ‘Deep Web’, targeting to a particular knowledge domain, integration of information and programmatic access. The analogy with research publication databases (where a ‘Shallow Web’/‘Deep Web’ split also confronts internet search engines) will be familiar to librarians.

The *ImageWeb Project* will be the first instantiation of the Data Web. It is the product of an extensive consortium, and now aims to link outputs to images from the free web. Publishers are involved, and are willing to share their *metadata* (they hold onto the objects). Starting with the published data means that it is working with safe data, and so avoids any confidentiality issues. We need forward and backward citation linking of data, so that we can see who has used our data, and we need to build in embargos in order to allay the fears of some scientists to releasing their data for reuse.

But the problems are not all within the eScience world, and indeed the challenges which eScience present can ironically reveal continuing challenges in the analogue data world which have still not been solved. For example, what about the analogue data which forms the historical scientific record? What should we save, and what throw away? We are in danger of losing the original data of pioneers of cell biology who are now retired or close to retirement. This suggests that we need large-scale re-education of librarians, archivists and domain scientists to address problems like this, or else the scientists of the future will blame us for waking up too late.

Research Councils have to start funding *infrastructure*, not just research. They can have a role in assisting the distributed structure. Chris Rusbridge quoted an Edinburgh colleague who said recently that a university's second most important asset (after its people) is its data. It is amazing therefore that we let it rot in the way we do. We clearly need to understand the motivations of scientists and researchers in order to create ways of achieving storage and curation.

In the afternoon session we heard from Peter Dukes of the Research Management Group at the Medical Research Council (MRC). He did not think that the MRC was particularly ahead of the other funders, except perhaps in genomics. He has responsibility for developing the MRC's data sharing and preservation initiative. Pro-active sharing and the preservation for it should be normal behaviour for researchers. The MRC is not just talking about the digital future; much of its asset base is in the past (punched cards, magnetic tape, microfilms, paper schedules, etc).



MRC data lifecycle model

The MRC has a data lifecycle model, shown above. There is little attention to preservation for sharing at present. Good practice exists in data management for primary research purposes, but not for preservation. The MRC now has a policy, but it lacks grassroots support for preservation. Researchers don't think about preservation. Instead, they think of an archive as a graveyard, while their more immediate concerns are that their data will be promiscuously mined and patient confidentiality will be endangered, or their data and work will be 'ripped off'. There is a lack of evidence of the cost-benefits, and a diversity of technological solutions. Another problem is that the data at present is controlled by Principal Investigators.

The MRC has launched a new two-year initiative which will come up with a definitive business model, working with a partner. They are currently working with two population-based longitudinal studies (the *National Survey of Health and Development* and the *Avon Longitudinal Study of Parents & Children*). They will hope to learn lessons and methods to apply in the future from these. All research proposals must now include a short Data Sharing & Preservation Plan (following the lead of the National Cancer Research Institute). The MRC *will pay for* data sharing and preservation (just as the Wellcome Trust pays for open access publishing of outputs) – but many proposals don't request it. Governance of access (patient and public rights) is also a big issue here, since it is key that users and creators trust each other's work, and that patients and the public trust the medical profession.

Jeremy Giles, Information Manager of the British Geological Survey (BGS) told us that data managers used to be happy that data stuffed in fileboxes would never be requested. Now that it's online, people do (this uncovers another analogy with libraries, which is the hugely different use of these once they are easily discoverable on the web). The high cost of data preservation can be justified by costs saved in, for example, not having to drill the same borehole twice. If we hold the data, we have the duty to provide means to search it. A grim confession was made to substantiate this point, and to emphasize the real importance of the work being done in this area. Jeremy Giles pointed out that the BGS had data (in an old manuscript form) which could have prevented the 1973 Lofthouse Colliery Disaster, in which seven men died. What's more, they had been asked for it previously, and had not tried to find it. This was a hard lesson.

Metadata must be supported by data collection and data transformation procedures. The NERC Data Policy requires that due consideration be given to post-project stewardship of data prior to approval being given for a project. Advocacy about the issues is hugely frustrating (as it is for eprints). We could learn from others about how to do this: he quoted the oil industry which has been doing this for a long time. Records managers are few and far between at the present time, which is putting sustainability at risk. Our paper record heritage (e.g. Kew) is massively expensive to maintain. The market for repository software is very diverse, and in a few years it is likely that there will be only a handful of big players. The reselling of data is one solution. Jeremy Giles reflected that the US gives more of its ordnance survey data away freely than we do – but then brings in greater income (from secondary services and corporate taxes).

The final presentation was by Michael Jubb, RIN Director, who summed up the fractured state of research data management with a Shakespearian quotation in the title of his presentation, *A Mote to Trouble the Mind's Eye?* We have substantial funds for research in the UK (£22b per annum), a substantial research community (200,000 researchers, half of them in universities) and a large number of experts in all fields. What we lack, however, is the coordination to make managerial sense of the data which emerges. The task which we all face, and which RIN is setting out to explore and to document, is to produce generic guidelines and principles which will regulate practice in the area of research data management generally, and to learn from the best examples from across a highly heterogeneous patchwork of practice at the present time.

John MacColl. 25 February .2007