

# **Rapporteur Notes – Mahendra Mahey**

Research Data Strand  
5 and 6 June 2006

## **Research data**

**Lead: Brian Matthews, STFC**

**Rapporteur: Mahendra Mahey, UKOLN / RRT**

Brian Matthews (BM) gave a brief introduction to the two days in the Data Strand section of the conference. He introduced himself as Head of the Information Management Group at the e-science Centre at the new Science and Technology Facilities Council.

He proposed that as there was a huge increase in the capacity to store primary research data, this could provide new opportunities in how we use it and study ways in how it may affect the way we do research, i.e. the way it modifies the research life cycle and what is going to be the role of digital repositories in this.

BM then summarised the programme briefly for the next few days, saying there were the following elements over the two days:

- Modifying the research lifecycle and data in terms of usage and long term preservation
- Generation of data for repositories
- Flavours of data repositories, e.g. images, geospatial data
- Panel discussion

## **Curation of Research Data**

**Presentation by Liz Lyon**

**Dealing with Data: Roles, Rights, Responsibilities & Relationships**

Overview

- Outcomes of a recent JISC-funded study by UKOLN
  - Institutions (repositories) and data centres
  - Roles, rights, responsibilities, relationships
  - High-level data-flow models
- Positioned in the UK context
  - 8 perspectives from Strategy to Practice
  - Examples of best practice
  - Recommendations

Presentation lasted around 20 minutes. A number of questions were asked

**When will the report be available?**

LL: After it is approved by the JISC board.

**Did Rights and Responsibilities idea originate from NERC?**

LL – Yes, they originated from Mark Thorley’s ideas of Roles, Rights and Responsibilities which he presented at a workshop that was organised.

**I can see how recommendations for research councils could be taken forward from your report, how will recommendations for institutions be taken forward apart from obvious ones to repository managers?**

LL -Recommendations of the report are going to JISC. (LL didn’t know much more than this apart from audit framework idea). She also mentioned Simon Coles was setting up a data and preservation working group at Southampton, which is where some of the ideas for the report have come from and are looking at all the issues around data within the institution.

**What about AHDS closing?**

LL. Stated that more infrastructure is needed to support data within institutions and that the funding council need to provide the funding for this infrastructure.

**Does the report summarize the funding councils’ position on institutions approach to data curation, preservation?**

LL – The only ones summarized were the ones that had already been visited. LL mentioned that her report was positioned below a recent Research Information Network report about datasets in research which is much more detailed and technical. LL’s report was trying to bridge a gap from basic reports and more detailed technical reports available.

**A question about the new generation of ‘Data Scientists’. Asking whether we were wasting our time about storing and sharing data if there are already Web 2.0 technologies which are being employed to do this. e.g. Simon Coles his use of blogs with research chemists, Open Source Science, and Wetware. Also mentioned that students were coming through as post grad, post doctorate with highly developed data manipulation skills .**

LL agreed that such graduates will have data manipulation and data curation skills, which may mean that they may use libraries less.

**Mentioned some research that was being carried on behalf of the British Library and JISC by UCL on the ‘Google generation’ and its impact on research behaviour. The conclusions from this research may have wider implications in understanding the way many researchers do their research today.**

Next presentation was given by Chris Rusbridge from the Digital Curation Centre entitled,

### **Curation of Scientific Data: Challenges for Repositories**

The presentation covered the following topics:

- Audience?
- Science and digital curation
- Why are data important?
- What kinds of data?
- What to do with data?
- Repository options
- Changing practice

The presentation can be found at:

Chris then answered the following questions:

**In reference to advocacy and Andy Powell's comments, if you get it right, people will want to use the service. I think we have to find this secret**

CR: We haven't yet. There are two Web 2.0 services that look in the data area. Only one of which I have used, Swivel – <http://www.swivel.co.uk> and another one is <http://www.gapminder.org/>. Swivel allows you to carry out cross correlations though is a long way from compelling, e.g. George Bush's popularity with drinking wine in the US. CR mentioned another example of the OECD publishing statistics through Swivel, however several statisticians from OECD were very upset about the fact that 'their' carefully crafted data was being publicly made available, there many attitudes need to be changed in terms sharing data.

**It is clear what kind of metadata is required for publications. What metadata is required for a cross correlation of data from domain A and domain B?**

CR: Lots, and we don't really know yet. Data for one community may be impenetrable rubbish for the another. There is a notion of 'designated community' and the first priority should be to curate the data for the designated community. Those who want to correlate the data are smart people and will try to work out how they can do this, despite the fact that there may be different names for the same concepts e.g. author, subject, intervals, time etc. What is important that there are some clear flags for outsiders to access the data, so some examples of good practice would be useful. The first priority is to make *data reusable within communities*. This will eventually lead to work *between communities* and this would mean this would be a lot better than we had so far.

## Generating and Depositing Research Data

Chair: Brian Matthews, STFC

### ***R4L: The Repository for the Laboratory'*** **Simon Coles R4L**

Simon's presentation is available at:

Summary of Simon's main points

Simon highlighted the main problems in the data generation sphere, namely:

Data Management – Small science generates much more data than Big Science

Data Deluge –so much data is being generated, it is difficult to deal with it

Publishing data – no accepted way to do this yet

R4L – being a possible solution to some of these issues

- essentially having a IR storing datasets being linked to publications

Simon talked about some of the problems and sustainability of the project over the future.

The following questions were asked:

**You seem to say two things that contradicted each other. Eprints and Fedora couldn't handle the data flow that was being used successfully by an RDF triple store, but you also said that you made the triple store fall over because of the flow of data.**

SC: This was about three years ago, so there was no way that an off the shelf repository was going to cope with the level of data flow we were dealing with, and certainly not in a scalable way and we encountered a problem that hadn't been foreseen in the triple store. The architecture was reengineered. But we are still using a triple store rather than eprints.

**But you said something about migrating to eprints?**

SC: No, eprints provides the mechanism for the deposit interface. I was trying to get across the same kind of thinking Andy Powell was saying that a repository is more around a set of infrastructure, it is not tied to once piece of architecture or one piece of software. We don't all rotate around eprints and use the same deposit interface. It is a set of services around a set of information resources, or data sources in the sector, it is a federation, a whole mish mash of whatever the institution needs to support its researchers. So you can pull in content from the RDF triple store or elsewhere. We are not trying to advocate it is one silo or one piece of software.

**The tools that you have written for automated blogging, how specific are they to Chemistry?**

SC: No, quite generic. What we have produced is an end to end demonstrator. We have tried to focus on bringing things to the desktop as much as possible, using tools that the end user is very familiar with. That in principle is generic. The workflow stuff and the structure of eprints could end up being one scientific data plugin for eprints, it can't be a catch all. It was quite a new thing to develop something for eprints where you could hang many records off a parent record. Eprints didn't originally handle this. We tried to be generic.

**Very interesting indeed but I sense a slight contradiction, in that you mention a bottleneck in that there is not enough time to write up papers (presentation and interpretation) for the data that is being generated. There needs to be a separation of the data that is being stored in a repository and the data that is being presented. R4I seems to represent a half way house, it is much more than just storing the data. Annotation tools e.g. the blogging tool. Are you not in fact inventing a tool that is a half way house between conventional repositories and publishers and what do publishers think about it.**

SC: I have presented this stuff to publishers, and may be I am a little bit guilty of miscommunication. May be the data deluge and bottle neck slide is a little confusing. Yes, we are building a set of tools and in general they are a set of lab tools. Not really intended for presentation. So we do need a steer from scientific publishing community for a standard template. In fact I scanned and surveyed PhD thesis to establish a template for the way data could be presented. Blog discussions were not intended to be public, as the blogs tended to be behind firewalls, it is for capturing and analysing data in the lab. The report generation is for distilling stuff from the original. Maybe it is a half way house, or may be slicing the two in half.

**Presumably a connection between published paper and underlying data**

SC: There would a link from a publication to the original data. At the moment it is in Southampton. Some publishers are happy for data to be in IR, others publishers want to store the data. On the whole publishers don't want to store the data. Royal Society of Chemistry – said it is too much trouble to curate data, or feel uncomfortable

**Alan Tonge, SPECTRa  
The SPECTRa Project :**

Available from

Summary of Presentation

Alan highlighted the problem that experimental Chemistry data was almost always omitted from traditional publishing.

Summarised a survey they had carried out with researchers in the following areas of Chemistry:

- **synthetic organic chemistry**
- **departmental crystallography services**
- **computational chemistry**

Main conclusions were:

- **A complex list of data file formats (particularly proprietary binary formats) being used**
- **Much data is not stored electronically (e.g. lab books, paper copies of spectra)**
- **A significant ignorance of digital repositories**
- **A requirement for restricted access to deposited experimental data**

Some outcomes and recommendations:

- **Data Management : No tradition amongst chemists (crystallographers apart) for organized deposition and re-use of experimental data.**
- **Data re-use : Additional analysis tools will be required to add value to large-scale data aggregates.**
- **Legacy Data : Did not appreciate the scale of non-conformance and changing standards for legacy file formats and data types.**
- **IPR : Who owns the deposited data? Guidelines for scientific data should be prepared by JISC in consultation with research funding bodies.**
- **Data Management : The project did not investigate the resource requirements for large-scale deposition and management of this experimental data**

One question was asked at the end of the presentation

**There is no real tradition for the reuse of experimental data. Why?**

AT: Thinks it is an educational matter. People need to be reinforced that what you have got is useful, what is needed is aggregations of datasets and a feeling of being able to do things to data that were not previously possible.

### **Publishing Research Data**

Chair: Brian Matthews, STFC

**Sam Pepler, CLADDIER**

**CLADDIER - Citation, Location and Deposition in Discipline and Institutional Repositories**

Presentation available from

A quick summary of the presentation is below:

The presentation covered the following areas:

- **CLADDIER aims and why data publishing is an issue for the project**
- **Data publishing issues**
- **Conclusions and future work**

Sam then defined some important concepts in this area:

Publication

Datasets

Making data permanently available

Encapsulation

There is no shared understanding for datasets.

Citation is a method of definition

Who defines the dataset?

***Data centres mean different things when they talk about Quality Control  
Researchers are more likely to cite data if they think its citable.***

***Can you peer review a dataset?***

- *Usability of the dataset.* This is the role of the data centres.
- *Usefulness of the dataset.* This is the role of domain experts.
- *Is the data well managed.* This is a data centre auditor role.

**Conclusions**

- As a data centre we need to define our datasets more precisely.
- There is no shared understanding for dataset definition.
- Mechanisms defining datasets in terms of semantics are needed.
- Desired citation is a way to define dataset, but you need to balance the needs of the citer and the cited
- The peer review mechanism may offer a way of creating the shared understanding and control the citer/cited balance.

There were no questions

**Graham Pryor and Ken Miller, StORe**

Summary of the presentation

- Aims and aspirations
- Practices and preferences
- A generic solution
- The StORe pilot
- Evaluation and prospects

## Aims

- Seamless transport from research data to research publications and vice versa
- Bi-directional links proven in social science e-research
- Capable of export to other domains

They carried out a survey of seven domains, some conclusions:

- Self sufficiency
- Risky data management practices
- Recognition of the value and demands of metadata
- Support for open access principle
- Caution over data access and ownership
- Personal networking

## Opportunities to

- Explore a deeper level of detail
- Supplement published papers
- Validate experiments
- Track the use and improvement of research output
- Identify collaborators
- Confirm completeness of information searches

## Potential risks from

- Uncertainty of peer review
- Premature dissemination
- Subversion of scholarly paper
- Scavenging
- Lack of interpretative data
- 

## Sufficient consensus to proceed with design

- Two-way links endorsed as advantageous (85%)
- Need to avoid bureaucratic structures and rules
- Sharing of data a fundamental principle but usually conducted on an individual basis
- Need simplicity and standards for metadata
- Simple Google-type searching preferred
- Self-management preferred to intermediation

## Features of Middleware

- Web 2.0 approach, similar to services like Flickr or MySpace, gives control to the researcher
  - Researchers determine which items are public / private
  - Researchers form collaborations with colleagues / 'friends'
  - Researchers select items for deposit and for publication
- Permanent links created between publications and underlying data
- Based on federations of institutional repositories and data archives
- Simple process for assignment of metadata

- Searchable metadata assigned at collection level inherited by items within the collection
- Collection owners add individual items plus minimum additional metadata (e.g. titles)

Ken Miller then gave a demonstration of the middleware that has been developed by the project.

No questions recorded

## **David Shotton** **Defining Image Access**

### **Presentation available from**

### **Summary of presentation**

The nature of scientific image data

- What data do we actually publish?

Integrating data from distributed resources

- Data webs
- ImageWeb

The JISC *Defining Image Access* Project

- Main achievements and conclusions
- Proposed future activities

What is special about images? – they are not self describing, metadata annotations are required to bridge the semantic gap.

Where to store the research data?

What data do we publish?

Original research data is never published

Described epochs in data integration

Need to improve this situation by the following:

We need to start treating experimental research data sets as **first class publication objects**, of equal value to the journal papers based upon them

- This includes recognition and reward for data publication
- 

We need to ensure they are saved with **good domain-specific metadata**

- This includes assisting researchers to capture metadata at the time of data creation, automatically if possible

We need to work towards **better interoperability between papers and data**

Talked about the semantic web and data web

The problems of achieving semantic interoperability between distributed heterogeneous archives of digital data are well known - several approaches used

None have applied to the problems of data integration the Semantic Web and Web 2.0 approaches which were described

They favour W3C standards in their solution

- RDF as the standard format for sharable metadata
- SPARQL as the universal query language for RDF
- Software such as D2R Server for abstracting RDF from relational databases in response to SPARQL queries
- OWL-DL as the standard web ontology language

For software:

- use of agile programming techniques
- Ruby or Python to provide a lightweight development environment
- loose coupling between the Model, View and Controller software components, based on a simple 'REST'-full approach to component integration (Fielding 2000, Representational State Transfer)

Data integration – the lightweight data web approach

Role of the data web

How might a data web improve on Google?

Web 2.0 aspects of data webs

**Image webs** are data webs for research images

The ImageWeb Consortium

Carried out Stakeholder analysis

Making the image web vision a reality

**Defining Image Access:**

Requirements for interoperable discovery and delivery of image data stored in DSpace, EPrints and Fedora institutional repositories using a data web approach

Activities

- Analysis of current institutional repository practice with particular reference to images
- Evaluation of repository software capabilities, particularly with respect to serving domain-specific metadata, to programmatic access, and to metadata harvesting using OAI-PMH
- Evaluation of third-party Semantic Web approaches and software tools to fulfil the functional requirements of a data web
- Evolution of our concept of what a data web for images might entail, driven by potential uses of data webs in other research projects

## Achievements

- have built a significant core body of knowledge and expertise concerning images in institutional repositories, and the problems and opportunities associated with integrating them, that is freely available for everyone to access from our wiki
- have held three very productive project workshops, that have served both to permit us to learn from other projects and also to publicize what we are doing
- have significantly refined our ideas about data web functionality
- have seen significant take-up of our data web idea for use in proposed new projects in the arts and the sciences

## Main conclusions – Data Web technology

- Semantic Web approaches and tools can be used to build a data web
- Data webs should comprise independent loosely coupled Web services
- The central data web aggregator, which will work with the schema registry and co-reference service to act as a query handler, needs to be created
- The degree to which it will be useful for the aggregator to pre-harvest core metadata into a central metadata registry is a subject for research
- The schema registry and co-reference service will simply store RDF data, but these will require hand-crafted alignment for each knowledge domain, since generic metadata is unlikely to be sufficient
- mSpace (<http://www.mspace.fm/>) seems well suited to provide semantically enabled browse and search discovery services
- Building a data web remains a significant implementation task

## Main conclusions - Repositories

- Institutional repositories currently contain few image collections
- Existing image collections mostly lack adequate domain-specific metadata
- Repository software is not equipped to serve domain-specific metadata, even if it existed
  - E-Prints: Possible by addition of two extra metadata fields to basic OAI-PMH DC profile, one providing URI for metadata and the other specifying its format
  - Fedora: Total flexibility, but you have to build your own interface!
  - DSpace: Serves basic OAI-PMH Dublin Core only
- These limitations make the creation of domain-specific inter-repository image webs both technically impossible and functionally unimportant

Went through proposed future activities

## Questions

**Images and copyright. Can you make it absolutely clear about the images that you create belong to community not the publishers.**

DS: It depends. Copyright of the images belongs to the creator. If the creators assign copyright to a third party, then it belongs to the third party. It is clear that the copyright is with the creator. Creative commons licenses are key to this.

**Anne Robertson**

**Scoping a Geospatial Repository for Academic Deposit and Extraction**

Presentation available from:

Summary

- Investigating and reporting on the technical and cultural issues surrounding the **reuse** of geospatial data
- Investigative in nature, not building a geospatial repository
- Particular focus on sharing and reuse of **derived** geospatial data

Why geospatial datasharing? – Demand seems to be there

Lots of examples of derived data – i.e. reuse of geospatial data

Went through workpackages

- Digital rights issues - when we consider the reuse of derived geospatial data concerns over data ownership, IPR and copyright are commonplace
- Investigate and make an assessment of informal mechanisms for geospatial data sharing
- Establish user based evidence for the requirements and functionality of a repository capable of managing licensed geospatial assets
- Debate over institutional repository – one size fits all? Cultural aspects of allegiance to discipline not institution
- Interoperability issues – how could a geospatial repository interact within JISC IE, how could it make its assets available to the GRID and eScience community

Main findings

- Repositories do have a part to play in assisting researchers share derived geospatial data
- A significant degree of informal geospatial data sharing occurs because of the lack of any formal mechanisms
- Community desire for a mechanism to legitimately share and reuse geospatial research data

- Main barriers to more formal geospatial data sharing within the community are:
  - perceived complexity of licensing and digital rights issues surrounding data (re)use in the UK
  - lack of quality metadata
  - concerns over the protection of depositors intellectual property
- Institutional repositories do not manage any geospatial content (and would not be capable of meeting the needs of those working with geospatial data currently)
- Geospatial community would support data reuse BUT not necessarily (at present) within an OA IR. More fine grained sharing mechanisms are preferred i.e. data sharing amongst peer group networks defined by the depositor

## Questions

### **Legal side. Given that when you ask two lawyers the same question and you can get two different answers**

AR: One person's interpretation can be different from another's. A separate study in the states, where the European Database directive doesn't apply came up with similar conclusions. Lawyers have come up with similar conclusions about copyright and geospatial images. Charlotte Webb from Carolina. Edina have passed on their data to JISC collections. Looking to keep this within Academia.

**Law of unintended consequences, OS may change the rules. We want them to grow up and allow reuse.**

## Panel session

**There were around 20 people in this session**

Chair: Brian Matthews, STFC

### **Panel Members**

- David Shotton, Defining Image Access
- Chris Rusbridge, DCC
- Anne Robertson, GRADE
- Ken Miller, StORe

BM introduced the session by saying that the panel session was an opportunity to get some feedback from the delegates of this strand and have a discussion. BM stated he would be asking the four panel members to provide a statement or impression of what they have heard during the conference and particularly through participation of the data strand session.

The panel members were introduced, all had previously presented during the strand.

BM then identified for him what he saw as the themes that were covered over the last few days in this strand. BM also thanked everyone for their interesting presentations over the last few days.

The main themes were:

How we put data into repositories, get it reused, and getting research and possibly new research from it is probably the ultimate aim.

Introduction from Chris Rusbridge and Liz Lyon, who have wide experience and have spoken to many people and seen a lot of systems and identified the key issues and policies and how we get a community engagement.

How we get the data actually into a repository and how it can be used, perspectives from the R4L and SPECTra projects, i.e. getting data from the desktop or getting stuff from the bench scientist and making it easy to get them into repositories.

Perspectives about publishing the data from the CLADDIER, STORE and GRADE projects, e.g. how to get data out there and how to get it linked and related issues of ownership and metadata for example.

Federations and linking, related projects being CLADDIER and STORE and also the Defining Image Access project, in how to get datasets to talk to each other so that they could share their information and all the related issues e.g. social problems around that including common formats and metadata.

BM then asked the question to the panel

What are the new opportunities for repositories to help us in the research life cycle research, what is missing, and what do we do next?

Each member of the panel then preceded to answer the question

Chris Rusbridge, Director of the Digital Curation Centre  
5 things he wanted to mention

1. If repositories to be useful, they have to exist.
2. 'Virtual federations' across institutional repositories. IR have tended to be generalist repositories tending to focus on eprints (scholarly works – full text) and haven't strayed too far into data sets, where domain knowledge is required. As the IRs role is to span across disciplines in an institutions and therefore they may not be able to fulfil the need for storing subject specific datasets. In order to tackle the issues that are arising because of the impact of the decision for AHDS funding to be removed by the AHRC there is a need to be able to build 'virtual federations' across institutional repositories that do have links to

domain specific knowledge in science and the humanities. This is a real opportunity for people to call out and say a federation across repositories or sets of federations spanning across research areas would tap into domain knowledge but allow for generalist set of platforms

3. Ease of use is and particularly easy deposit. Many repositories are virtually empty. SDFC has around 20, 000 publications but this could be because this is a less collegiate and more managerial type organisation where there is much more of an explicit demand and requirement for people to deposit, this may not however translate readily in a more traditional organisation such as a university. There needs to be an easier way to deposit which is less challenging and be able to get more value from it.
4. In order to recognise the importance of good management of data, researchers don't see that as their role and what they are paid for. However there needs to be a recognition that research staff have to manage their data well in order to do good research, even if the benefit would only be for themselves. If good management of data is to be rewarded the academic currency for this is citations, so better mechanisms are needed for the citation of data in the digital domain to work.
5. Universities are beginning to build large repositories for research data, for example Storage Area Networks (SANs), but by and large these are large data dumps. Need to engage with institutions in how they should be curating this data rather than seeing it as going into a 'dump'

## **David Shotton**

Mentioned that two of the items on Chris's list were on his list.

1. Citation of datasets – he stressed a lot of the issues on ease of use and recognition an academic currency really only come from peer reviewed articles. This narrow focus will need to be broadened if we expect academics to deposit data related to their research. There is a need to identify proper mechanisms for identifying datasets, i.e. the citation of datasets, this is an issue that clearly needs to be resolved (i.e. identifiers)
2. Repositories need to move beyond OAIPMH as a method of delivering data, to embrace the semantic web and create SPARQL end points. David mentioned that Les Carr (leads Common Repositories Interfaces Working Group – CRIG) has applied for funding to JISC to do this. David implored the audience to try and put pressure on JISC to do this also, i.e. for current repositories to generate SPARQL endpoints. It wasn't realistic for institutions to convert their data into RDF, but there

should be a requirement to access legacy formats in such a way that they are accessible to the semantic web.

3. CLADDIER has demonstrated the importance to be able to query datasets and papers simultaneously from different domains, but what is also important is the ability of academic papers to cite datasets in a mutually enforcing and updatable way and therefore version control, which is a difficult problem, will need to be addressed.

### **Ken Miller, Store Project**

Stated that one thing that is missing is content.

Also there need to be ways of making repositories 'sexy', for example, using Web2.0 type technologies.

Another issue is being able to catch content early, it should be part of workflow of a researcher and should be automatically collected. Also there should be a way of automatically trawling content and a set of rules which could be used to decide what data needs to be preserved for a longer period of time, may be based on the type of file or the associated workflow. And at the other end focusing on the publications which is obviously the main focus for researchers as this is what they get brownie points for there could be direct links with IR and publishers so that the actual process of publishing data is somehow connected via the IR, i.e. data publication is delivered by the IR

David Shotton also mentioned that he was surprised how open publishers are open to the idea of data integration.

### **Anne Robertson: Grade Project**

Decided to come from the Geospatial angle and said that there was a great opportunity to ensure that geographic information is added to the metadata for data objects. Essentially datasets are created in a place, therefore the geographic location could be easily extracted from an IR. Even simple things like geographic place names are useful as there are services which provide a Geo parsing service such a GeoXWalk which can take a place name and translate it into specific co-ordinates. 80 % of data has a geographic component and it could become a critical component in the way we search for information. There are already services such as google maps and google earth, the so called 'Google Effect' which have already been utilised through mash ups which can give interesting ways of searching for information. These tools have the ability to engage people in new ways in terms of searching for information. Fundamentally another method for searching for information.

David Shotton mentioned some work which relates to the idea of using Web2.0 type technologies and 'Citizen Science' to open up repositories, to allow or external annotation for people on the 'outside'. This is from a UK e-

information group meeting in Manchester. We all have views that we are the experts and we own the data, we want to guard it and we don't want anyone else to have access to it or tamper with it. He mentioned an example of a 'herbarium'; curator who is in the process of digitising his collection and he is getting experts from around the world to add metadata to it. The process of digitisation involves digital images of the plants as well as scans of delicate hand written notes. He making all of this accessible and people are volunteering (mainly interested botanists around the world) to transcribe the notes into metadata fields. David felt this change of mindset in academia would be enormously valuable for repositories.

BM mentioned how there is also a huge amount of supporting materials that accompanies annotations and also Simon Coles Chemistry Dataset blog was creating new research

David mentioned that a student of his was developing a prototype ontology of animal welfare, she is going to blog this and allow people to develop the ontology in an open way rather than the closed way they tend to be developed.

Howard Noble asked about two things; ease of use and good metadata and admitted that these two things can be contradictory. HW mentioned that some metadata schema development is open to web 2.0 type development, however, what about schemas that are highly structured and usually require skilled experts to create them, would a web 2.0 way to develop these types of schemas be possible, or not?

David Shotton, mentioned one of the best examples he could think of, where a highly structured metadata schema is National centre for e-Social Science Manchester (ESRC). Connected to policy grid in Manchester? Developed an interface that was backed by an ontology. There was a controlled vocabulary for entering the metadata, and tag cloud of terms other people used when they were searching and didn't find what they were looking for. A user could use their own tag, tag from tag cloud or a term from the formal ontology vocabulary. Demonstrates that the informal and formal mechanisms coming together, the informal methods can inform the ontology Alistair Miles who works with BM also has clever ideas how this can be done.

Howard Noble mentioned an example of epidemiology datasets, where there where many expert users using it and cataloguing, though there was a hierarchy in terms of what changes users could make to the record, gradually improving the quality of the metadata.

Brian mentioned that in reality people are writing metadata all the time without realising it, throwing it away or locking it in forms that can't be used. E.g. users are very willing to add metadata in research proposals, reports to boss, even e-mail and text messages e.g. I have found this really interesting thing and it is about....or in lab note books, or typed up into journal paper, locked up in pdf,

CR: Stated that repositories are not good enough at capturing Metadata, systems are not yet smart enough to provide hints such as 'We think the title of your paper is...X' and your name is 'Y'. Those kinds of things would make it easier.

Eprints repository at Edinburgh is extremely easy to use, just need to e-mail somebody the files and they do all the rest. So even as simple as this there are issues about getting enough content.

Not enough domain specific tools, too much emphasis on Learning objects and VLEs. We need to develop domain specific tools. i.e. the practioners need to lead this.

CR said that no one sees the logic of depositing to a repository or advantage of doing this. Mentioned the situation of writing a chapter for a book that cannot be published for six months, trivial activity requires that in six months time I have to remember to make it visible. So embargoing and more specifically conditional embargoing should be easier.

Comment: We are not making it easy to get data from experiments and research into a repository.

Grade – mentioned that repository demonstrator asked for the bare minimum of metadata for geospatial data, tools available to automatically populate specific metadata fields, e.g. bounding box.

Fundamental problem here is that a repository is store where you put a completed item of work, rather it should be embedded more in the whole working process. If they were more like part of your desktop, file store, a mounted disk where you could store items, a public space, .kind of 'MS Sharepoint' way of going about things, you are managing your documents through the entire process, i.e. deposit metadata as you are going along. A repository should be something in inverted commas, not one single entity, sets of services based around.

The problem then becomes how do you preserve certain parts of it.

It depends on the object you want to preserve and then you have to do some final tidying up.

AR – part of the demo deposit process, so that it is minimum, taken from deposit title.

CR asked the audience how many habitually type in property information into documents, e.g. author, title

Anne mentioned that if you work in an organisation where the policy states that you should do this then it will happen.

In MS Office 2007 there is a publish wizard, Lotus Notes does the same.

In much more controlled environments would work but unlikely in an academic environment.

CR mentioned that Glasgow University did some research to find out how many extra clicks would committee clerks accept in order to make their materials preservable (CR). The answer was zero. Then a system called CDOC (long in the tooth now), which uses a 'negative' number of clicks. Presented with a few additional buttons in their work, which lets them fill in property boxes, but then automatically rewards them by automatically doing a whole bunch of other things later on in the process, turns documents into word, rtf and xml and produces HTML versions and puts them up on a website, by dint of already saying who they got all this extra stuff for free, so for a negative number of clicks you get a preservable records management of the committee process. So it is about changing the mind set, i.e. doing a bit more work will ultimately mean in the future you will have ended up doing less work. So if you could build systems like that, very much like built into the work flow of academic and researchers.

Coming back to data, there is no equivalent of a property box for a dataset.

SP – that depends on the data, for example in images, there are lots of properties that can be defined and added. Asked how metadata was added in biological images.

DS – for example it is possible to obtain metadata automatically from a microscope which is capturing the image. Not easy ways for a person to add metadata to an image.

There are ways to add metadata to specific image formats such as Geotiff – i.e. adding geographic information and other types of information such as file formats, conventions etc, however this is not widely used.

Simon Coles then wanted to pick up on the point that CR has made about institutions having large stores, SANS, 'data dumps'. Terabyte stores are becoming a reality. He is on a university working group trying to assess the data output that is coming from the institutions, so that they can up with a strategy for preservation policy, so that they can convince senate so that their institutional repositories can be supported. There is also a realisation that there is going to be a massive problem of the size, diversity and heterogeneity of the data. So how do we approach this problem? To preserve this, you can't have structured schema, you are going to have to get metadata from depositors, either provide and to make it a requirement. Question to ask is whether a 'put' and 'get' mass store a bad thing?

CR said that is it a good thing, better than put and get 'local disk'. IR managers have to start having dialogues with the people who are generating the policies for these large disk storage places e.g. IT Services, and also with the scientists and researchers who are starting to use them, getting people to

engage with the them and use them. IR are not a library silo, it is institutional wide, however it is often stuck in a library silo librarians need to get out more!

SC there must be minimum set of metadata for storing data, so that the objects can be preserved. It has to exist first and then it has to be considered by someone if it is preservable. If there metadata isn't there to make it preservable then the object shouldn't be preserved. (BADC person) Sam Pepler.

SP have to have a minimum set , got data and you have to look at . SC hugely time consuming process (and library budgets will not stretch to account for this)

SP there is huge amount of work that still needs to done

Chris – Reading, have to follow the money, if they contain nothing or if they contain something they are driven by the motivation of the person who wants to deposit as opposed to the needs of the reader or consumer. It is going to cost a lot of money to fully populate each institutions repository

Massive quality assurance in research agenda universities are having to address, funder and government agencies will only fund research where it can be proved that the research and the process of collecting the data is vigorously managed in the first place stored securely, fully attributable, trustworthy and not doctored after the event, and managed according to schema. If Institutions don't do that then it is going to remove research funding. So there is a need for Inst to manage data properly, right from the collection phase and through the workflow of the research process. These are some of the drivers which will contribute to whether subject based or institutions based repositories will dominate.

ERC fund UK data archive and hold back some money until the data for underlying research is deposited into the data archive. We list quite bureaucratically lengthy metadata to ensure that the data can be preserved and that it will be reusable

Only depositors fill out the forms is that the ERC are holding back some money.

SC, stated that the reason why Epubs has so many publications in it is because you can't get more instrument time unless you have proved that there was output from your previous usage and the only way to prove that is a record in Epubs

Sam Pepler from NERC – said that this doesn't happen, there isn't a mechanism to enforce adding metadata and deposit. NERC should inform us

CD- NERC should have a Keith Jeffreys CRIS, grant, RSS feeds, take them seriously, i.e. conditions of grant. I.e. don't have a condition of grant or take it seriously one or the other

CR Stefan, Norway funding next years research

Stefan mentioned about the commissions study look at what is meant by data publication, use in inverted study, empirical study across a number of subject disciplines. How it is in practice and how they manage their data in terms of disseminating it, what stage in the process is data published. Need to look at the issues, quality issues, important of data is recognised, and citation, in discussions with publishers a lot of interested in data is published. Not just looking at data as a precursor to published outputs, or could be replaced, or how data complements published output, more information on website

Key Perspectives are doing this, basis for further study, research assessments, current JISC call have (CR) has an innovation, small projects to innovative on repositories using additional repositories, trying to make things better, mash ups and do it.

Get out more!

BM Concluded the session.