Fourth IEEE International Conference on eScience

Rethinking Metadata Creation and Management in a Data-Driven Research World

Andrew Treloar
Australian National Data Service Establishment Project, Monash University
andrew.treloar@its.monash.edu.au

Ross Wilkinson
Australian National Data Service Establishment Project, CSIRO
ross.wilkinson@csiro.au

Abstract

Research data collections are tremendously important and thus need good curation. However data collections are significantly different to publication repositories and so we need to ensure that these differences are taken into account when managing research data. We believe that a good way of approaching this problem is to articulate the needs of research data stakeholders – particularly users and creators. Consequently we have described an analysis of these needs and then examined costs in the light of these varying needs – it is important to note that costs are often incurred by different people to the beneficiaries. We finish the paper by showing practically how incurring software costs can provide valuable savings for both data creators and data managers.

1. Introduction

Publications have a very well understood publication process that produces the information that is needed for finding these documents in reliable ways. This process has a heavy human involvement both by necessity and history, but is well embedded in the architecture of scholarly communication [5].

Data associated with a research project are different to typical publications describing the same project on a number of dimensions:

- **Size** of each file: data files are often orders of magnitude larger than the corresponding document(s)
- **Numbers** of files: a research project will often give rise to many tens (or hundreds) of data files, as opposed to a small number of publications
- Variations in **file formats**: data files come in a bewildering variety of file formats, both proprietary and open
- **Lack of human involvement** in their creation: increasingly data files are generated automatically from instruments or sensor networks
- **Internal complexity**: rather than a single sequence of text (typical of most publications), many data files contain significant internal complexity
- **Human readability**: a data file may need significant extra metadata to explain the contents (for instance, the row and column identifiers, and a key to the meaning of the data values)
- **Inherent semantics**: documents often need no external object to understand meaning, whereas data almost always needs significant metadata, ontologies and other artefacts to interpret them

The recognition of the importance of data began in the sciences [2, 15] but has now moved outside these disciplinary boundaries. An increasing amount of data in the social sciences and humanities is being created digitally [1]. Some of these data collections can be of significant size, particularly where audio or video are involved. In order to deal with these large collections, it is necessary to manage them effectively for the long-term, and use some form of information retrieval service to discover them.

This analysis of the metadata creation space is written from the perspective of the Australian National Data Service or ANDS (http://ands.org.au/). This has just been established by the Australian Commonwealth Government to facilitate better data management, sharing and discovery across all Australian research. ANDS assumes data storage in institutionally-supported repositories, and is building a range of discovery services to facilitate discover across these stores.

2. An information needs based approach

Drawing on the traditions of library science and information science, Ingwersen & Järvelin[9] have recently provided a new perspective on why it is very important to know why information is being sought, not just what the information is. This wider context should be explicitly modeled and used in evaluation of effectiveness. This is just as true for data discovery as document retrieval. For example, if a researcher is formulating an experiment, and needs to find out all relevant data sets to the experiment, then it is important

978-0-7695-3535-7/08 $25.00 © 2008 IEEE
DOI 10.1109/eScience.2008.41

to get wide coverage using any system to discover the data. On the other hand, if a researcher is in the middle of an experiment, and needs to access research data from a colleague, then they need to get exactly that data set. Naïvely, exactly the same query might be issued of some discovery service, despite the widely differing information needs. Less naively, exactly the same query might be issued to different discovery services, and each need would be appropriately satisfied.

3. Information needs

We believe that a good way of addressing these issues is to concentrate on information needs in the first instance, and then the information that is needed to satisfy those needs, before finally turning to the technologies and data that might be needed to support these needs.

Needs come first – clearly a dominant need is simply access to known data, held in a known location. The reasons for this access may vary however – there is a need for the creator to access the data to check their work, to re-analyse using different tools and parameters. Other researchers access data to apply their techniques to existing data, and to agglomerate the data. Importantly there is a need for independent researchers to access data to test the claims of a research outcome. For all of these needs, the metadata stored will be targeted at interpretation of the data and its acquisition environment – there is no requirement for discovery metadata.

There are other needs that might be best supported by a wide variety of access methods, each with their own metadata requirements. In that context, we elicited views from Australian researchers and data managers on the information needs, the information, and finally the sorts of technologies that are required (there is as yet no public report describing this work).

We found a wide range of needs:

- Finding a specific resource within a discipline
- Finding a specific resource across disciplines
- Alerts to new data as it becomes available
- A data review that would be used as part of the start of any research program, just as a literature search is carried out – by its nature this requires coverage, rather than item search
- A perspective beyond the domain of inquiry - users might be issuing queries within a domain, but are provided with references to collections that are beyond the domain but potentially relevant using cross-walks provided by a collections registry – the needs are met within the domain but exploiting a discovery service that walks across very high level ontologies to find other potential collections
- Researchers needing information that enables them to connect with other researchers and their data to expand their capability
- An overview of collections as a whole
- An overview of the collections together with the research creators, institutional custodians, and data services available
- Novel information perspectives to support innovation that comes from accessing information created for one purpose being used by another person for a different purpose (so no having barriers, such as domain language and access methods, is important)
- A view of research outside any particular discipline to support cross disciplinary awareness through data awareness

In order to satisfy these needs, we will need access methods with a variety of features, again with implications for metadata: we see that sometimes the researcher will be a specialist, with metadata that is particular to the discipline, and other times, researchers outside the discipline for whom domain specific taxonomies might mean little. When researchers are exploring across different spaces to get an overview, there is a need for language that might interpret domain specific information. The data itself might be in a variety of formats, so it might be important to have metadata that enables transformation into a common format. For streaming data, and frequently updated data, it will be important to know temporal information in order to issue alerts. We have seen research needs where data in a discipline may be relevant to an information seeker but not in a language understandable by the information seeker – this has significant implications for metadata approaches.

Notice also that it is not only data but researchers, or research projects that researchers are seeking – descriptions of people and projects, with corresponding metadata might be important. ISO 2146 [10], an international standard currently under development by ISO TC46 SC4 WG7 to operate as a framework for building registry services for libraries and related organisations has recognised the need to support these expanded information needs. ISO 2146 recognises Collection (an aggregation of physical or digital objects), Party **(**a person or group), Activity **(**something occurring over time that generates one or more outputs) and Service **(**a physical or electronic interface that provides its users with benefits such as work done by a party or access to a collection or activity) as first class objects.

4. Metadata and its importance for discovery

Metadata can be characterized according to the attributes of the object it encodes. This leads to the following different types (although there is no one agreed metadata typology):

- Descriptive Metadata: identifies the object and describes its contents
- Technical Metadata: often derives from how a digital object was created, captures its format-specific technical characteristics
- Structural Metadata: encodes either the internal structure of the object or the structural relationships between this object and related objects
- Preservation Metadata: encodes information about the object that will be needed for later preservation/curation activity
- Provenance Metadata: describes the transformations that have been applied to the data (this can be useful for later audit processing)
- Rights Metadata: describes the restrictions (or lack of them) that apply to various uses that might be made of the data.

While all of this metadata is potentially searchable, the descriptive metadata is often essential when dealing with data objects by acting as a proxy for the object. This still enables discovery where direct indexing of the object is not possible. Example object types might include images, audio and video, numerical data and a whole range of proprietary instrument formats. Discovery of these data objects requires metadata.

5. Scholarly communication decomposed

We have examined needs from an information seeker perspective; however we can also look at needs from a provider perspective. Roosendaal and Geurts[18] argue that scholarly communication models fulfill five basic functional requirements:

- Registration, which allows claims of precedence for a scholarly finding
- Certification, which establishes the validity of a registered scholarly claim
- Awareness, which allows actors in the scholarly system to remain aware of new claims and findings
- Archiving, which preserves the scholarly record over time
- Rewarding, which rewards actors for their performance in the communication system based on metrics derived from that system (restatement in Van de Sompel et al. [20].

In the case of publications in journals, the first four requirements are met by a single artefact (the journal) and take place within a small number of organisations (publisher, indexing/abstracting service, library). There is also a well-structured set of roles performed by various players (authors, editors, reviewers, readers). The costs are either made explicit, or are shared under well-understood arrangements.

These needs or requirements hold true for data providers also. In the case of data objects, the requirements are not concentrated in a single artefact, the roles are much less agreed and more fluid, and the rewards do not accrue to the same place as the costs. As a result, it is important to consider the cost implications of decisions made about the creation of metadata that support those needs.

6. Making cost-effective metadata decisions

From the above discussion it is clear that some level of metadata is essential when working with data objects. And yet, the cost of creating this metadata (particularly descriptive, but the other types listed as well) can be very high. The cost per object may remain static, but the number of objects (and hence the total cost) will continue to increase, as more and more instruments (with increasing capacity) come on line and as data-driven research assumes greater importance.

However there are potentially very significant costs in creation of the contextual infrastructure and then the capture of the relevant metadata to fully describe the context. With a fixed budget, how does one decide what metadata to capture? Information systems researchers have extensively investigated a variety of frameworks that enable a cost/benefits approach to deciding on features of an information system. A good paper on this approach by Delone and McLean [6] describes the many costs and benefits that need to be considered. However the overwhelming costs are usually human costs – those of data creators, those of data managers and curators, those of data discoverers, and (often importantly for the research context) those of data transformers. Comprehensive metadata capture loads effort on the creators, managers, and curators, whilst minimal metadata, loads the costs on data discoverers. How does one strike the right balance?

One possible answer to the question above is to capture as much as you can afford, and derive the benefits that are possible with that level of capture. However, we think that the range of benefits are considerable, and that effort put into some forms of capture may provide no benefit for the area that is important, or becomes important. We need to recognise that there is

a very limited cost that we can incur – far less than is necessary to fully support all of the possible needs. For this reason, we believe a useful starting point is to enunciate benefits that are sought first, as above, and then look at capture costs in the context of the benefits.

6.1. Overview of current practice

The process of determining what metadata to keep based on expected benefits is well established in the records management and archival communities where business benefits are routinely analysed as part of determining metadata needs [4]. In many cases this is made somewhat easier as all of the costs are borne by the one organization.

In the case of archives, there are often legislative requirements meaning that some metadata becomes mandatory. In the case of libraries, there is again a well established process of determining metadata based on needs – because the needs are so enduring the metadata requirements can be simply derived from the needs, rather than subject to a cost-benefit equation. Most information systems undergo requirements specifications based on both business needs and benefits, however they are looking at a broader set of requirements within the context of a single enterprise.

All of these areas are relevant to understanding and analyzing costs and benefits for research data and yet we cannot obviously directly draw a process from any one of them that enables us to balance emerging information needs from a wide variety of users against the costs that could be borne by a wide variety of organizations. The analysis performed by the Research Information Network [16] shows that the needs vary across research disciplines. The Classics (where raw data is rarely made directly available, but annotated data may be a significant contribution) have very different approaches to that of Astronomy (where data is shared according to well articulated international agreements).

6.2. Cost-Benefits analysis

It is common to undertake a cost-benefit analysis from a single institutional perspective – this enables a research library to answer a question such as do we purchase an on-line service, or explore trade-offs between increase metadata creation and discoverability of data. A very good example of such an exploration is that of Beagrie et al. [3] which examines how to keep research data safe from a UK library perspective.

They show classes of benefit accrue from preserving research data: avoiding the cost of re-creation, using research data for new purposes, promoting the institute through increased access to their data, and improved research processes – enabling for example, validation of results, or determination of research value.

They then describe a cost model which involved describing an activity model, the resources needed for the activities, and the costs associated with the resources – they note that costs may vary considerably over time, but using this model they estimate that the cost of running a data archive may be an order of magnitude greater than running a publications repository. This indicates the need for great care when determining research data management approaches.

Their work is derived from their own investigations, but also draws on the work of NASA in developing their Cost Estimation Toolkit [7], and the OAIS Reference Model [11] amongst others.

From the perspective of metadata creation and management, we can see that the above analysis applies from an institutional perspective. However Borgman [5] describes an information lifecycle where information value may accrue. As we will explore further, metadata creation can occur at many points in the data lifecycle, and so the costs can be borne by many different people and institutions. By developing appropriate software, it may be possible to capture metadata at the point of creation. The researcher might also manually enter metadata at that time. Metadata might be entered – either manually or automatically, at the time the data enters a research group data management system, and further metadata might be added at the time it enters either an institutional or disciplinary data archive. Metadata can be added at many times after that. The costs might be borne by many different individuals and institutions over the data lifecycle, and the benefits might accrue to many different individuals and institutions.

It is beyond the scope of this paper and the authors' expertise to develop a cost model for metadata that deals with these complexities. However it is important to note here that there is a very wide range of benefits and costs, and that a single institution perspective on metadata capture and management will not suffice.

6.3. Reduce or share costs?

We have seen from the report by Beagrie et al. [3] that costs for data can be very large so it is clearly desirable to reduce these costs or share them, since we have also seen that the benefits are shared widely. From a metadata perspective, storage and computational costs are not likely to overwhelm, so we are able to concentrate on human costs. Who bears the costs?

There are a set of people who might do so: the researchers creating the data, the data custodians, who

are often information professionals such as data managers, librarians and archivists. There are the data user who could annotate data, and there are information technologists who might create software to automate metadata capture, develop data mining tools, or develop tools that improve curation efficiency.

It is often the case that the cheapest point of metadata capture is at the time of creation - this might be a combination of instrument metadata generation, default metadata values that apply because of context, and certain metadata terms that a researcher might apply from taxonomies that make sense locally. However many of the benefits do not accrue here. Consequently, without appropriate technical or policy intervention there may be little incentive for a researcher or research group to capture metadata beyond the needs of the person or group.

As has been indicated in the Beagrie report [3], there are significant benefits that accrue to the institutional custodian of the data, so there is a corresponding interest in metadata capture by the information professionals who manage the data. It is at this point that metadata for uses beyond the domain might be applied, as well as metadata for the purposes of management. Again, tools can help here, but there is likely to be human effort needed, and this cost can be substantial. See the Beagrie report for sample costs.

Collections registries managers, national research organizations and international research data organisations might also contribute metadata for data beyond their control. They might use data mining tools that extract explicit metadata. They will crawl data collections using software such as DataFountains (http://ivia.ucr.edu/#DataFountains). They may well independently index the data using content analysis tools much as Google might do with pdf documents.

Finally we have the users of the data – there are a couple of viable ways that effort can be reasonably expended here. The first is that users may simply expend more effort discovering data that is relevant to their needs. This will occur if the metadata has been designed for the original purposes and there are no taxonomic translations available. Another viable alternative is that the metadata associated with a data set is simply the documents that describe that data – quite possibly the research proposal that describes what data would be collected, and secondly any research reports or papers that are derived from that data. This method has been proposed for some chemical research data by Murray-Rust [14]. Note that some of the needs we have described are better suited to this approach than others. Researchers needing to connect to other researchers, or to access a collection they know about, but not its location may well have a very good success rate – search engines are pretty good at these tasks. However researchers who are attempting to do a comprehensive survey run into the standard problem of search engines – they are poor at recall – full coverage of a topic,. Also researchers who are attempting to link into new disciplines might not use the terms of the discipline, even though the concepts are appropriate face another well understood limitation of search engines – the vocabulary mismatch problem.

The second way users can contribute is by external annotation of the data. This activity has been fundamental to many disciplines – law, religious studies, literary studies amongst them. However we have recently seen the immense power of annotation when there are enough participants – Flickr uses community annotation of photographs to great effect. Research is currently being conducted into means of most efficiently supporting controlled annotation within a research community [8].

We see thus that there are many options for managing the costs and benefits associated with metadata capture. It is clear that the optimal solution for different participants in the information lifecycle will be different. Any local solution is unlikely to be the best global solution. Consequently it appears to us to be the case that there is a clear need for national and international approaches to optimizing the research data systems. We do not expect that the same solution will be applied universally, as there will be national and disciplinary needs that vary.

Nevertheless we see a need for drivers that influence the point of capture of metadata and associated tool development that drives down the total cost of metadata capture as our demands for wider uses of research data grows. Policy is one driver here; research funding agencies might well be concerned to achieve both the cheapest, and most efficient, environments for discovery of existing data sets. A possible solution might be an agreement for appropriate sharing of effort between research data creators, curators, and users, in an analogous way to the sharing of effort among different players in the research publication system.

7. ARCHER case study

We now examine a practical example of how costs can be shifted and reduced in order to capture metadata in the crystallography domain. Notice that the transfer of costs is unlikely to be justified by the benefits for a single research group – in this case a national initiative applied.

7.1. ARCHER overview

The ARCHER Project (Australian ResearCH Enabling enviRonment) was funded by the Australian Government's Department of Education, Science and Technology in 2006 as part of the Backing Australia's Ability II program. ARCHER supports data collection and management, as well as collaboration over the data. The underlying storage technologies ensure that data remains well curated and publication-ready, with appropriate metadata, provenance, and authorisation. ARCHER has produced a suite of tools developed jointly by Monash University, James Cook University, and the University of Queensland, drawing on, integrating and extending existing open source toolkits. It provides infrastructure to assist researchers in collecting, managing, storing, collaborating on, and publishing scientific data. ARCHER completed its tool suite in September 2008, and has made its products and source code openly available at http://archer.edu.au/. ARCHER was a significant investment in software development costs in order to remove costs from research data creators, and research data managers.

While ARCHER has developed a number of different components, two are particularly relevant to this discussion: DIMSIM and XDMS.

7.2. DIMSIM (Distributed Integrated Multi-Sensor & Instrument Middleware)

New scientific instruments are producing and collecting data at very high and increasing rates, and conventional practices, such as storing the collected research data on CDs or portable hard drives, will not suffice to ensure long term storage and management. Other potential challenges include: dealing with complex and distributed instruments; determining the status of a remote experiment; transferring data from a remote instrument to the desired data store; and starting an analysis while the experiment is still running.

DIMSIM solves these problems, and allows multiple sensors to be easily integrated. It is built on CIMA (Common Instrument Middleware Architecture) [12], which allows instruments to be abstracted and exposed over a network. This facilitates direct deposition of collected research data into a network data store without human intervention remote control and remote telemetry. ARCHER has designed DIMSIM to deposit the collected research data directly into a data repository based on SRB [13] that supports rich metadata through another ARCHER extension This allows curation processes for the collected/generated data to begin at collection time, improving curation quality and with the intention of substantially reducing its costs.

7.3. XDMS for data management

XDMS is the web based data management component of the ARCHER suite of e-research infrastructure tools, and sits on top of ARCHER's data repository. It promotes good data management practises and provides researchers with data access, data deposit, data export, curation facilities, search and discovery services, and the ability to associate persistent identifiers with datasets.

XDMS provides two levels of metadata support: a generic core metadata profile, applicable across disciplines, based on the CCLRC (now STFC) Scientific Metadata Model; and a domain-specific metadata profile, which is user-configurable, and editable by the ARCHER Metadata Editor (another ARCHER component – not discussed here). Metadata associated with the various levels within the CCLRC metadata hierarchy, including discipline-specific metadata, can be searched and browsed, enabling researchers to easily locate objects and collections.

XDMS provides support for the deposition of research data, and can automatically extract a datafile's metadata from its header and associate it with the deposited datafile. XDMS can export research data in both native file format and packaged into a METS format, in readiness for ingestion into a publication repository. Publication repositories are where data is made available to a general audience rather than the collaboration group [19], with a guarantee of long-term persistence. These are typically provided by institutionally supported repositories, and use technologies such as Fedora and Dspace rather than SRB; so packaging is necessary for transferring the data across. The ARCHER project has been working with researchers at Monash University to ease the migration of publication-ready datasets across the curation boundary. This includes migrating and augmenting the metadata as appropriate.

7.4. Crystallography example

Do these tools actually shift and reduce costs? One end result of applying the ARCHER toolkit model and the associated migration process is a paper by Rosado et al. in *Science* [17], where the final published version points to a dataset that has been migrated across the curation boundary [19] into the ARROW Repository (http://arrow.monash.edu.au/hdl/1959.1/5863/). This process was somewhat ad-hoc the first time, and involved a lot of manual work and creative problem-

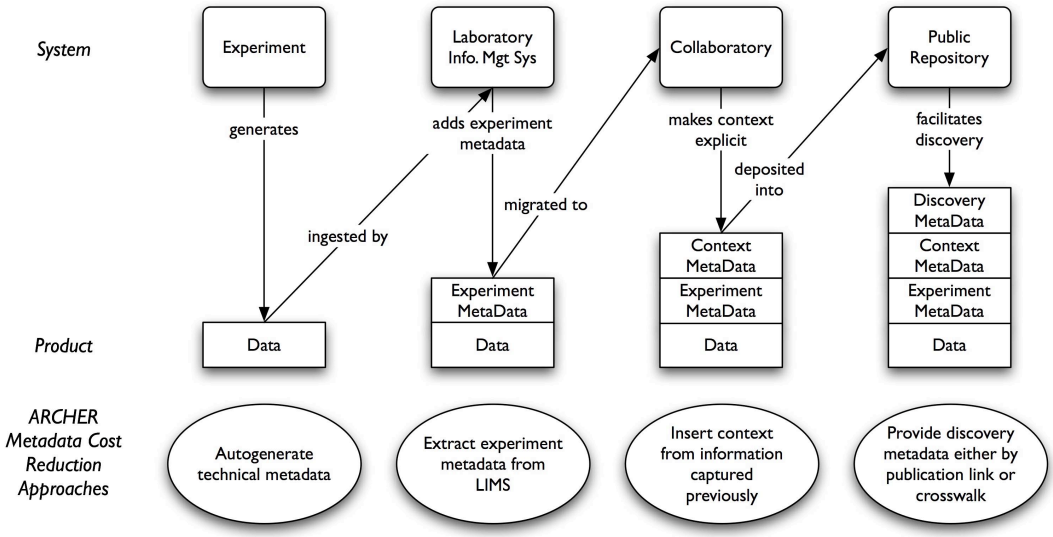

Figure 1: Data and Metadata Flows

solving. Setup of the first data set took between at least 24 hours to complete over a calendar week (and weekend). Once ftp access was provided to the content it took about two hours to upload and prepare the metadata per record. Most of the effort was in trying to keep a mental map of the many different data-streams straight and then double checking work several times to be sure that the complex structure was complete.

Procedures are now being put in place that will allow the researchers themselves to undertake much of the work of lodging the dataset objects, with the library staff performing more of a quality control and authorisation function. Under this approach, the researchers will provide the quality control over the technical metadata and the library staff will review (and augment) the descriptive metadata. It is expected that it should take between 5 and 15 mins to approve submissions that have no obvious dc metadata issues and where the number of compressed file parts match a count. Library Staff will be unable to verify CCLRC metadata or the data in the compressed files. Problem files will be returned to the submitter for correction and resubmission.

Fig. 1 shows a schematic of how this process was applied in ARCHER. The systems at the top are those that generate or manage the research outputs. The collaboratory is a collaborative environment which supports a defined group of fellow researchers who are sharing access to a set of data. The system needs to make explicit to this group the context for the data (which may be implicit in either the laboratory information management system or the culture of the research team that generated the data).

The products are the outputs of the systems and/or the objects that they manage. Note that at each migration stage, the product from the previous stage is augmented with additional information in the form of metadata relevant at that stage.

The diagonal flows represent processes that move data between systems, augmenting the metadata at each stage, and (by implication) making a selection of which objects to migrate [19]. The metadata cost-reduction bubbles show possible approaches to shift or ameliorate the cost of generating the metadata at each stage.

8. Conclusions

We have argued that in approaching metadata capture and management for research data that it is important to recognise significant differences in both the nature of the information, and the nature of the needs of researchers when compared to publication repositories. We have seen that substantially different access methods are required to support these needs, and that metadata is needed for these varying access methods. However creators of research data also have needs that must be satisfied, and this can be looked at from a scholarly communication approach.

Having described the needs, we turned our attention to costs. These are very substantial and choices have to be made; which needs will be met, and who incurs the costs? Importantly the people who are best placed to capture metadata are often not the beneficiaries of the value of that metadata. We thus discussed how costs can be both transferred and reduced.

Finally we gave an example of both cost shifting and cost reduction in crystallography where metadata

was captured that would meet both evidenciary and discovery needs.

It is clear that a range of approaches are available to reduce metadata costs. The cost that is borne elsewhere (in writing a publication) can be used to provide very low-cost discovery metadata by linking to the publication, but in a way where certain types of discovery (search by chemical structure rather than name) will be more expensive. In a different approach the cost of capturing experiment and context metadata can be transferred from a metadata expert to a one-off software development cost plus a small amount of researcher time per dataset. Not only will costs be reduced for the data capture phase, but higher quality metadata will be captured.

We believe that the growth in data-intensive research, and the ability of instruments and sensors to generate increasing data volumes will require the use of these sorts of innovative techniques to avoid a potential metadata bottleneck.

9. References

[1] American Council of Learned Societies Commission on CyberInfrastructure for the Humanities and Social Sciences, *Our Cultural Commonwealth*. Retrieved July 25, 2008 from http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf, 2006

[2] Association of Research Libraries, *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering*, Association of Research Libraries, 2006. www.arl.org/pp/access/nsfworkshop.shtml

[3] Beagrie, N., Chruszcz J. & Lavoie, B., "Keeping Research Data Safe, a cost model and guidance for UK universities", JISC, 2008, Retrieved from www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.doc on July 25th, 2008

[4] Bearman, D. *Item level control and electronic recordkeeping*, Archives and Museum Informatics, Volume 10, Number 3 / December, 1996

[5] Borgman, C. Scholarship in the Digital Age: Information, Infrastructure and the Internet. MIT Press, Cambridge, MA., 2007

[6] DeLone, W. H. & McLean, E. R.. Information Systems Success: The Quest for the Dependent Variable. *Information Systems Research*. Volume 3, Issue 1,60-96. 1992

[7] Hunolt, G., *Technical Description Document Cost Estimation Toolkit (CET)*, version 2.1 September 2006. Retrieved 3/1/08, from http://opensource.gsfc.nasa.gov/projects/CET/Doc.zip

[8] Hunter, J., "Harvesting community tags and annotations to augment institutional repository metadata", *Proceedings of eResearch Australasia*, 2007

[9] Ingwersen, P. and Järvelin, K. *The Turn: Integration of Information Seeking and Retrieval in Context* (The Information Retrieval Series). Springer-Verlag New York, Inc., 2007.

[10] ISO 2146 Information and Documentation - Registry Services for Libraries and Related Organisations (ISO TC46 SC4 Working Draft, 13 December 2005). Retrieved July 30, 2008, from http://www.nla.gov.au/wgroups/ISO2146/n197.doc

[11] Lavoie, B., *The Open Archival Information System Reference Model: Introductory Guide*, DPC Technology Watch Series Report 04-01 January 2004. Retrieved 3/1/08 from http://www.dpconline.org/docs/lavoie_OAIS.pdf

[12] McMullen, D.F. and Huffman, K., "Connecting users to instruments and sensors: portals as multi-user GUIs for instrument and sensor facilities", *Concurrency Computat.: Pract. Exper.* 2006; 18:1-11.

[13] Moore, R. "Integrating Data and Information Management", *International Supercomputer Conference,* June, 2004. Available online at http://www.sdsc.edu/dice/Pubs/ISC2004.doc

[14] Murray-Rust, P. "Data-driven science - a scientist's view", *NSF/JISC Repositories Workshop*, April 10, 2007

[15] National Science Board, *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century,* National Science Foundation, 2005. www.nsf.gov/pubs/2005/nsb0540/start.jsp

[16] Research Information Network, *"To share or not to share: Publication and Quality Assurance of Research data Outputs",* 2008, Retrieved July 25, 2008 from http://www.rin.ac.uk/files/Data%20publication%20report,%20annex%20-%20final.pdf

[17] Rosado, C. J. et al. (2007), "A Common Fold Mediates Vertebrate Defense and Bacterial Attack", *Science Express*, August 23 2007. Science DOI: 10.1126/science.1144706

[18] Roosendaal, H., and Geurts, P., "Forces and functions in scientific communication: an analysis of their interplay". *Cooperative Research Information Systems in Physics*, August 31—September 4 1997, Oldenburg, Germany. http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html .

[19] Treloar, A. and Harboe-Ree, C. (2008). "Data management and the curation continuum: how the Monash experience is informing repository relationships". *Proceedings of VALA 2008*, Melbourne, February, 2008. http://www.valaconf.org.au/vala2008/ papers2008/111_Treloar_Final.pdf

[20] Van de Sompel at al., "Rethinking Scholarly Communication: Building the System that Scholars Deserve", *DLib* September, 2004. doi:10.1045/ september2004-vandesompel