Research data and repository metadata. Policy and technical issues at the University of Sydney Library.

Research data and repository metadata. Policy and technical issues at the University of Sydney Library.

Rowan Brownlee
Digital Project Analyst
University of Sydney Library
r.brownlee@library.usyd.edu.au
http://escholarship.library.usyd.edu.au/dpa/

Abstract
The University of Sydney Library's repository contains research outputs primarily comprising traditional publication types. Many academics manage data collections within databases and spreadsheets using metadata dissimilar to the repository's Dublin Core schema. During 2007 and 2008 the author explored issues surrounding submission of a small range of research data collections and associated metadata. Native metadata structures were analysed and mapped to DC and scripts translated, packaged and transferred collections. This paper discusses metadata management and repository service levels and sustainability. It describes the Library's approach to defining service requirements and includes discussion of various metadata management options. It also describes related activities within the University of Sydney to develop eResearch services and to harmonise the roles and relationships of eResearch support service providers[1].

Keywords
research data collections, metadata translation, batch processing, repository service levels, repository service sustainability, repository service policy development.

Introduction
The number of discrete data collections created and managed within academic units at the University of Sydney is unknown, but can be reasonably assumed to be very large[2]. Also unknown is the range of formats, applications and tools used to capture, manage, render and manipulate the data[3]. The University of Sydney Library supports research, learning and teaching through a variety of initiatives and collaborative activities with academics. Some activities concern data collections which are intended to be long-lived. During 2007 and 2008 the author sought to explore issues regarding management of research data within the Library's repository by examining collections related to several partnership activities. The aim was to develop guidelines to support a consistent and sustainable approach to dealing with requests to manage these types of materials within the repository[4]. No selection criteria were adopted with regard to the collections under discussion. It was generally the case that the author had been working with particular academics who were interested in ensuring ongoing access to their data or that of their department.

Description of data collections

The collections were created by an individual academic, research project team, or a group of academics working within a discipline. Academics represented within the projects under discussion are aware of the facility of descriptive metadata for categorising and interrogating datasets. They adopt or modify domain standards or create rich and often highly granular tag sets to suit project requirements. They are less aware of technical or preservation metadata. The collections are not large, generally in the range of tens or hundreds of gigabytes. Metadata is typically held in databases including Filemaker and MySQL or spreadsheet applications such as Microsoft Excel, with associated data objects housed on personal computer or departmental file systems. Collections under discussion arise from the School of Geosciences, Sydney College of the Arts, Department of Archaeology and School of Biological Sciences

School of Geosciences theses and datasets

Administrative records are contained within a spreadsheet. A thesis may contain hundreds of associated data files comprising images, spreadsheets, text, and video. There may also be data hosted on GIS systems. Metadata is minimal and readily mapped to DC.

Sydney College of the Arts research archive

The research archive is a small but growing component of SCA Images Online[5], an image management service developed by Jacqueline Spedding through a partnership with the University Library and the Power Institute Visual Resources Library[6]. Metadata is based on VRA Core 3.0[7] and the collection is managed through Filemaker and presented using MDID2[8].

Sarah Colley archaeological fish-bone collection

Comprising digitised images of fish bone specimens, metadata includes a taxonomy developed by Sarah to support her research. Identification of such tiny specimens is very difficult and future metadata developments may involve the incorporation of tags from taxonomies used for describing shapes. Sarah's collection management system incorporates Filemaker.

eBot plant sciences collection

eBot is being developed as a cross-faculty database of plant sciences objects to support research, learning and teaching and represents collaboration between the School of Biological Sciences, the Library and the Faculty of Agriculture, Food and Natural Resources. Metadata is based on HISPID[9] and includes preservation metadata recommended by the National Library of New Zealand Metadata Standards Framework[10].

University Library eScholarship DSpace[11] repository

The University of Sydney Library's repository has been in operation for several years offering a digital preservation service for research outputs. Managed by Sten Christensen[12], the service provides open access to approximately 2000 items, and

during the month of June 2008 recorded 32,000 item views. A key aim of service development has been to position the repository within the University as a critical infrastructure component, with a view to eventually seeing routine submission of research outputs across the institution. Through partnership with the University Research Office and government research assessment initiatives, service profile is growing. Schools and faculties are increasingly initiating contact, and the growing momentum regarding open access to research outputs has seen an improving rate of service uptake[13].

Repository metadata

DSpace offers Dublin Core (DC) as a default descriptive metadata schema and with the exception of a small number of additional qualifiers DC is largely unaltered within the Library's implementation. DC is suitable for bibliographic description of most items, as the collection mainly comprises traditional publication formats such as research articles and conference papers.

Defining metadata management requirements for data collections within the repository

The first step saw consideration of general requirements for a service to enable management within the repository of research data records and their associated objects. Four particular concerns were identified.

- Retain the granularity of the native record.
- Enable export, including Open Archives Initiative (OAI) harvesting, of records in DC and native format[14].
- Enable development of schema-specific search interfaces, whether through repository tools or integration with other services.
- Ensure service sustainability.

The activity did not seek to address the distinct though related issue of relational database preservation. The focus was instead on metadata management issues surrounding incorporation of records and objects supplied by other data management services.

Considering options for metadata management

Four approaches to metadata management were identified and are described below.

1. Map native metadata to existing DC elements.

Native metadata records are mapped to DC and transferred to the repository as standard DC records. Native metadata records are not retained within the repository.

Positives

- Relatively low submission cost and low ongoing maintenance cost
- Requires no configuration or maintenance of DSpace index keys, customised metadata schemas or OAI crosswalks.

- Records would be fully searchable through default DC indexing and harvestable via default OAI.

Issues

- Loss of metadata granularity and inability to recreate the original records.
- Many items of metadata would not be meaningful without contextual information provided by their native tags.
- Does not support provision of a traditional field-based advanced search reflective of the granularity of the original records.

2. Map native metadata to DC elements and create new custom qualifiers for standard DC tags

Native metadata records are mapped to DC and transferred to the repository as standard DC records. The granularity of non-DC elements is retained through mapping to customised qualifiers of standard DC tags.

Positives

- Retains the granularity of the native records, supporting recreation of the original metadata records. Also retains contextual information conveyed by the original tags.
- Requires no configuration or maintenance of DSpace index keys, customised metadata schemas or OAI crosswalks.
- Records would be fully searchable via default DC indexing and harvestable via default OAI.

Issues

- Higher submission and maintenance costs than option 1, requiring additional and ongoing recordkeeping and maintenance procedures.
- As DC qualifiers proliferate, management of the central registry may pose challenges.

3. Create a custom schema identical to the native metadata set

A custom schema separate to DC is implemented within the repository. Metadata records are transferred to the repository in their native format.

Positives

- Avoids the DC registry management problems of option 2, by enabling partitioning and separate maintenance of each custom schema.
- May enable future provision of a collection, community or schema-level traditional field-based advanced search reflective of the granularity of the original records[15].

Issues

- Requires configuration and ongoing maintenance of DSpace index keys, customised metadata schemas and OAI crosswalks.

- May result in a proliferation of project-specific schemas requiring accompanying recordkeeping and maintenance.
- Will not assist in the management of hierarchical metadata schemas, as these are not supported by DSpace.

4. Generate DC records as abstractions of the native metadata records and submit the native metadata records as digital object bit-streams.
DC records act as bibliographic descriptions of the native metadata records. The original records are submitted as accompanying bit-streams.

Positives
- Relatively low submission cost and low ongoing maintenance cost
- Requires no configuration or maintenance of DSpace index keys, customised metadata schemas or OAI crosswalks.
- Depending on how much of the original metadata is mapped to standard DC, records could be keyword searchable via default DC indexing.
- DC versions of the records would be harvestable via default OAI.
- Avoids the DC registry management problems of option 2 and the schema proliferation issues of option 3.
- Retains the original metadata records in their native format.

Issues
- Would not support future provision of a collection, community or schema-level traditional field-based advanced search reflective of the granularity of the original records. Would require indexing of the accompanying native metadata file
- Would not readily enable harvesting of native metadata records.

Discussion of approach selected
The Library's adoption of option 4 as a general guideline was informed by policy regarding the purpose of the repository service and capacity to maintain agreed service levels. Option 1 is least likely to satisfy requirements for preservation and reuse of research data metadata. Option 3 would be the most expensive, though it may support the greatest level of interactivity and flexibility regarding presentation. Option 4 might allow the least flexibility regarding user interactivity, though even this is unclear as there are uncertainties regarding how best to transfer, support and reflect degrees of functionality offered by native collection management systems, within the repository service. Option 4 is however coherent with the repository's primary preservation function and is likely to make least additional demands on resources. Although the service will continue to review the feasibility of implementing a set of domain-specific schemas, it is doubtful that current funding would enable a guarantee of ongoing maintenance of multiple schemas.

Metadata mapping

Metadata from the source databases was mapped to DC to enable simple keyword searching within DSpace and DC-based OAI harvesting. In the case of the SCA collection, VRA to DC mapping was guided by Tony Green[16], Visual Resources Librarian at the Power Institute. For the fish-bones and plant sciences collections, most fields were mapped to dc.description and dc.subject.

Metadata transfer

Two of the collections are managed by their owners using Filemaker database software. Records were exported from Filemaker as CSV files, each record comprising a row in the file. The author created a Python[17] script which wrote each row to two files. One was a DC XML file and the other a native metadata file. The script also packaged the metadata and associated data files in a format suitable for submission to DSpace using the repository's itemImport utility. A selection of records were manually sampled and compared and additional scripting ensured that all records were correctly transferred.

The mapping of many discrete fields to dc.description and dc.subject threatened a loss of important contextual information provided by field labels. This was mitigated by the incorporation of scripting instructions to prefix each data element with its source fieldname supplied by the original database. The outcome was a searchable DC record within DSpace and submission of associated granular native metadata records. (A less desirable outcome was the inclusion of contextual prefixes within the repository search index).

Figure 1. DSpace record from Sarah Colley's fish-bones collection, illustrating the addition of field labels within dc.description.

Title:	Collection: ANU. Fish taxon code: CAR-SE-GR. Fish anatomy code: PMX.
Authors:	Russell Workman
Keywords:	Carangidae -- Seriola -- grandis (Yellowtail Kingfish (2)) | CAR-SE-GR Trevallies | CAR-ALL
Issue Date:	26-Sep-2008
Description:	Fish specimen origin: Modern reference specimen Fish anatomy label: Premaxilla | PMX Collection name: Department of Archaeology and Natural History, Research School of Pacific and Asian Studies, Australian National University, Canberra. (ANU) Fish anatomy view: two views Source database record number: 200060 Collection Accession ID01: 284 Collection Accession ID02: 35

Figure 2. The same information in a more granular native XML format.


```
<fishbone_record>
<image_reference_number>200060</image_reference_number>
<image_date_photo_taken>1/06/2007</image_date_photo_taken>
<image_photographer>Russell Workman</image_photographer>
<collection_name_code>ANU</collection_name_code>
<collection_name_label>Department of Archaeology and Natural History, Research School of Pacific and Asian Studies, Australian National
<collection_accession_ID_01>284</collection_accession_ID_01>
<collection_accession_ID_02>35</collection_accession_ID_02>
<fish_specimen_origin>Modern reference specimen</fish_specimen_origin>
<fish_taxon_code>CAR-SE-GR</fish_taxon_code>
<fish_taxon_family>Carangidae</fish_taxon_family>
<fish_taxon_genus>Seriola</fish_taxon_genus>
<fish_taxon_species>grandis</fish_taxon_species>
<fish_taxon_common_name>Yellowtail Kingfish (2)</fish_taxon_common_name>
<fish_taxon_group_code>CAR-ALL</fish_taxon_group_code>
<fish_taxon_group_label>Trevallies</fish_taxon_group_label>
<fish_anatomy_code>PMX</fish_anatomy_code>
<fish_anatomy_label>Premaxilla</fish_anatomy_label>
<fish_anatomy_view>two views</fish_anatomy_view>
</fishbone_record>
```


Outcomes and future directions

These activities provided an opportunity for the Library to consider issues regarding metadata management of non-DC collections within the repository. Development of a general guideline was informed by an understanding of the fundamental preservation requirements of the repository and capacity to ensure service sustainability over time. The selected approach captures DC and native metadata in separate files and future activity may see the use of METS[18] or OAI-ORE[19] to relate categories of metadata and associated objects to each other.

Although on a small-scale and dealing with a limited range of collections, the experience of working with academics on data management activities has highlighted a need for eResearch support services. Beyond discipline-specific requirements for customised data interrogation, manipulation and presentation tools, there may be a common need for services enabling submission and user-defined structured description of research data collections. Such services would provide tools enabling academics to securely share data with nominated colleagues, and may also capture administrative metadata to support systematic transfer of collections to local and/or remote preservation services.

Activities described in this paper commenced prior to current University of Sydney initiatives to develop institutional eResearch support services incorporating research data management. The Library is a partner in a University of Sydney Information Communications Technology Services (ICT) sponsored project lead by Jim Richardson[20] to explore tools and frameworks for eResearch support. This includes examination of options for providing data storage services for researchers. One model for consideration is that of the Large Research Data Storage service (LaRDS), established at Monash University[21]. Chris Rusbridge has written of a need to bring the repository upstream, to make it an integrated component of researcher workflows[22]. The author believes that future service development at the University of Sydney will see a repository service located within a content management workflow relating volatile researcher workspaces with archiving services.

There are challenges for the University of Sydney Library in considering data management roles. Beyond technology and policy, the author believes that data management requires knowledge of data and associated documentation standards of a type described by Alma Swan in her classification of a data scientist[23]. Although the Library has expertise in particular types of content, associated documentation standards and access mechanisms, the primary focus is on delivery of broad-based services rather than specialist infrastructure. Cataloguing staff know a great deal about a particular documentation standard and its relatives, and content within the repository and the library system is of a format suitable for description by these known standards. In addition most content has already been subject to an external review process from a trusted agency such as a publisher or academic examination board. Library services rest on maintaining investment in particular capabilities while leveraging trusted relationships with external content providers and reviewing agencies.

The Library has several librarians with research degrees who possess a high level of domain knowledge, who contribute to collaborative academic projects and who guide librarians possessing lesser levels of subject knowledge. These scholar librarians could become data scientists within their field and might play a key role within institutional domain-specific data management and digital preservation services. More generally University of Sydney Library involvement in data management or digital preservation services will rely on partnerships with practitioners who have the required subject-based data management skills. These partnerships will require institutional backing to ensure persistence over time. The ability to support an archiving service assumes that requirements are understood regarding documentation and data integrity, and that access to this knowledge will be maintained and developed for the expected lifespan of the submitted content.

Another focus for the Library is the value-adding of metadata through integration of repository services with flexible presentation, interrogation and auditing tools[24]. The author is currently migrating a copy of the eBot research data image collection to an XML platform incorporating the California Digital Library's eXtensible Text Framework (XTF)[25]. The collection uses a very rich metadata set underpinned by an extensive taxonomy and XTF offers excellent support for precise metadata indexing and rendering of schema-specific search and display interfaces. A future aim of the migration is to leverage the collection's existing taxonomy to enable online presentation of fully navigable hyperlinked pathways.

The Library has access to a range of platforms, tools and services through which it may publish collections. Selection and integration of services is challenging, but is informed by collaboration and communication with client groups to understand their requirements. In partnership with other University agencies such as the Research Office, ICT, and Archives and Records Management, and within the context of the Australian National Data Service[26] and the Intersect eResearch Support Institute[27], the Library is seeking to understand requirements for eResearch support services in general and information and

data management and preservation in particular.

---

1 The current paper builds on a discussion document created for the University of Sydney Library titled “Research Data Management and Repository Metadata”, March 2008, http://escholarship.library.usyd.edu.au/dpa/meta.html

2 Abed Kassis, Arts Faculty Web and Information Systems Manager estimates thousands of Filemaker databases within that Faculty alone.

3 Edwina Tanner, Associate Lecturer School of Geosciences, has recently secured funding from the Australian Partnership for Sustainable Repositories (http://www.apsr.edu.au) to develop a data auditing tool. Activities will be informed by the JISC Data Audit Framework Development Project (http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/dataauditframework.aspx)

4 This was intended to provide a supplemental preservation service which did not seek to match or replace the functionality of tools an academic may use on their desktop or through a departmental data management service.

5 SCA Images Online, http://www.usyd.edu.au/sca/learning_teaching/projects/sca_images_online.shtml

6 Power Visual Resources Library, http://www.arts.usyd.edu.au/centres/power/?page=Power%20Library

7 VRA Core 3.0, http://www.vraweb.org/resources/datastandards/vracore3/index.html

8 Power Visual Resources Library MDID2 installation, http://mdid.arts.usyd.edu.au/

9 HISPID is a standard format for the interchange of electronic herbarium specimen information. http://plantnet.rbgsyd.nsw.gov.au/HISCOM/HISPID/HISPID3/hispidright.html

10 Metadata Standards Framework – Preservation Metadata (revised), http://www.natlib.govt.nz/catalogues/library-documents/preservation-metadata-revised

11 DSpace, http://www.dspace.org

12 Sten Christensen, Repository Coordinator, University of Sydney Library, http://ses.library.usyd.edu.au/

13 For further information regarding integration of repositories with research support systems at the University of Sydney, see Sten Christensen “From Research Management System to Digital Repository : Managing and Storing Research Outputs at the University of Sydney” (poster presented at eResearch Australasia, September 29-October 3, 2008, http://www.eresearch.edu.au/posters)

14 Throughout this paper, the term ‘native format’ is used in reference to the metadata structure adopted by a researcher within their local database.

15 DSpace technologies such as Manakin may support the development of customised collection or community-based interfaces offering greater interactivity and drawing upon customised metadata. Integration with external services may also offer enhanced user interface tools. The author has not investigated these areas.

16 Tony Green, http://www.arts.usyd.edu.au/centres/power/?page=staff

17 Python was chosen as it is familiar to the author and offers excellent text processing facilities. http://www.python.org/

18 Metadata Encoding & Transmission Standard, http://www.loc.gov/standards/mets/

19 Open Archives Initiative Object Reuse and Exchange, http://www.openarchives.org/ore/

20 Dr. Jim Richardson, University of Sydney Information Communications and Technology eResearch relationship manager, http://www.usyd.edu.au/ict/relation/contacts.shtml

21 Monash University Large Research Data Storage (LaRDS), http://www.monash.edu.au/eresearch/services/lards/

22 Chris Rusbridge,”Moving the repository upstream” (paper presented at the ARROW Repositories Day, Brisbane, Australia, October 14, 2008, http://www.arrow.edu.au/news/event2008program.php)

23 Alma Swan and Sheridan Brown 2008, “The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs”, http://eprints.ecs.soton.ac.uk/16675/

24 In particular, partnerships with APSR on development of a DSpace Open Journal System connector, harvesting of collection metadata for a research collections registry and implementation of improved statistical tools for measuring repository content use. http://www.apsr.edu.au

25 Development version of eBot-XTF online at http://pictor.library.usyd.edu.au:8080/xtf-2.1/search California Digital Library eXtensible Text Framework, http://cdlib.org/inside/projects/xtf/
26 Australian National Data Service, http://ands.org.au/
27 Intersect eResearch Support Institute, http://www.intersect.org.au/