An Innovative Shared Preservation Service: The SHERPA-DP Project: Securing a Hybrid Environment for Research Preservation and Access - Digital Preservation

Andrew Wilson, Arts and Humanities Data Service, London UK

Abstract

This paper will provide insights into an innovative preservation environment being developed by the SHERPA-DP project. SHERPA-DP, led by the UK Arts and Humanities Data Service (AHDS), builds on the work of the original SHERPA project and aims to create a collaborative, shared preservation environment for e-prints repositories framed around the OAIS Reference Model. The project brings together a number of academic institutional repository systems with the existing preservation repository established by AHDS, to create an environment that fully addresses the requirements of the different phases within the life cycle of digital information.

Introduction

I will begin with a brief mention of the Open Archival Information System Reference Model (OAIS)[1], ISO 14721, but I do not intend to spend much time on details of the standard. I will assume that most of my audience is, if not actually familiar with OAIS, aware of the framework model, and I don't propose to discuss it in any more depth here. I do, however, want to bring your attention to some statements in that document about the scope and aims of the reference model which are relevant. The Reference Model says at the outset that it is "designed as a conceptual framework in which to discuss and compare archives" (p. 1-3). It sets out a high-level model for establishing an "organisation of people and systems" which has the responsibility to preserve digital resources and make them available (p. 1-1). While the model identifies the high-level activities necessary for an archive to be a conforming open archives information system, it explicitly does not "specify a design or an implementation. Actual implementations may group or break out functionality differently" (p. 1-2). To reinforce the message, I is clear from the OAIS model below that what the standard sets out is a basic framework for modelling all the activities needed to capture, maintain, and make accessible digital resources over time. Nothing about real-life implementation is implied by the diagram or by the model itself.

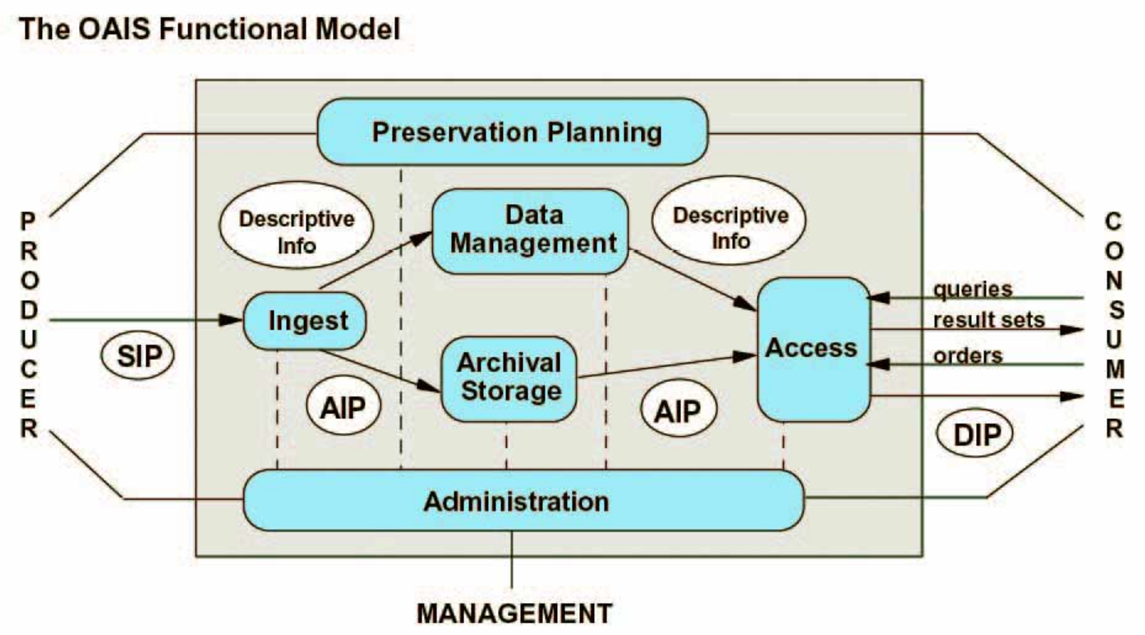

***Figure 1** OAIS Functional Model*

I've provided the quotes above because they indicate very well the flexibility inherent in the OAIS reference model, and the unwillingness of the authors to constrain implementers through specifying implementation approaches. This is all to the good and leads directly to the Sherpa-DP project, although the connection will not be made clear until later in this paper, after some more general remarks.

Project Background

It is apparent that within the library and archive community there has been growing an awareness that responsibility for managing information held in digital formats will need to be distributed in new ways. Monolithic structures and approaches do not necessarily satisfy the needs of small institutional repositories, and in some cases will not provide the best strategies even for large public institutions.[2] A way forward recently envisioned and becoming more articulated in the literature is to disaggregate the tasks undertaken by a digital repository, so that not all digital repositories need undertake all tasks implied by the OAIS framework.

This separation of various components of the archival service should allow for distributed open access content repositories to be maintained independently of value-added services, which can be provided discretely by multiple service providers.[3]

In a mixed environment of non-standardised institutional repositories, focusing on the easy submission and dissemination of resources, and archival repositories with an emphasis on long-term retention, not all repositories will be able to, or necessarily want to, develop a full set of value-added services. Instead, repositories may find it more efficient and cost-effective to seek access to value added services through collaborative arrangements, or by sub-contracting to specialist service providers.

Digital preservation is a vital 'value-added' service, but one that, in the present context of institutional repository developments in the UK at least (for scholarly communications, e-learning or other areas), few institutions are well equipped to deal with alone. As observed in the UK Joint Information Systems Committee (JISC) funding call 4/04, and earlier by Neil Beagrie, coordinated effort is likely to be needed in order to ensure the preservation of the increasing volume of digital material held by UK institutions of Higher and Further education.[4]

Much of the research to date has focused on establishing requirements for digital preservation, and models for, and implementations, incorporating digital preservation into a single repository. Even where research does deal with distributed content, as in the Preservation Risk Management for Web Resources research (http://www.library.cornell.edu/iris/research/prism/prism-research.html), part of the Cornell University Project Prism (http://www.library.cornell.edu/iris/research/prism/index.html) project, and the JISC funded Hybrid Archives Project (part of the Focus on Access to Institutional Resources (FAIR) Programme) undertaken by the AHDS in 2002-05, archival tasks remain centralised. It seems far more likely that multiple archives, libraries and IT centres will emerge in the future, offering a heterogeneous array of preservation services. By way of a simple example, an institution may rely on The National Archives (UK) PRONOM database to check files for obsolescence, but if files are found to be at risk, it will then need to pass them to a separate migration service for remedial preservation action.

Existing work relevant to the distributed management of digital material is mainly concerned with design of interoperable systems for the discovery and delivery of digital material. Conceptual frameworks such as the architecture for the Joint Information Systems Committee (JISC) Information Environment (http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/), and the SCORM (Sharable Content Object Reference Model) for e-learning have emerged to tie together work on the development of common standards and profiles for the use of standards, software tools, shared terminology, and the identification of roles and functions.

No comparable framework for the distributed management of digital preservation services yet exists. However, many of the elements of such a framework already exist, in the outcomes of many other research projects both in Europe and North America. What is needed now is a *distributed preservation model* that draws existing research and practical work into a coherent structure for disaggregating the framework OAIS model, and which is viable for often inadequately funded UK HE institutional repositories.

Institutional repositories are still a relatively new and high profile area in the UK, often championed as the way forward for making the research outputs of Higher Education available to a wider public. In recognition of this the JISC funded the establishment of a number of institutional repositories in the UK as part of its 2002 FAIR Programme. The initial focus of activity under this program was on the process of establishing institutional repositories – installing appropriate software and establishing policies and procedures; encouraging deposit of articles and dealing with the associated rights issues; and working to effect the cultural change needed for the successful development and population of repositories.

Given the experimental and project-based nature of much of this activity, it is not surprising that so far less attention has been paid to preservation, and that no UK institutional repository established to date in the HE sector would claim to be 'doing preservation'. Of course, the original Securing a Hybrid Environment for Research Preservation and Access (SHERPA) project, funded under the FAIR program mentioned above, of which Sherpa-DP is a direct successor, had a specific remit to investigate the requirements for preservation and produced some valuable outputs. However, this was a secondary aim of the project and has not resulted in the establishment of any coherent and long-term preservation regime for the institutional repositories involved in the project.

A recent JISC-funded Feasibility and Requirements Study for Preservation of E-Prints argued that there is a unique window of opportunity to address the preservation requirements of repositories at the beginning of their adoption rather than leaving it until the lack of preservation management becomes an issue, and content is no longer accessible.[5] The study noted that the scarcity of staff and services with practical digital preservation skills and expertise made it difficult for HE institutional repositories to also manage and run a preservation environment based upon the OAIS Reference Model. The study suggested that a sensible way forward would be to look to disaggregate the functions and activities identified in the OAIS Model, and to seek collaborative arrangements between repositories and specialist services with each taking responsibility for different functions.

The Sherpa-DP Project

When in 2004, the Joint Information Systems Committee (JISC) program called for projects for funding under its Supporting Digital Preservation and Asset Management in Institutions program, the opportunity arose to address an identified need to provide institutions in the UK Higher Education sector (HE) with practical support in effective digital preservation and asset management, and to ensure the ongoing availability and future accessibility of digital information of value to the UK Higher Education community. The AHDS managed Sherpa-DP project is funded under the so-called 4/04 program and evolved quite naturally from an earlier Sherpa project, also JISC funded. This original Sherpa project had the specific aim of developing a number of academic institutional e-prints repositories, and initiating the populating of the repositories with content. Sherpa was a collaborative project lead by the University of Nottingham, and involving 17 other UK Universities, the British Library and the AHDS.

The Sherpa-DP project is a logical extension of the original Sherpa project and aims to advance research through a study of the practical implementation of the recommendations made in the Feasibility study cited above. By extending collaboration into a full preservation service the project's overall purpose is to remove from each individual institutional repository the burden of adding a preservation layer to its repository services. Sherpa-DP is a collaborative project and involves a subset of the original Sherpa partners: the universities of Nottingham, Glasgow, Edinburgh, the White Rose Consortium (Leeds, Sheffield, York), and the London Leap Consortium (University College London, Imperial, Birkbeck, Kings College, Royal Holloway, and the School of Oriental and African Studies). Work on the project started in March 2005 and is due for completion by the end of February 2007.

The primary aims of Sherpa-DP are to:

- develop a prototype preservation environment for the Sherpa-DP partners based on the OAIS reference model, including shared protocols and software tools;
- establish a comprehensive workflow and set of procedures to suit the needs of institutional repositories and the preservation service;
- provide guidance on the ingest process in order to encourage the deposit of file formats that will minimise long-term operational costs and maximise preservation potential;
- develop an exemplar for an outsourced preservation service;
- create a digital repositories handbook that will set out best practice standards and processes for resource creation and ingest, preservation planning and management, and provision of access for the holdings of institutional e-print repositories in the UK.

In essence the project will deliver an implementation of a distributed preservation service for use by the partner institutional repositories whose holdings are primarily e-prints in both pre- and post-print form. The project will investigate the business case for the proposed disaggregated model and will seek to establish an economic cost model that could be used to ensure its long-term sustainability. Although development of the exact nature of the model that will be adopted is underway, the final form of the service will not be finalised until later in 2006.

The first question addressed at the outset is the most obvious: why disaggregate preservation from the other components of the OAIS reference model? Through development of a conceptual paper that set out the variation of the OAIS model that Sherpa-DP will seek to implement, we began to understand some of the issues and barriers that could encourage UK institutional repositories to seek outside providers for supply of components of the OAIS model. These include such considerations as: institutional repositories lack the time to implement preservation; there is, certainly in the UK and probably elsewhere, a scarcity of staff with the necessary digital preservation skills and expertise; to remove inefficiencies involved in duplication of services; there is a potential cost saving in terms of staff time and equipment; and because preservation is not an inherent functionality in most repository software there will be a significant extra resource burden involved in adding it to existing institutional repositories.

Development of this conceptual paper (available at: http://ahds.ac.uk/about/projects/sherpa-dp/), established that the OAIS model is flexible enough to support the disaggregated model we had been considering. Hence the specific quotes provided in the Introduction which indicate that the developers of the OAIS reference model had no intention of being prescriptive about how implementers should apply the model in real life situations. Our conceptualisation of the variant model being developed by Sherpa-DP is shown in figure 2 below.

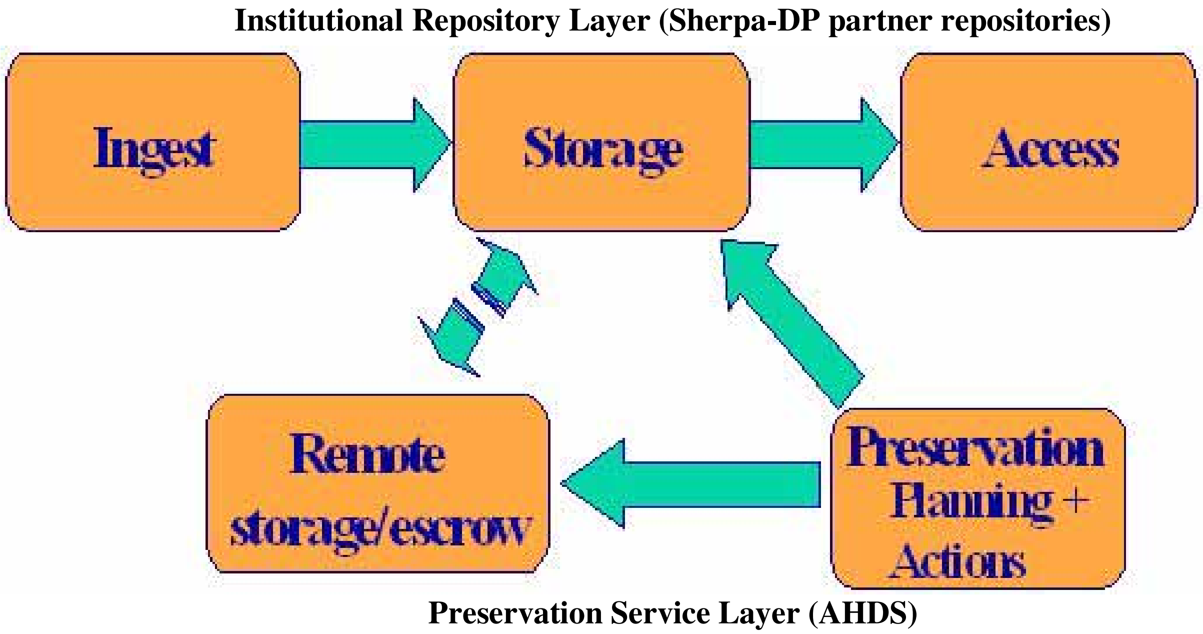
Institutional Repository Layer (Sherpa-DP partner repositories)
Ingest
Storage
Access
Remote storage/escrow
Preservation Planning + Actions
Preservation Service Layer (AHDS)

Figure 2. *The Sherpa-DP conceptual model*

A companion paper to that referred to above, and available at the same URL, sets out in more detail the various responsibilities of the institutional repository and the AHDS preservation service. Put simply, the institutional repositories will be responsible for ingest, access and depositor relations (as they are now), the AHDS will be responsible for the mechanisms for transferring data and metadata in both directions between the IRS and the AHDS, and will undertake preservation planning and all preservation actions. Mechanisms will be established to allow retrieval of preservation copies if and when they are needed by the institutional repository to replace corrupt or deleted content, and repository requests for retrievals will be handled by a semi-automated process.

The major challenge of the project will be to implement the disaggregated model successfully with the different repository software solutions chosen by the institutional repositories involved, taking into account the individual policies and approaches in regard to content and metadata. So one of the first set of work activities almost completed, was to investigate the repository landscape, ie. the technical infrastructure of partner repositories. Sherpa-DP is especially interested in exploring the use of open source software and standards to implement the preservation environment, including XML, Fedora, DSpace, METS, OAI and, possibly, grid technologies. To this end the project has investigated and implement automated networked transfers of test data and metadata between DSpace/Eprints systems, used to set-up the great majority of institutional repositories in the UK, and the AHDS preservation repository which is moving to implement of Fedora. The aim of this work is to enable automatic synchronisation of data and metadata resources with a remote preservation repository in order to enable resources to be preserved and maintained within an OAIS framework. Solutions to this problem at several different functional levels will be investigated. As a corollary, the project will investigate the use of METS as the framework for combining and packaging metadata, and as a possible transfer mechanism for both metadata and e-print content.

The most common repository applications, DSpace and E-Prints (one partner repository uses DSpace, all the others use E-Prints) have been analysed for necessary functionality and for possible options for extending functionality to incorporate the mechanisms necessary to allow the operation of a remote preservation service. This work has included research on transfer mechanisms between institutional repositories and the preservation service – analysis of DSpace and Eprints APIs, storage layers and module add-on capabilities, management of versioning and synchronisation issues, and access to repository content. Draft papers on these issues are under consideration now, and a final report on the preferred solutions to the various infrastructure issues will be released around the middle of 2006.[6]

A closely related work package has been investigating what is required for sustainable preservation actions. It is envisaged that the Sherpa-DP approach will be based on format migration strategies. While much of this work is only in preliminary stages, we have examined various preservation approaches and have been undertaking investigations of existing or required automated tools to perform preservation actions. The automated tools will need to address several crucial issues:

- File format conversion. Automated file format often produces variable results, according to the complexity of the file format. It is suitable for simple conversions, but may be problematic when data is deposited in unusual or diverse file formats, such as CorelDraw.
- Methods to monitor e-prints for obsolescence, and perform integrity checks. It is envisaged that this process will be performed at the preservation service, rather than by remotely monitoring individual resources. Obsolescence checks may be performed by the preservation service, or through a third-party, such as the UK National Archive's PRONOM tool.
- Processes required to enable changes and updates to e-print content that ensures their long-term integrity and preservation.

An extra component of these investigations has been to examine existing metadata held in the e-print repositories to establish if additional metadata needs to be collected or added. Work on developing preservation metadata for maintenance and accessibility to e-prints over time is in its final stages and a minimum metadata set based on the PREMIS data dictionary will soon be released to partner repositories for their comments.

Preliminary work on a workflow model for the provision of a disaggregated preservation service has also been undertaken. This work uses our understanding of the technical environment and metadata needs to construct a high level statement of the desirable workflow process. This is only a draft model and may not represent the final set of processes implemented by the Sherpa-DP implementation solution to be tested in late 2006-2007. At a high-level, the workflow will consist of six broad steps (see Figure 3.):

- The depositor (producer in OAIS terminology) submits a submission information package (SIP), consisting of an e-print and associated metadata, to the institutional repository.
- Institutional repository staff refine the descriptive metadata that accompanies the e-print, as defined by their internal requirements.
- On a pre-determined schedule, the updated SIP is transmitted to the preservation service, which generates an archival information package (AIP) intended for preservation.
- The AIP is stored within the archival store at the AHDS and an appropriate backup strategy is implemented. A copy of the AIP is returned to the institutional repository.
- The institutional repository generates a dissemination version intended for use by their user community (designated community) and makes it available via their search catalogue.
- A user is able to request and download a copy of the dissemination information package from the institutional repository.

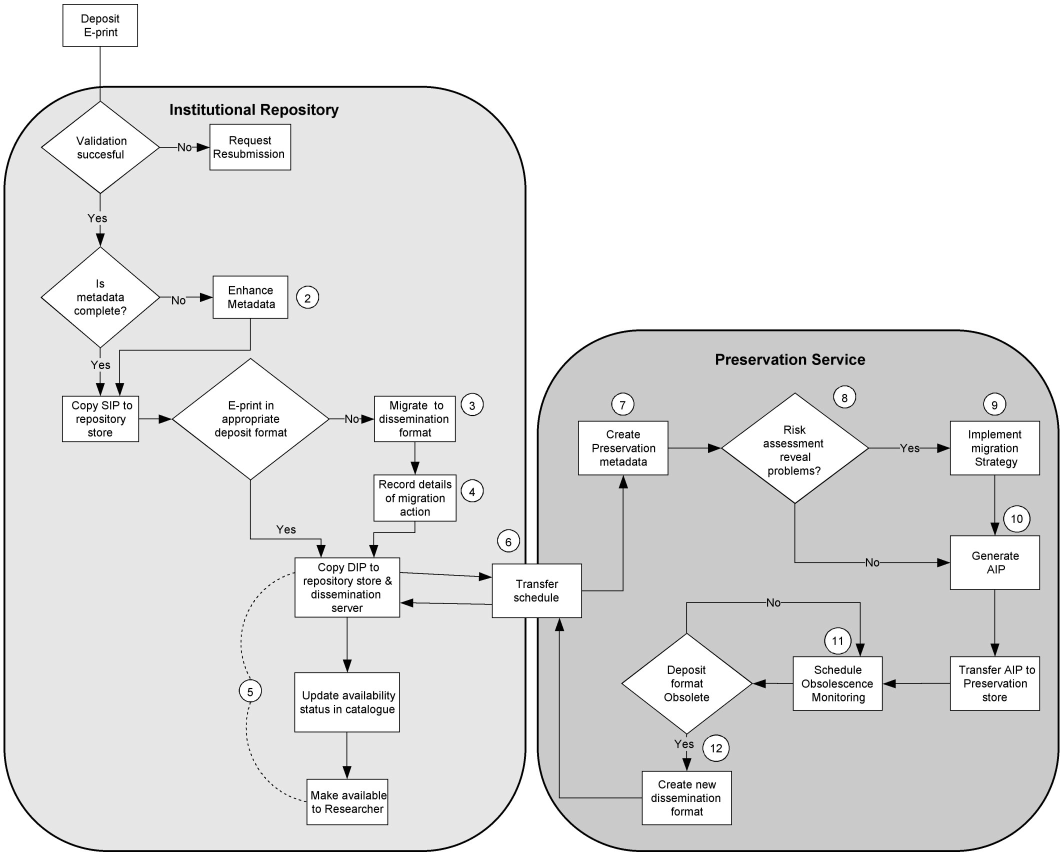
Deposit E-print
Institutional Repository
Validation succesful
No
Request Resubmission
Yes
Is metadata complete?
No
Enhance Metadata
2
Yes
Copy SIP to repository store
E-print in appropriate deposit format
No
Migrate to dissemination format
3
Record details of migration action
4
Yes
6
Copy DIP to repository store & dissemination server
Transfer schedule
5
Update availability status in catalogue
Make available to Researcher
Preservation Service
7
Create Preservation metadata
8
Risk assessment reveal problems?
Yes
9
Implement migration Strategy
10
No
Generate AIP
No
11
Deposit format Obsolete
Schedule Obsolescence Monitoring
Transfer AIP to Preservation store
Yes
12
Create new dissemination format

Figure 3 *Sherpa-DP preliminary workflow model*

Conclusion

In summary, the project work to date has concentrated on investigating and understanding the technical environment within which e-prints are collected and stored, and developing a practical implementation model that takes into account the technical environment and conforms to the OAIS framework.

Future work to be undertaken by the project will focus on providing a generic model that may be applied to other preservation services; finalising the workflow and procedures to suit the needs of institutional repositories and the preservation service; providing guidance on the ingest process, to encourage the deposit of formats that will minimise long-term operational costs; developing/refining software tools to enable communication between the institutional repository and the preservation service. The culmination of the project work will be the development and publication of a User Guide that recommends standards, best practice, protocols and processes for use in the management and preservation of, and access to, e-print repositories.

References

[1] All page references to OAIS are to the 'Blue Book' published in January 2002, and available at: http://public.ccsds.org/publications/archive/650x0b1.pdf [last checked 13 February 2006].

[2] See, for example, R. Crow, *The Case for Institutional Repositories: A SPARC Position Paper*, SPARC 2002; Beebe & Meyers, "The Unsettled State of Archiving.", *The Journal of Electronic Publishing*, Vol. 4, No. 4. 1999; C. Lynch, "Institutional Repositories: Essential Infrastructure for Scholarship Age." *ARL*, 226, February 2003, 1-7.

[3] I am indebted to my predecessor Hamish James, and to Sheila Anderson, AHDS Director, for much of the information in this and the following paragraphs.

[4] This is the basis of the JISC 4/04 program; N. Beagrie, *JISC Continuing Access and Digital Preservation Strategy 2002-5*, JISC 2002, p. A13.

[5] Hamish James, Raivo Ruusalepp, Sheila Anderson and Stephen Pinfield, *Feasibility and requirements study on preservation of e-prints*, version 1.0, 9 May 2003: http://www.jisc.ac.uk/uploaded_documents/e-prints_report_1-0.pdf.

[6] Thanks are due to my colleague Gareth Knight for the summary on which the following paragraphs are based.

Biography

Andrew Wilson is Preservation Services and Projects Manager at the Arts and Humanities Data Service, based at King's College London. Prior to joining AHDS in 2005, he worked at the National Archives of Australia on metadata and digital preservation projects. He is a member of the Dublin Core Usage Board and co-chairs the Dublin Core Agents Working Group. He holds a post-gradate qualification in Archives Administration, and a Masters Degree and PhD in Ancient History.