

DSpace Project Status Report

Date: June 2004

Prepared by: Michele Huston

This document describes the work done between Oct 2003 and June 2004 towards establishing an institutional repository at ANU based on the software platform, DSpace.

Project Background:

There is a growing institutional repository movement that is interested in both the technical and cultural aspects of coordinating the management and distribution of an institutions valued research and teaching materials. The success of any repository is dependent upon adopting appropriate strategies to ensure that it is populated and developing a repository infrastructure that will be flexible enough to support new functionality as awareness grows and new uses are found. Developing within a standards-based environment that facilitates interoperability is an important part of this strategy.

It is early days for institutional repositories and there is much yet to be done in establishing technical standards and a policy framework that encourages population of a repository. Maintaining a watching brief on international developments in this area is important to this project.

The Australian National University to establish a broad-spectrum digital repository that will allow ANU Faculties, Schools and Centres to manage their valued digital collections in the context of a professionally managed archive.

The repository will be based on the DSpace platform, an open source software development, championed by the HP-MIT alliance. The ANU is collaborating in the DSpace open source community as the best way to achieve a standards-based environment that will allow the ANU to participate in federated information services and to grow in functionality as new requirements for the repository are identified.

About DSpace

DSpace is a broad-spectrum digital repository system that captures content directly from the creators and systematically distributes this content within a rights management framework to a world-wide audience. Important to the DSpace philosophy is compliance to international standards and best practice in archival practices to ensure interoperability with other scholarly information services.

The DSpace Federation

The DSpace software development is coordinated by the DSpace Federation. The DSpace Federation aims to oversee governance of the code through an initial group of system architects and developers that will be expanded over time. The Federation will also examine copyright issues as well as licensing of the code & the DSpace trademark. It will foster new and innovative uses of the platform, provide a forum for institutional repository issues such as preservation and access.

Reporting against milestones

Area of work: DSpace System Evaluation

Objective: To assess the strengths and weaknesses of the DSpace 1.x architecture based on an ANU needs assessment.

Oct 2003 – Mar 2004: DSpace system testing and evaluation. The DSpace system was installed in Oct 2003 and a number of targeted demonstrators were designed to assess the strengths and limitations of the DSpace system. Three representative collections were identified that characterised aspects of a typical ANU collection: a collection of images, a PhD thesis and a music CD.

Demonstrator 1 - Image collection: To examine issues related to the use of DSpace to manage a collection of images.

A collection of 326 digitised objects from the Noel Butlin and University Archives were identified. The images and metadata were suitable for ingestion into a digital archive. The metadata was both substantial and consistently structured. The collection size was small, but large enough to be a useful test of the DSpace batch upload tools and the DSpace information model. All items were cleared of any copyright restrictions. Additional functionality was added to DSpace to support thumbnails and a web-view image. An automated system was created to generate these derivatives from the archival tiff format.

Demonstrator 2 - PhD Thesis: To examine issues related to the ingestion of a PhD thesis. It was already known that the DSpace system was able to accept and display text documents in pdf format. An alternate and arguably preferable preservation format for text documents is XML. This demonstrator looks at the conversion from MS Word to XML in the DSpace environment.

A PhD thesis from a humanities area was identified. The thesis was provided in MS Word format with styles applied consistently throughout the document. The document was composed of images and text structured into a number of chapters. The DSpace system was already proven as a repository of pdf documents. This demonstrator aimed to assess integration of the repository with an XML publishing environment. In addition to DSpace, a range of XML publishing tools were examined. The flexibility of the XML for display in multiple formats was demonstrated using the Cocoon framework. Automation of conversion of the documents from MS Word format to DocBook XML was also explored.

Demonstrator 3 - Music CD: To examine issues related to managing a complex object in DSpace. This demonstrator examines issues related to the ingestion of a collection of audio files, text and image files which are related to each other through their association on a CD.

The ANU School of Music publishes an anthology of Australian music annually in CD format. The CD has a number of compositions and each composition is made up of a number of movements. The individual tracks are simple objects in the context of the repository whereas the CD is a complex object made up of audio, text and image files. Wav was identified as the preservation format and Real Audio for the access format for the audio files. Issues related to the integration of a streaming server and DSpace were examined. The DSpace interface was modified to provide seamless access to the Real Media files on the streaming server via the DSpace interface. An rdf file was created to contain metadata describing the complex relationship between the files and their aggregation on the CD.

Area of Work: Repurposing of archived content

Objective: To examine issues related to the repurposing of content held in DSpace.

October 2003 – March 2004 - Items held in DSpace in XML format are ideally suited for repurposing. The content held in XML format can be manipulated using XSLT for multi-format display, e.g. as pdf, html, print, etc. It was considered that clients may want to place the digital objects in DSpace to take advantage of preservation treatments but that they may also want to republish this information within other contexts such as their own websites. A demonstrator was prepared to display text files in XML format combined with image files held in DSpace with a non-DSpace look and feel for publication on a Faculty website.

Area of Work: DSpace Production Environment

Goal: The goal of the DSpace production environment is to further an understanding of the cultural issues involved in establishing an institutional repository at ANU; such as shared responsibility for curation of collections and appropriate policy development that will encourage population of the repository.

March 2004 – Production environment launched - A production environment was established and initial communities were introduced in March 2004. A standard installation of DSpace was used with a

small number of minor modifications to the interface to support the image collections. A DSpace development/early adopters environment was installed. An ANU DSpace production environment was installed and commissioned.

May 2004 – ANU Look & Feel: The DSpace interface was redeveloped to apply an ANU look and feel. The new look & feel has been applied to dspace-dev for user testing.

Area of work: Collection Development

Goal: To encouraging population of DSpace through promotional activities and policy development.

March 2004 – June 2004 - The ANU has the largest DSpace implementation. The following collections were ingested into DSpace:

- 9 image collections – 40,000 images used for art history teaching.
- 2,000 pre-prints documents were uploaded into 200+ subject based collections
- 9 collections of images from the Noel Butlin Archives
- 1 collection of images from the University Archives

DSpace-dev was used to introduce new communities. Areas currently exploring the use of DSpace include; ePress, Department of Mathematics, CPAS, CNMA, Faculty of Asian Studies, RSPAC – photo collection, Centre for Atom-photonics, RSISE, NITA Library.

April 2004 – June 2004 - Materials Access Program - ANU Quality Review. A new instance of DSpace was installed to provide restricted access to 5,678 documents for use by 480 - 500 registered clients. Documents were distributed among 110 communities. Scripts were developed to create community and collections with appropriate access permissions from metadata provided by the client. Scripts were developed to ingest documents and associated metadata. Scripts were developed to register users based on metadata provided by the client. Regular usage statistics were provided to the Quality Review Task Force.

Area of Work: DSpace Federation

Goal: To establish the ANU as an active member of the DSpace Federation.

May 2004 – XML Interface - Building on earlier experiences with integrating the Cocoon framework, the DSpace interface was redeveloped using XML. There are significant advantages in using XML technologies in terms of separation of content, look & feel and scripting. This project was undertaken as a proof of concept for work being done by the DSpace Federation on the DSpace V2.0 architecture.

October 2003 – June 2004 –DSpace Discussion Lists – DRS staff are active participates in discussions with the DSpace community via the DSpace discussion lists.

Area of Work: DSpace Developers Group

Goal: To establish the ANU as an active member of the DSpace developers group.

A DSpace developers group was established to examine DSpace architecture issues and for governance of the code base. Planning is underway for V2.0 architecture that aims to strengthen DSpace commitment to preservation issues, support the open source development with a more modular structure and look at scalability.

May 2004 – Contributions to DSpace V2.0 discussion. A select group of six organisations has been identified as core DSpace developers including the ANU's DRS group. The ANU is an active participant in high-level architecture discussions and undertaking new developments. The ANU has contributed to the V2.0 discussion through a proof of concept project looking at applying XML-based technologies to the DSpace user interface.

Area of Work: Federated Services

Goal: To examine issues related to federation of ANU DSpace with other quality information providers.

April 2004 - PictureAustralia – The Noel Butlin and University archives collections were suitable for inclusion in PictureAustralia. The strength of the DSpace system is in its support for international standards particularly the OAI metadata harvesting protocol. Support for this protocol made it very easy to participate in this federation service.

March to June 2004 - Google Search Project – At the DSpace User Group meeting (10-11 Mar 2004) a pilot project to develop a search restricted to scholarly material was outlined. Google offered its technology and the DSpace community was approached for participants. During the testing and development phases of this project the harvesting would be restricted to DSpace repositories but it was agreed that once the concept was proven that all repositories of scholarly material would be included. OCLC as an independent group offered to set up a registry of scholarly repositories. Initial harvesting of content has been completed.