APSR Image Workflow Final Report

Submitted to:
Scott Yeadon
APSR Project Co-Ordinator
Australian National University

Prepared By:
Tristan Gutsche
Technical Lead
iSpheres Project
University of Sydney

December 14, 2007

Table of Contents

Acknowledgments

The project team would like to thank all those who supported the project and provided valuable contributions.

Rowan Brownlee , Nik , Gary Brown, Ross Coleman, Tony Green. Jacqui Spedding, Su Hanfling, Peter Raftos, Scott Yeadon, Chris Blackwell and the MAMS team at Macquarie University.

Project Overview

Project Summary – Goals and project background

The project was setup to develop a process for submission, sustainable storage, and reproduction, of image collections within the RIFF environment, which will provide the higher education community with greater resources and capabilities for the utilization of digital media.

The project will address the issues of, mapping image collection metadata to the standard NLA METS profile, as well as and fundamental presentation functionality for user interaction with image collections. Initially image collections where to be referred to as “Image Albums”, however this was latter dropped by the NLA committee.

To demonstrate the presentation of image collections, templates where developed in Manakin by ANU, and the iSpheres repository system developed a connector allowing access to Dspace objects and metadata via web services. By this means, objects from Dspace and other repositories, may be presented throguh custom web and desktop applications.

The project was scheduled to be complete in 2007 and will be demonstrated at the Clever Collections conference in November.

Stakeholders

Name	Project role	Phone	Email
Ross Coleman	Director, Sydney eScholarship	02 9351 3352	r.coleman@library.usyd.edu.au
Daniel Burn	ISphere Project Manager	02 9351 7775	daniel@med.usyd.edu.au
Tristan Gutsche	ISphere Developer	02 9351 6686	tristan@med.usyd.edu.au
Phillip Nguyen	ISphereCore Developer		pnguyen@med.usyd.edu.au
Scott Yeadon			
Peter Raftos			
Rowan Brownlee	Library Consultant		

Milestones & Deliverables

Milestone/Deliverable	Responsibility	Status
Project Description Document	USYD	Complete
Image Collections Workflow Requirements Document	USYD	Complete
Reference Group Meeting	USYD	Complete

Milestone/Deliverable	Responsibility	Status
Development of interface in liaison with METS and RIFF teams	USYD, ANU METS & RIFF teams	Complete
Implementation of metadata interchange protocols (including NLA METS, OAI-PMH, and MAMS authentication) that provide for appropriate metadata, including rights metadata, to be wrapped with the images extracted.	ANU METS & RIFF teams	Complete
Development of tools within the service that enable extraction, transformation and rendering within xml/xsl dissemination framework.	ANU METS & RIFF teams	Complete
Testing of image extraction and transformations across repositories.	USYD, NLA, ANU METS & RIFF teams	Complete
Write implementation document	USYD, ANU	Complete
Documentation and uploading of code to SourceForge	USYD, ANU	Complete
Presentation at APSR Interoperability Conference – Clever collections. The iSpheres project was demonstrated by Scott Yeadon as part of his overall presentation.	All partners	Complete

Requirements for Images and Image Collections

Image collections are built with many different intents and fit into a wide variety of roles in the education sector, not exclusivly to learning and research. Images are repurposed regularly for instance, educational material is used in promotion via websites and publications for institutions. Sustainable image collections utilising the CORDRA model, provide the higher education sector with a highly repurposeable resource giving the proper management and a means to utilise them.

Reusing and repurposing image collections are now the immediate goals of active institutional repositories. The wide scope of uses for images presented a challenge to define a base set of workflow requirements. We set out to dicern,

1. What types of images and information are being captured?
2. What purposes image collections serve?
3. How are images collections being manged?
4. How are images and collections being utilised?
5. Are images being repurposed, and through what methods?
6. Are images being shared with other institutions and what where the associated issues?
7. What means of discovery are in place?

In the initial round we investigated a series of repositories in different stages of development and use. The business cases gathered data on, their purpose, their intended audiences, aspects of collection management, their data sets,

cases of repurposing and examples of shared use. We studied image collections in the areas of medicine, botanical siences, the arts, archeaology, architecture, and marketing.

This was followed up with the formation of a reference group promrising of people working with large image repositories. The discussions focused on repository management needs, deslivery interfaces, workflows, security and legal constraints.

The findings from the reference group and busines cases was backed up with research into projects overseas in similar areas, discussion papers and input from the educational community.

This process provided us with valuable information. The summary of results are:

Many of the image collections had initially been setup as digitisations of existing slide collections. Some collections where setup as a “virtual” representations of physical subjects, such is the case for the medical specimen collections. Some collections where a compilation of commissioned works for unrelated projects and stored in a community image collection. Many of the collections did not have basic recommended archival metadata.

Workflows dealing with images that are archived, where considered distinctly separate to those with images in daily use. Images used in daily processes, tended to be compressed, in lossy formats, and compatible with a wider range of software. Below is a summary of issues mentioned. It is by no means a complete list of all issues associated with working/archival images.

Archive Image	Working Image
Very large files	Smaller size
Hi quality	Lesser quality
Act as a Master template	Multiple formats, and dimensions for different views based on the master image.
High cost	May never be archived.
	Low cost
	Image whos contract conditions required removal of images from all systems where not archived

There was a lot of consideration put into archiving, but in practic, it was more often left to the backup policies of the IT units managing the servers. Most collection manages recognised the need to use recommended archival formats as primary storage for their images.

In all cases collections had high levels of customised datasets, with minimal usage of standards such as Dublin Core or MODS.. There is a great variety in the taxonomies, and level of information associated with images. In some cases no metadata was stored on the images as their purtpose only became clear within the context in which they were presented users.

There are moderate levels of content sharing but under strict agreements. Manageing shared collections predeominantly involved copying entire collections to new locations then enabling the original collection owners to manage the content. Managing content sharing is considered to be part of the source collection managers responsibility.

Methods for shareing where not generally available in the software systems that where being used by many collections. In order to share collections within the RIFF environment an interface module would be required.

Repurposing was common place. The infrastructure for doing so was not in place, however it is the initiative of users, taking images from one collection and using them in different forms and mediums.

Out of scope of this project, but a driving future was the role of collaborative tools for sharing, such as comments or annotations and discussion

9 out of 10 collections were kept behind closed doors. Bound by various legal constraints, exposure was permissiable only to certain audiences. This is possibly the greaters limitation for sharing and repurposing image collections as in many cases the desire to share is there but the legal ability is not.

Delivery methods where not as widely varied as expected. There are only so many ways an image can be shown. The primary modes of presentation included, single image viewing, lists of images, slideshows of images, comparative presentations and embeding images in a variety of conext containers, such as google maps.

Images within a collection are commonly semantically grouped under a variety of subsets and multiple images will be found in multiple subsets. An example of this is the MESH mapping for medical images. MESH provides a descriptive medical vocabulary, which is being used in medical image collections to describe and categoriese images.

The collection as an object itself had minimal standard descriptors. For the profile we took the basics of who, what, and when to describe collections, their management, tracking information, contacts and legal information.

The on techincal concerns the reproduction of images in their correct colour space was raised as it was not uncommon to encounter artists stipulating in their agreements highly specific reproduction forms for images.

From this study we identified the key requirements for image collections within the RIFF enironment.

1. Identify a common metadata set for both images and the collections they are contined within.
2. Produce and interface for collections to be shared by transfering one site to another
3. Provide a means for interfaces to embed image sets by referencing directly from the source colleciton..
4. Define how METS encoding will need cover areas of :
    1. Copyright and authorisation management
    2. Mapping of custom data to prefered METS standards,
    3. Containment of custom dataset structures, with or without set

schemas.

4. Archival information
5. Technical information
6. Identification of physcal sources for images.

5. Develeop common methods for search, retrieval and display of images. These interfaces should be able to reflect the structure of the collections and objects therein.
6. Identify a process for manageing authentication and authorisation when access images
7. Identify an infrastructure to enable the reuse and repurposing of image collections.
8. Identify the different and cater for, the needs of images in archival lifecycles and in daily use lifecycles.

Metadata Mapping – Images, Collections and Standards

As stated eariler, most image collections being used in our samples used customised datasets particular to their purpose as a result most do not use a standard as these standards are may be too generic, to detailed in areas of no concern, or perceived as too bulky to impliment. The NLA working group recommended MODS, PREMIS and Dublin Core as standard for collections to be exchanged in. In order to map image metadata, we reviewed a number of alternative standards.

Of significant interest was Mpeg-7. MPEG-7 is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group). While the Mpeg-7 did cover images it was more directed towards the description of content then the workflow of the objects. It may be a useful standard to use as a means of storing descriptive metadata for AV collections.

For rights management we considered the Justice Sector Metadata Standard, however it significantly lacked componnets of workflow, such as “login required”, “permission required”.

Dublin Core though extensable was considered to be too generic to adequately express the complexities of many of the collections.

MODS, combined with PREMIS and MIX provided a better overall model with the same structural foundation.

Image repository metadata was broken down into 3 primary components. That of the repository that is responsible for storing the images, the collection to which they belong, and the objects themselves. Repository metadata was considered out of scope beyond information needed to connect, and access the image collections as the focus was on mapping collections of images whether they be from a single repository or sourced from multiple repositories.

Data was then broken down into the following categories:

- Technical and archive data
    - Processing information

    - Processing details
- Software details
- Authentication and authorisation
- Collection content description (or structure)
- Common descriptive metadata
    - Identifiers
    - Rights and ownership
    - Common descriptive metadata
- Record management data
- Related object mapping

These areas where then mapped to corresponding elements in MODS, PREMIS and MIX accordingly. Data that did not fit within these schemas were preserved in their custom form and referenced within the METS package.

Presentation of Images – Interface Design

We identified early on, that interface design is a very subjective area. There was a risk that when engaging people on their perception of image functionality that it could easily be misinterpreted as a discussion on interface preferences. To mittigate this problem, we focused our attention of image presentation and functionality to business cases and reviews, and kept discussion of interface functionality to very specific tasks.

The result was a good outline of the basic workflows associated with images and image collection management.

Search and retrieval – Almost all search and retrieval of images was based on metadata used to describe the content. There were no collections in our sample that used image recognition systems to identify shapes/objects/patterns in images. Images were in 100% of cases presented in small thumbnail images no smaller then 100x100 and no larger then 300x300. The presentation of the thumbnails may or may not be coupled with a small amount of desscriptive metadata, and would be presented as listsor tables. Slideshows of images at this sizeonly occurred as a navigational tool to view larger sized images, but were not common.

Image views – Images were presented with varying amount of contextual metadata. Some images where used jointly for comparative views, and some where presented singularly, without metadata but with the option for metadata display. Most commonly images were presented with metadata and a series of functions that users may enact upon the image.

Image functions: Resize was the primary function that all repositories had in common. Users are presented with a preview of the full size image. The preview was generally greater then 500x500 but no larger then 800x800. Users have the option to view images of greater and lesser sizes, and view the original image.

Collections of images where generally presented in the same form as search

results. Search results technically being a collection of images that are matched under a set of terms.

Presentation of object collection structure is reflected in interfaces predominatnyl through the navigation of the set. Where an object in a repository for instance, had 4 images associated with it, the image relationships determined the navigation methods and roles the navigation had. Purpose built interfaces naturally benefit from an end to end design maximising the contextual representation of the images in the repository. Generic interfaces suffer without mechanisms to identify the relationships between the images in a meaningful way.

A dynamic interface was considered out of scope for manakin within the project, however was developed for iSpheres. The initial version of the interface still lacked adequate representation of relationships between images. This is something that is peged for future development.

ISphere interoperability with the RIFF environment

ISpheres is a digital repository system that is designed to stand on its own or act as a front-end to existing databases. It aims to provide a standard interface to repositories of images, sound, video, and other types of data.

iSpheres uses the iSpheresCore server application to handle standard repository tasks such as storage, searching, and transforming the objects. The iSpheresCore acts as an abstraction layer over the underlying databases, and provides a web services API for interacting with the data.

The iSpheres web services API allows front-end applications to be built that concentrate on being a first-rate user interface, without needing to re-implement all the services provided by the core. This allows us to develop a range of front-ends that fit the needs of the user communities, rather than shoehorning users into a one-size-fits-all system.

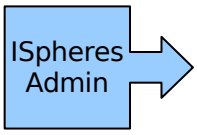
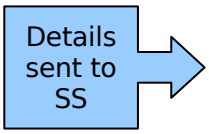
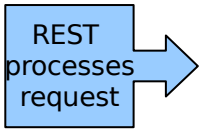
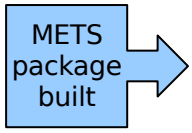
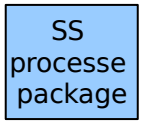
In order to connect the iSpheres repository system to the RIFF environment, the project team needed to overcome, mapping multiple data sets, from different collections to the METS package, providing a workflow for collection managers to submit their collections to other repository systems such as Dspace.

Working with Scott Yeadon we were able to quickly devise a means to transfer collections from an iSpheres to the submission service. Initial prototyping showed that using non-standard packages where causing rebuilding and new configurations of the submission service. This wasn't a feasible solution so we looked for a standard means. We decided to sync work on image mets package with the image collection transfers. This solution provided then submission service a single image collection handler for any collection from any source, not just iSpheres

The submission service does not use the SOAP service for web transactions. Instead it uses methods based on the REST architecture which was designed to take advantage of the HTTP inbuilt processes for data transfer. The solution was to build a REST interface which could pass through large packages via

HTTP attachments. Image objects where passed on demand rather then bundled with the METS package to minimise traffic, improve processing times and provide destination repositories with a full metadata set which may be presented to a user and the images sourced as needed.

The process is as follows:

1 	2 	3 	4 
From the iSpheresimage Administration interface, a user selects the image or images, the destination repository and collection, and a login where required.	ISpheres generates a REST call to be passed to the submission service to source the images.	The details are submitted to the service, if all the options are correct	The service makes the call to the rest service.
5 	6 	7 	
The REST service processes the call and requests the data from the iSphereCore server.	The result from the server is then converted into a METS package and returned the the submission service.	The submission service then processes the package, submitting calls to the REST service as the images are required.	

In order to map different datasets from multiple collections, iSpheres developed a metadata transformation engine which allows collection owners to register transformations for their collections. This allows for the flexibility of mapping to as many different datasets as needed.

The iSpheres team also investigated various means of handling the authentication and authorisation issues. The main issues were, how to enforce restrictions, and how to track adherence to these restrictions. We looked to the fields of federated identify management for potential solutions. There are a number of projects dealing with this problem. The one that aligned most closely with what the iSpheres group was trying to achieve was the Meta Access Management System (MAMS). We engaged MAMS with a number of discussions about potential options and decided that at this time it was out of our scope to implement a MAMS solution, however we believe that this is the way to proceed. Their solution provides collection manages to enforce what

can be done and by whom.

Outcomes & Recomendations

The project completed all the milestones and delivereables it set out to achive. Along the way we managed to raise awareness for repository managers of other collections and the issues that they face. We also manage to raise awareness of the issues and potential solutions to archiving and the copyright problems faced by all image repositories.

The requirements for the image profile, identified a common set of management, technical and structural data that can be associated with images. It is recommended that image collections upgrade their basic metadata sets to include these basic requirements. Customisation is not discouraged, but embraced as an essential for effective image collections.

The iSpheres software, with the development of the interoperability with the RIFF environment is able to utilise Dspace as an archival repository for institutional collections. ISpheresImage is now bundled with the and interface for the RIFF Submission Service, enabling collection manages to easily share their collections through RIFFS. The iSpheres transformation engine and REST services allows iSpheres to offer harvesters a means based on the OAI model to gather data in any available format. ISpheresImage is sighted for further development of context representation for multi sourced image collections.

The CORDA project states that reuse and re-purposing of objects are issues for the future concerns for repository development. The results of the APSR projects are a demonstration that reuse and re-purposing is already well underway, as an immediate concern. At the time of writing this report, iSpheres is currently implementing processes for embedding images into MediaWiki and enabling users of the FSKEditor embedded web html editor to use iSpheres as the content management system. The aim is to provide tools for users to reuse resources in different contexts and purposes.

The project has highlighted the need for better management of copyright and contract restrictions on images and their use with the goal for greater sharing. The recommendations in the OAK-Law report provide a strong step in this direction.

There is also a strong drive for more collaborative environments, which provide an added layer of context and sharing from the users perspective rather then the data store replication. This influence is primarily from the huge number of collaborative sites online just as Flickr, Facebook and mySpace. There are many projects already investigating this area for the educational sector and it would be wise to monitor and adopt where it reflects the target audiences needs.

References

CORDA (http://cordra.net/)

EDNA Groups http://www.edna.edu.au/edna/go

FSKEditor http://wiki.fckeditor.net/

Justice Sector Metadata Standard.
http://info.lawaccess.nsw.gov.au/lawaccess/lawaccess.nsf/pages/jsms

MAMS http://www.melcoe.mq.edu.au/projects/MAMS/

MediaWiki http://www.mediawiki.org/wiki/MediaWiki

METS http://www.loc.gov/standards/mets/

MIX http://www.loc.gov/standards/mix/

MODS http://www.loc.gov/standards/mods/

MPEG-7 Moving Picture Experts Group
http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm#E12E8

Open Access to Knowledge http://www.oaklaw.qut.edu.au/

PREMIS http://www.loc.gov/standards/premis/