Software Architecture Meeting, Wednesday, October 31st , 2007

(Ananthan, Beard, Ellis, Geng, Jantz, Liew, Mills, Triggs , Yu; Guests: Gardner R., Hartman K.)

Agenda

1) Investigation of different ways of exposing elements of the repository to Google, Triggs. Members of Faculty Services will attend for this topic.
2) JP2000 R4.5 spec (http://rucore.libraries.rutgers.edu/collab/ref/spc_sawg_r4_5_jpeg2000.pdf)
3) Investigation of Fedora's rel/ext and diringest, Geng
4) Digital Docs/Administrative Space R5.1 Specification Review
5) Handle server, single handle pointing to multiple objects

General Announcements

Kalaivani and Chad identified three duplicate Collection ID/Name entries in the dlr.collections table on MSS3; 'RUL', 'RULGriffis' and 'NPLRNG'. It is thought that originally the collection objects were purged from dlr/EDIT and when this purge occurs the dlrcollection table does not cleanup purged entries. The collection object was then re-ingested from WMS and a duplicate entry then was created. As of R4.5 a collection object ingest will check for existing Collection ID/Names in the dlr.collections table and not allow the duplicate entry to occur. So would have had to of occurred before the R4.5 release. In the future release cleanup of the dlr.collections table will occur when a collection is purged using dlr/EDIT. The duplicate entries will need to be removed from the dlr.collections table as well and the process for that will be explored by Kalaivani and Chad.

First Fedora install for NJVID is schedule for first quarter, 2008.

Agenda Items

1) Investigation of different ways of exposing elements of the repository to Google, Triggs. Members of Faculty Services will attend for this topic.

Karen Hartman and Rebecca Gardner from Faculty Services sat in for the discussion of the write up Jeffery sent out regarding Google Indexing. Using Google Sitemaps capability will be a better solution for exposing objects to be indexed in the repository by Google and other search engines than using OAI harvesting. Google Sitemaps can be constructed to limit indexing to certain objects whereas OAI harvesting provides an all or nothing solution with no control.

Three implementation scenario's were identified:

1) Enable showfed to FOLLOW only PDF links in objects. Object metadata and PDF's will be indexed then.
2) Enable showfed to NOFOLLOW any links and expose the full metadata of the record to Google to index. This will direct users to a full metadata record view and a link back to the brief record view with link to the objects would need to be created.
3) Enable showfed with Descriptiove Data in the Meta tag in the header of the page and FOLLOW only PDF link in the objects. Object metadata and PDF's will be indexed then.

Scenraio 3 proved to be the most likely to implement out of the three. It exposes more metadata for indexing than scenario 1 while still providing the ability to have PDF links FOLLOW'ed and indexed. Scenario 2 would also require the user to know to click on the link to get back to the main showfed page which might not be apparent and desirable.

Lastly, by exposing the PDF for indexing it might jump a searcher into a PDF without any context of the rest of the record. It will be explored if search results in Google might be nested when displayed, maybe through Sitemap nesting. Triggs will investigate further with a Google Sitemap on some lefty objects. Google Sitemap is also compatible with other major search engines such as Yahoo and MSN.

2) JP2000 R4.5 spec (http://rucore.libraries.rutgers.edu/collab/ref/spc_sawg_r4_5_jpeg2000.pdf)

Isaiah presented the JP2000(JP2K) conversion specification. Some issues are if a conversion of TIFF to JP2K were to happen should the TIFF be deleted after conversion? What would happen if the JP2K was not created properly and the TIFF had been deleted? Isaiah mentioned there is no commercial OCR engine available that could read in JP2K files so TIFF's would still need to be delivered for OCR. Most likely the implementation of JP2K as a archive would need to occur in many steps over many years.

Phase One: Ingest JP2K into an object that was created outside of the pipeline.
Phase Two: Create and ingest a JP2K file from a TIFF in the pipeline.
Phase Three: Convert old objects with TIFF to JP2K. Implement JP2K use in the search display.

Restructuring of ARCH1 would need to happen as well since the JP2K file would be used for presentation and archiving.

3) Investigation of Fedora's rel/ext and diringest, Geng

Jie presented documents relating to her Directory Ingest and Relationship services investigation work. For Directory ingest all metadata describing the files in the ZIP is in a single MET.xml document in the root of the ZIP file. For REL:EXT, contextual information about the parent and children nodes can be expressed in the RELS:EXT section, not just object ID's. This would provide a better display of child node information to the end user when looking at the parent node than just object ID's. Using RELS:EXT without Directory Ingest would be troublesome. Locking of object ID creation is not possible right now and when ingesting a child or parent object related objects ID's need to be known before ingesting the object. Confidence in the being able to determine those object ID before they have been create is low. A child object would need to be ingested first and the object ID would be obtained. Then the parent object would be ingested and the child object ID would be inserted in it. The child object would then need to be edited and updated with the newly ingested parent object ID. An identical process would need to happen when siblings exist.

In general using RELS:EXT would be troublesome from a purging of object perspective. Ron mentioned Fedora does not have any native API purge REL:EXT object arguments so a purge process would need to be created by us.

An alternative was suggested. If for ETD's supplementary files were just included as separate datastreams in one object with separate metadata datastreams sections for each supplementary file. See diagram.

Mods.Main
Mods.Suppl1
Mods.Suppl2
Main
Suppl1
Suppl2

Search interface enhancements would need to be made to direct a searcher to the appropriate datastream for a search hit, this would not be trivial. Also the question was raised if XACML allows for restricting permission at the datastream level? In a multi-object scenario XACML would provide for restriction of the whole supplementary file object if it were needed, but can this restriction still occur if everything were one large object?

4) Digital Docs/Administrative Space R5.1 Specification Review

Next meeting.

5) Handle server, single handle pointing to multiple objects

Next Meeting.

Next Meeting

The next Software Architecture Meeting is scheduled for Thursday, November 15th at 9:30 A.M. at the SCC. An initial, proposed agenda is below:

1) Video Streaming investigation paper, Beard & Hoover.
2) XACML – Can permission be defined at the datastream level?
3) Handle server, single handle pointing to multiple objects
4) Digital Docs/Administrative Space R5.1 Specification Review