Software Architecture Working Group **July 21, 2009**

Minutes of July 9, 2009 Meeting

Agenda

1. Quick updates and Announcements
2. Migration and testing status
3. PDF text specification
4. License document requirements
5. Review of NJEDL status

Announcements and Quick Updates

We will begin the NJVid training process with a WMS and metadata session on Wednesday, July 15. Grace will provide the WMS training and Isaiah will set up a video that can be ingested on lefty64.

R5.0 Migration and Testing Status

Dave reported that all support code has been compiled and he will proceed with installing R5.0 software on mss4. Isaiah and Jeffery will do final testing on mss2 to verify the recent bug fixes related to WMS edit. In addition, we will do one more final video test. Assuming mss2 tests are OK, we will proceed with final sanity testing on mss4 beginning Friday and proceeding into early next week. Kalaivani will also put the controlled vocabularies in place. Jie will do sanity testing on the faculty deposit software. While Kalaivani is on vacation next week, Isaiah will be the contact for final testing. Assuming all is well, our target for public release is Wednesday, July 15.

PDF Text Specification

We discussed the issues related to generating the proper xml1 datastream with "div" markers. This datastream is required in order to do full text searching across multiple text documents. The only scenario that works at this point is one in which the pipeline does ocr on multiple tiff files. Note that there are two issues related to this functionality which we will not deal with at present: a) the page limit for ocr is now at 75 – the limit is imposed because of the time delay incurred when doing ingest, and b) we may want to standardize the approach with the other scenarios – see below.

We concluded that more testing is required to make sure we understand the problem. The problems and scenarios to be tested are outlined below:

Faculty Submission. The scenarios to be tested include: 1) A single Word file is submitted. Check to see if a) an xml1 datastream is created and b) if the datastream has div markers, 2) Two Word documents are submitted. Test to see if WMS can be configured to create a single PDF or it can be configured to create a PDF from each Word file. In both cases, tests a) and b) as above should be run and 3) Tests a) and b) should be run when a single PDF is submitted without a text layer.

ETDs. For ETDs, the following tests should be run: 1) For an ETD that originates as a Word document, test to see if div markers are included in the resulting xml1 file and 2) do the same test when an ETD is submitted as a PDF without text (is this possible?).

Regarding implementation, the pdf2xml CGI script can be used to insert the div markers and produce the proper xml file, including properly encoded special characters (script developed by Jeffery). It appears that all changes resulting from the above tests would be made in WMS. As mentioned earlier, we may want to consider standardizing the process for tiff-based documents as well.

Handling of License Documents

Ron reviewed a preliminary requirements specification for license documents that covered two basic scenarios: 1) each resource has a unique license document (LD) and 2) a license document is used for multiple resources. Access, search, and presentation scenarios were discussed briefly. Although there is more discussion required, we concluded the following regarding implementation: a) the LD will be a separate object, b) two-way pointers are required between the resource and the LD, c) pointers will be implemented using rels-ext and d) related services for creating ontologies, and creating/deleting relationships will need to be developed. This capability will be targeted for R5.2, although we may need to handle LDs in an ad hoc fashion in R5.1.

NJEDL Status

Vince has been working with Jeffery and Kalaivani to test WMS batch ingest with NJEDL objects. He is encountering some strange problems related to doing ocr on tiff images. There is some speculation that this could be related to the ocr server. Vince will continue to try to isolate the source of this problem. However, we did conclude that we can safely reuse all of the ocr-ed text from the 2700 objects that were ingested originally into Fedora. Vince will investigate to make sure this is possible and also examine the remaining objects in NJEDL to determine how many require ocr. Given the ocr page limit, we should probably abandon the approach of using the pipeline for ocr ingesting NJEDL objects.

Pending for Next Meeting

- Results of PDF text tests
- Requirements for license documents
- Launch studies of “Perl to PHP”, java bridge and mss3 to rucore transition
- Determination of date for celebration of release R5.0
- NJDEDL status

rcj - -7/021/2009