Software Architecture Meeting, Thursday June 12th, 2008

(Ananthan, Beard, Colonna, Daniel, Ellis, Geng, Liew, Marker, Mills, Nakagama, Triggs, Yu)

Agenda

1) Shibboleth/Apache attribute passing
2) Finalizing R5.0 Migration and XACML specifications, on RUcore site currently
3) R5.0 Content model specification draft, distributed by Jeffery
4) R5.1 Statistics Specification draft, distributed by Jie
5) Faculty Deposits that need text layers and OCR
6) NFS directories/configuration, ongoing discussion

Postponed until next meeting
- NFS directories/configuration, ongoing discussion

General Items

- Next meeting is Thursday June 26th.
- Sujay Daniel of NJEDge.net was welcomed. Introductions were made along with brief explanations of everyone's role in the software development and deployment of RUcore.
- It was determined that in the near future part of a Software Architecture meeting should be dedicated to giving a more in-depth explanation of the applications we have developed, how they interact with Fedora and each other.

1. Shibboleth/Apache attribute passing.

- No schema has been defined but will be needed in general for Shibboleth implementation and attribute passing.
- There is a Shibboleth Working Group and it will need to be determined how to communicate the needs for specifications from that group to this one.
- The group finds the need for an initial draft framework diagram outlining the interaction of authenticated and authorized users with applications that will need A/A.
- The current status of a test Shibboleth site was mentioned, however it was unknown at the time.

2. Finalizing R5.0 Migration and XACML specifications, on RUcore site currently

- No concerns were raised regarding finalizing the two R5.0 specifications. Finalized.

3. R5.0 Content model specification draft, distributed by Jeffery

- Proposed in R5.0 new and existing Content Models will be edited/created in dlr/EDIT.
- Proposed in R5.1 new and existing Content Models will be edited/created in Open Source WMS by possibly using the File Policy builder. It was agreed this would be a Rutgers specific tool as it is so policy driven and RUcore specific.
- Validation of Content Models is unknown at this time due to lack of understanding Fedora 3.0 features. When we get closer to have a final Fedora 3.0 distribution these features will be known.
- Guidelines for testing Content Models is unknown at this time and will be expressed in the specification document.

4. R5.1 Statistics Specification draft, distributed by Jie

- A line item will be added that mentions WMS involvement in the statistics package. This is intended for editing and purging of objects using the WMS Fedora edit.
- Marker will investigate the length of time the IP addresses need to be preserved as outlined by the Patriot Act.

- Marker mentioned that an agreement of responsibility regarding the use and access to user information. It will be brought up at a Cyber Infrastructure meeting.
- The overall collection hierarchy of the repository is currently fragmented. Both NJDH and DLR/RUcore starting collection nodes are not related to each other by a super/root collection. It was discussed and agreed that the need for a root collection node with NJDH and DLR/RUcore as its children should be added. This will create the ability to completely render the entire repository collection hierarchy. This root collection will be created and called "root" with the DLR and NJDH collections becoming children of it and having the corresponding mods:relatedItem fields in the collection objects updated.

5. Faculty Deposits that need text layers and OCR

- Currently in faculty deposit a few scenario's exist where OCR and text layer extraction needs to occur and is not.

1) Deposited PDF's need to be examined to determine if they have a text layer. If no text layer exists the PDF will need to be OCRed. A text layer will be added to the presentation PDF and that text layer will need to be extracted for full text searching.

2) Deposited PDF's with a text layer in them will need to have that text layer extracted to provide full text searching.

3) Deposited Office Documents(Word, Excel, Powerpoint) that have a PDF generated have a text layer in them and they will also need the text layer extracted for full text searching.

- Currently TIFF's that are uploaded are OCRed in the pipeline and the text layer is extracted.
- An inconsistency was found in ETD where some have full text datastreams and some do not. Ellis and Ananthan will investigate.
- Mills will investigate the possibility of using the PDF server to determine if a submitted PDF has a text layer in it and also the process of taking a PDF without a text layer, converting it to TIFF, OCRing it and generating a PDF with a text layer from it.

Topics for next meeting

1. NFS directories/configuration, ongoing discussion
2. Resource Index issue with migrating objects to Fedora 3.0, raised by Nakagama
3. Fedora/Application architecture briefing