Software Architecture Meeting Minutes - February 3, 2011

Attending: Kalaivani Ananthan, Isaiah Beard, Jie Geng, Dave Hoover, Linda Langschied (guest), Rhonda Marker, Chad Mills, Sho Nakagama, Jeffery Triggs (recorder), Yang Yu.

Agenda:
1) Updates
- Google Scholar (Triggs)
- Data Portal (Ananthan, Mills, Womack)
- Book disseminator/sym links patch (Triggs)
- R5.2 Data Portal Specification update, sent last week (Mills)
- R5.2 Development Status

2) Continued discussion of Object updates – XML-1 for ETDs and thumbnails for PDFs (All)
3) Planning for JPE migration (All)
4) Performance of handle server (All)
5) Handling modifcation requests (Ananthan, Mills)

1) Updates
— Google Scholar.
The situation is the same as at the time of the last meeting. There are 1,330 objects in Google Scholar and about 10,300 objects in the regular Google index. We are still awaiting the full refresh of the Google Scholar index itself that should happen in the next few weeks. In the meantime, Jeffery and Chad will work with Rhonda to find a place or places to link to the scholar browse page from the RUcore site. This will allow Google to discover the scholarly objects through its natural link-following process. Since the browse page is based around year dates, we may want to make mods:dateCreated a mandatory field in the faculty deposit interface at least.

— Data portal.
Since the last meeting, RUresearch went live on the production server, though at first it was still pointing at resources on the development server. These objects have been exported and are in the process of being migrated. We have had to patch a few files to handle the new kinds of datastreams used in the data objects. This data patch is now ready to go. For future releases we will look into generating the list of acceptable mime types dynamically. We need to put more test objects on the Solaris staging server so that new functions can be tested there before moving to the production server. For instance, we need to have an SPSSPOR-1 datastream on the Solaris staging server. We initially used the datastream ID SPSS-POR-1, but decided it would be better to drop the intermediate hyphen.

— Book disseminator patch.
The book table of contents disseminator patch is packaged, tested, and ready for delivery. Sho and Dave have started deleting synonyms from the TMP directory.

— R5.2 data portal specification update.
There have been minor updates to the document. The diagram on first page was edited so that it now shows the major “project collection” rather than the “view object”. Nothing

needs to be aggregated in the near term, so for now we will put in a "contact us for aggregate data" note, and consider how best to handle aggregation in R5.2.1. Currently descriptive events are not closely coupled with RELS statements. The current implementation specification for R5.2 is not longer apposite, since the prototype has been pushed out in advance of the next release. The current specification suggests using "rulib_data " as an authority attribute for data genres. There was some discussion about this. It is hard to control all the genre vocabularies, but we need a place to search against the terms. In the future, it may be better to use RELS statements in indexing, but for now genre with human intervention may be the best way to go. Dome people wondered if we should use "rulib_subtype" instead of rulib_data. We need to consider these questions, but come up with something specific for data. Chad will take the document to CISC next.

— R5.2 development update.
We are supposed to be done by the end of this month, though it begins to look as if we will need more time. If some developers are able to finish sooner, however, testing could begin first on those parts of the release. We estimate that we will need a three week extension, but will go ahead with testing as elements become available. We may be able to let lower priority elements slide to the next release.

2) Continued discussion of Object updates — XML-1 for ETDs and thumbnails for PDfs.
Linda and Jeffery have been looking at search functionality, and discovered that missing XML datastreams used for full text searching have probably caused many if not most of the seeming search anomalies. These need to be fixed before we can do further analysis. The problems with ETD objects are now fixed, requiring only the update of objects with legacy issues. The problems with Faculty Deposit objects are not necessarily fixed in the current implementation. It was noted that for the Faculty Deposit objects, the PDF server on the production system has not been OCR'ing PDFs, though it continues to work on the development PDF server. This problem may have started in December. We will investigate this further, as many of the Faculty Deposit PDFs were presumably created from word processed files, and thus would already have had text and should not have required OCR by the PDF server. One issue may be a handful of PDFs that have had their contents "locked", so that their text cannot be copied. We will proceed with a cleanup process, creating XML-1 files for the objects identified as having a PDF with text but not an XML-1 datastream. Jeffery will test the process first on the development system. We should be able to generate the new files for the production server within the next two weeks.

We discussed how we should create thumbnails for new objects as well as older objects that need them. We need file policies for how to generate thumbnails. Kalaivani will look at the current content models and enable thumbnails where appropriate. It was agreed that for objects with multiple TIFF files, we will use the first TIFF if possible. If an object has only a PDF, we will generate a thumbnail from the first page of the PDF using the hash number syntax in ImageMagick. We will stick with the first page of a given object for now. Audio objects will have to be treated like Video objects with manually created and uploaded thumbnails. Older objects may need to be updated. Jeffery will work on a report based on objects in the development server to determine which objects need thumbnail addition or repair.