Software Architecture Working Group
Meeting Minutes
March 17, 2011

Present: K. Ananthan, I.Beard (recorder), J. Geng, D. Hoover, R. Marker, C. Mills (co-chair), S. Nakagama, J. Triggs, Y. Yu

Agenda:

- 1) Updates
  - Google Scholar (Triggs)
  - R5.2 development status (All)
  - Shibboleth on production (Hoover)
  - CISC report (Mills)

- 2) Object Update - XML-1 datastreams and thumbnails (Triggs)
  - Continuation of previous discussion concentrating on XML-1 datastreams for objects that do not have one. Also focusing on adding text layers to PDF's & DjVU that do not have one that should.

- 3) Handle server specification (Nakagama)
  - Continuation of previous discussion. Expected outcome is determining solution paths and scheduling those solutions into next version of RUcore

- 4) Handling Modification Request - (Ananthan, Mills)

- 5) Large file support in Fedora and WMS - (Marker)

- **Updates**

  - **Google Scholar (Triggs)**
    - Jeffery reports that there are now 2,280 objects found in Google Scholar
    - Pazzani article is now associated with the RUcore/mss3 version
    - Total expectation has been met
    - An e-mail will be sent by the subgroup to CISC announcing this development

  - **R5.2 development status** • Code Complete Date scheduled for March 21
    - Jeffery reports that his code, and MARC export, will be ready. Some improvements made to showfed.
    - Jie is waiting on faculty submission portions, but other non-dependent pieces of code should be ready.
      - Stats testing: Dave will dump production tables to the staging server for a test run.

- Yang: reports about 90% complete with code revisions and new specs related to 5.2, but has not had time to do regression testing before releasing the code to the group for testing. Will be able to demo new workflow to MDWG on Monday.
    - Concerned about the changes to a collection-oriented approach and how it relates to access restrictions and authentication. Authentication is currently organization-based, and will require some extra time to re-align.
    - Addressing bugs began this week. Some bugs will be fixed as part of code changes, but others will take additional time.
    - Estimates a need of about 1-2 weeks additional time to complete code and conduct a full round of regression testing.
    - Rhonda suggest possibly have a few people assist with regression testing for Yang, particularly Kalaivani. Both agree this would be helpful.

- Chad also needs some additional time, to about April 6.

  A new code freeze for these components needing more time has been agreed upon for April 6, 2011, but developers should continue to work towards finishing their code and making components available for testing sooner than that. Most work, including Jie and Jeffery's work, will be completed by the March 21 date. Kalaivani has asked that developers give notification as to what is ready as soon as it's available, so that testing can commence ASAP.

- **CISC meeting update**
    - Three R5.2 specs were presented at the last CISC meeting (March 9):
        - The data portal specification was approved with no changes.
        - A new date specification was approved with no changes
        - A revised Portal specification was discussed. Some minor revisions are needed, mostly related to how we refer to items that currently fall under what we call "cultural and historical heritage" collections. CISC is currently voting on what to name this subsection of objects.
        - Future development will be needed for implementing Descriptive Events (in MODS) to describe reLS connections. Grace is drafting a spec, that will at some point in the future be presented to SW_ARCH for review. Grace is aware that this will be considered after current issues are resolved, at some later date further down the development pipeline.
    - **Other issues**
        - Item-level portal associations (Dynamic Collections)
            - Kalaivani raised concerns about the number of objects being ingested, and the need for notification/coordination when ingests are taking place.
            - DC-Identfier: Yang and Jeffery will need to work together on test objects in order to check dynamic portal functionality.

        - License document – Rights Event Linking Capability
            - This functionality would permit license documents, release forms, deposit agreements, and similar documents that pertain to rights and permissions for collections and objects to be stored in RUcore and associated with pertinent object/collections. Objects will need to be embargoed as they will not be available to public users, but can be examined by appropriate admin users. One possibility would be access through DLR/edit or other similar authentication-controlled

interface.
This functionality is slated on development roadmaps for Release 5.2.

- Apparently, it was suggested that no WMS changes would be necessary? This turns out not to be true.
- More details and information are needed to proceed. Chad will look up Ron's write-up on this functionality to get more information. In the meantime, some manual implementation may be required while a more permanent functionality is developed.

- **Shibboleth (Dave Hoover)**
  - Shibboleth is still not working on the production server. Initially this was due to a key opening error, which has since been resolved through collaboration between Dave and Chuck Hedrick. Now there is a new error: "signature not verified." Dave will continue to work with Chuck to find a solution for this problem.
  - Shibboleth *does* work properly on the staging server. A last-resort scenario could be to temporarily host objects requiring Shibb (such as Journey to Planet Earth videos) on the staging server where authentication is known to be working correctly.
  - Dave will continue to keep us apprised as progress is made.

- **FFMPEG**
  - FFMPEG, the library currently being explored as a solution for frame-grabbing of videos in the annotation tool, is also not working; it will not compile under Solaris. Attempts to compile using a minimal install of libraries (eliminating unnecessary items like the media player and hardware driver support) did not have any effect.
  - The problem may resolve itself when a migration to new servers (under linux) occurs, but solutions will continue to be explored. So far, FFMPEG seems to be the library of choice for video manipulation and transcoding in a *NIX environment.

- **Journey to Planet Earth Migration Document (Kalaivani Ananthan / Jeffery Triggs)**
  - Document present to SW_ARCH; KA would prefer to hold off on migration until the aforementioned Shibboleth issues are resolved.

- XML-1 Datastream Update (Jeffery Triggs)
  - We identified many more than 45 objects that needed datastreams that could be generated (and in fact they have been generated and are waiting on lefty64 to be added). But we also identified a small set of objects (about 45 perhaps) that should have XML datastreams based on the text type, but where we could not generate the XML from either a PDF or a DjVu file. Spot checking some of these suggests that they may not really be files that should have XML texts - there are some postcards, for instance, with handwriting on the back given the mods:typeOfResource "text". Jeffery recommended we save these to look at later and perhaps create the XML by hand if possible and needed, and get on with adding the files for the majority of objects missing them and having creatable texts. ~~Jeffery has been analyzing the nature of objects lacking XML-1 datastreams. It appears that about 45 documents lack the datastream but do in fact have usable OCR text.~~ This

suggest that correcting the XML-1 datastream issue will have to occur on fewer files than previously anticipated~~a relatively small subset of files~~, and can be handled on a manual basis.
    - Some additional testing of the required scripts will be performed on the staging server before proceeding on the production system. Jeffery will touch base with Dave to get what he needs in order to make this happen.
- Another observation involves sets of documents that were scanned by a particular outside vendor. The vendor applied a dithering algorithm to each bitonal scan, causing text to be rendered in a very fine but fragmented dot pattern that confuses most OCR software and makes it nearly impossible to correctly recognize the text. The only recourse for these documents is to completely re-scan them.

- **Handle Server Issues (Sho Nakagama)**
    - In order for handle generation to be sustainable and robust for ongoing growth of RUcore collections, it is necessary to make a number of changes
        - A new handle server needs to be brought online at TAS for production, running a MySQL database and new security keys for which we have the passwords and can effectively administer
        - We are using an older handle prefix and format. We will need a new handle prefix and should begin generating new handles based on currently-established CNRI formats.

    - We can continue to support old handles by ultimately remapping them to newly generated handles for legacy objects, causing users of the old handles to redirect to new ones.
    - Intervention with staff at CNRI will be necessary to effect the needed changes, obtain a new prefix and to recognize the new server as authoritative. We anticipate there could be a downtime period of up to a week where at the very least, ingestion of objects will not be possible while the new handle server is established and brought online.
    - Sho will write up a procedure/specification document which will detail needed steps and actions, and what impacts may occur as a result.

Next meeting: Thursday, March 31, 2011 @ 10:00 am • Heyer Conference Room, SCC