

State of UFDC Programming

June 30th, 2008

Mark Sullivan

Recent Changes to UFDC

The UFDC web presence has undergone major changes over the last couple months. These have been a result of optimization efforts as well as a usability study conducted by Marilyn Ochoa and Patrick Reakes. These changes included a complete refactoring of the resultant html pages and ASP.net C# code for their creation. Basic design decisions were examined in light of the new amount of data and process flows, resulting in quicker execution. Support for Spanish and French was added to the application, as highlighted by the Digital Library of the Caribbean collection and interface. (<http://www.dloc.com>) In addition, the incoming URL is now parsed, allowing the dloc.com domain to point directly to the UFDC folder on the web server, removing any separation which existed before.

A more complete list of the changes appears on page 4 of this document.

File Consolidation and Verification

We are extracting all digital resource files from the Greenstone collection hierarchy structure. All the files will be consolidated into a single `image_files` folder on each server. This consolidation is nearly complete, with just one collection left to collapse. This should complete today, and then the process of verification of the copy can occur. Only when this verification is complete, will we delete the old files from each collection folder.

A new version of the UFDC Loader/PreLoader has been tested on the developer's computer which will correctly handle the new structure. However, no loading or preloading can occur until the data move has been verified.

Cross-Server Duplication Verification

Once the move has been confirmed and the pending items have been preloaded and loaded, another short pause will occur while it is verified that all the data on the production server and development server are identical. Any errant files will be evaluated and either duplicated or deleted.

UFDC Optimization Phase Nears Completion!

Optimization of UFDC was addressed and coding began a little more than two months ago. This resulted in many changes to both the code and the resulting html pages. The only change left for the web application is the conversion of to the latest version of the .NET Framework.

We have found UFDC is experiencing timeouts on database server requests, especially when the server is running prescheduled tasks. We are currently examining the use of another database server, and have limited our options to either utilizing SQL Express 2005 on the UFDC Audio-Video server or utilizing the current SQL 2005 server on SMATHERS2K3SQL.

All of this work should complete this week. Continued examination of the application shows that all memory leaks have been corrected, and the application is much quicker than before, due to code (html and ASP.net) changes.

UFDC Builder Emerges!

The new version of the UFDC PreLoader and Loader combines the functions of these two separate applications... as the UFDC Builder. It is expected that this application will run fairly continuously, examining incoming packages at a quicker pace, but still only building the Greenstone search indexes on a daily basis. The process and code is currently undergoing a refactoring process. The currently envisioned process appears as the last page of this document, but highlights of the process change are included below.

With the new UFDC Builder, the incoming packages will be examined more often and metadata update packages will be handled as soon as they arrive with priority throughout the process. Along with corresponding changes to the UFDC application, mentioned below, the new Builder will allow metadata updates to reflect in the online citation information more quickly than changes currently appear. New and replacement items will be handled after metadata updates, but they will also reflect on-line more quickly. While the items and changes will be viewable on-line, they will not be searchable within UFDC until the search indexes are built.

The current system causes there to be discrepancies (some temporary, some permanent) between the structures on the two Greenstone servers. The new builder will treat the production server and greenstone server identically. The only difference will be the server upon which the search indexes are built. The development server will be considered the primary server for this purpose, but the structure and content of the server's data folders will be identical to production.

Each time a collection is built it will be created from scratch, by copying the necessary data files (currently the Greenstone `doc.xml` files) into a newly created `TEMP` collection. Once the indexes are built for this collection, it will be published to both Greenstone servers simultaneously.

Upcoming UFDC Changes

Prior to examination of a new content management system for searching and discovery of resources, all current tickets for UFDC changes have been examined. As many of these are metadata related issues, they will be resolved under the current system (Greenstone), in the hopes of reaching some metadata stability before starting any conversion process. These issues include adding additional fields in the METS file, adding new search indexes, and allowing new sorts of the resulting search results.

- Metadata Issues
 - Add *Journal Title*
 - Hierarchy Geography should have a type attribute (SPR-309)
 - Include TOC in 'full citation'
- Search Indexes
 - dateIssued, ufdc:temporal, ufdc:type, hierarchicalGeographic (SPR-397)
 - Format (Material Type) (SPR-386)
 - Place of Publication (SPR-387)
 - Attribution

- Sorts
 - Sort by Country in Maps (SPR-319)
 - New items sort by date added (SPR-317)
 - Allow thumbnails to be sorted

In addition, basic display issues will be addressed, including the inclusion of serial information in the title box for each item (SPR-309). UFDC will begin to read a shortened, bibliographic METS file to display all citation information. Currently, it relies on the content management system to disseminate the complete record. This allows for more relational data to be displayed, and relieves the content management system from having to be aware of every bit of metadata. This will also allow metadata changes to reflect online more quickly. Making changes to a metadata scheme which does not affect basic search issues will be much simpler.

Several other user interface issues will be addressed, including the addition of alternate search types for newspapers (SPR-311) and maps. Inclusion of a Google map for single items and for search results will be researched. And the use of tree-type rollups will be added to search results (SPR-369), just as it was added to the multi-volume display for large items.

Content Management System

In attempt to improve the quality and speed of searches, an alternate content management system will be examined for indexing and resource discovery. Currently, UFDC is built on a Greenstone data layer, but we will begin to test the use of an underlying Fedora data layer. Research indicates this will drastically improve the search result quality and speed. In addition, it should allow for quicker inclusion of changes into the indexes used for searching.

We hope to finish testing and any necessary conversions by the end of the current year.

Remaining Work

There still remains more work to be done in the future on UFDC. We hope to build an ALTO-aware image server to provide text highlighting. And we also are looking at alternate JPEG2000 server technologies, with the hope to add drag and drop features.

UFDC Application Recent Changes

- Optimization Completed Steps
 - Completed conversion of user controls to classes
 - Corrected a memory leak issue
 - Completed examination of html, view state, and css for speed optimization
 - Completed redesign, based on NDNP usability and work above
- Complete refactoring of the ASP.net code
 - Separation of much of the UFDC code into a library file
- Interfaces
 - Interfaces now language specific
 - Loaded dynamically when needed, rather than being preloaded.
 - Most common interfaces, UFDC (en), dLOC (fr, en, es), are preloaded however
 - Multiple language headers and footers now supported
- Collection Groups, Collections, SubCollections, Institutions
 - Original structure of folders can be replaced with a configuration XML file
 - Application doesn't have to look through structure folders for each possible file.
 - Application doesn't have to peek into the info and browse files to get the codes.
 - Info, browses, and home pages all support multiple languages.
 - Rotating highlights
- Added all translations to the UFDC interface. Now fully supports English, French, and Spanish.
- Added support for www.dloc.com
 - Query URL is parsed. If it is for dloc...
 - 'dLOC' interface becomes the default
 - Collection = dLOC
- Static pages for search engines (<http://www.uflib.ufl.edu/ufdc2>)
- RSS feeds added to UFDC

UFDC Builder Proposed Process

Pre-Building Functions

(Runs fairly continuously)

1. Gets initial settings from the UFDC database and configuration file
2. Step through each of the institutions *inbound* folder
 - a. Check each package for 'acceptability'
 - i. A single METS file package (metadata update) 1 minutes old?
 - ii. A complete package an hour old
 - b. Move the package to the corresponding institution's *processing* folder
3. Step through each of the institutions *processing* folder
 - a. Does the package have a METS file and only one METS file?
 - b. Did the METS file pass the schema test?
 - c. Did the METS file pass the UFDC standard validation?
 - d. Categorize as type of METS file (metadata update, delete, other)
4. Handle all metadata updates first
 - a. Save shortened METS file (bibliographic data only)
 - b. Save Greenstone XML file (with alternate directory for text files)
 - c. Move package to server's image location(s), by primary collection
 - d. Save this package to the database
5. Handle all deletes next
 - a. Save delete to the database
 - b. Move existing package to the *delete* folder on servers
6. Any package in non-UF folder, save copy of entire package. Do this by moving files which would otherwise not be used, and copying all files which will be used.
7. Move all special files to their respective locations (JPEG2000, AV Server, etc...)
8. Handle all remaining packages (new and replacements)
 - a. Save shortened METS file (bibliographic data only)
 - b. Create the attributes.xml file
 - c. Save Greenstone XML file
 - d. Move package to server's image location(s), by primary collection
 - e. Save this package to the database. (If new, set built flag to false)
9. Create Failure Log for all items which failed during the above tests

Building Functions

(Runs daily or nightly)

1. Iterate through each Greenstone collection waiting for build
 - a. Delete TEMP collection, if there is one
 - b. Build TEMP collection
 - i. Copy metadata and configuration files from existing collection or TEMPLATE
 - ii. Copy Greenstone XML for each item linked to collection in database
 - iii. Generate archive.inf file
 - iv. Build this collection through WinSCP script execution
 - v. Adds the collection building log to the UFDC database to indicate the building start time and end time
 - vi. Publish collection by copying to both development and production boxes
 - vii. Save new greenstone name in UFDC database, turn off new package flag, and reset UFDC cache
 - viii. Delete old collection on both production and development

Weekly Maintenance

(Runs weekly)

1. Perform cross-server validation for any newly altered items
2. Regenerate all Greenstone XML files and bibliographic METS files
3. Regenerate all static pages