Digital Access Committee (DAC) Meeting

October 12th, 2010

Today we had a CRRA Digital Access Committee (DAC) meeting via the telephone. Attendees included:

- Ann Hanlon
- Demian Katz
- Eric Frierson
- Eric Morgan
- Kevin Cawley
- Pat Lawton
- Susan Leister
- Thomas Leonhardt

I (ELM) did a bit of “Portal” show & tell demonstrating the work done to date on indexing EAD files. (See the previous blog posting.) We then discussed ways the indexing/display could be improved. Suggestions included:

- putting the words “Archival material” into the format field of the Solr index thus allowing better faceting
- reading the value of langmaterials and using it as the value for Solr’s language fields, again allowing for better faceting
- reading all of the fields associated with a given container-level element and putting them into Solr’s allfields field to improve indexing
- extracting the last value of our current “title”, using it as our title, and using the remaining values as some sort of supplemental description or alternatively, simply reversing the “title” string

We then brainstormed ways to resolve character encoding issues, the feasibility of making our metadata available via Web servers, and the status of the metadata guidelines.

We felt we had discussed it all, so the meeting was over.

Posted in Uncategorized | 1 Comment »

Indexing MARC and EAD in VUFind with Solr for the CRRA

October 12th, 2010

This posting outlines how I am currently indexing MARC and EAD files in VUFind with Solr for the CRRA. (Boy, there are a lot of acronyms in that sentence!)

Background

The Catholic Research Resources Alliance (CRRA) is a member-driven organization with the purpose of making available "rare, unique, and uncommon" research materials for Catholic scholarship. Presently the membership is primarily made up of libraries and archives who pool together their metadata records, have them indexed, and provide access to the index. My responsibility is to build and maintain the technical infrastructure supporting this endeavor.

A couple of years ago much of the CRRA metadata was manifested as MARC, and at that time VUFind was selected as the tool we would use to index, search, and display this content. About six months ago the Alliance realized the growing necessity of including EAD files as well. At the same time, the ability of accomodate non-MARC metadata was increasingly becoming a VUFind reality. New ground still had to be broken; processes needed to be implemented allowing VUFind (and the underlying Solr indexer) to understand how to work with materials which were not book-like.

The balance of this posting describes in greater detail how I am beginning to accomodate MARC as well as EAD metadata into VUFind's interface with Solr.

Assumptions

The system runs on a number of assumptions. First, it is assumed it is the members' responsibility to create and maintain their metadata. Second, it is my responsibility to index it and make it available for display. Moreover, it is assumed each metadata record incudes at least three values: 1) a unique identifier, 2) a human-readable description of an item, and 3) an address pointing to the location of the item. For MARC records, these things reside in the 001, 245, and 099 fields. For EAD files, they have been designated as the id attribute of unitid elements, the content of unititle elements, and the url attribute of the eadid element and from there the location of the item.

Additionally, it is assumed all metadata records, whether MARC or EAD, are available for harvesting from a Web server. In other words, each member who wants to have their MARC records available in the CRRA needs to export their records to a single file and make them accessible via a URL. Similarly, all EAD files which are intended to be indexed need to be in a single Web-accessible directory and the URL of the directory needs to be known. Making member metadata accessible via a Web server has three benefits: 1) it facilitates automation, 2) it distributes the responsibility of archiving metadata across the membership, 3) it enables the metadata to be harvested by other applications and used for other things. "Can you say 'linked data?'"

Files and Perl scripts

Given these assumptions, the following sets of files and Perl scripts are used to do the work. The first set is core the both of the other two:

- libraries.db – A “database” of CRRA participants consisting of their names, libraries, and URLs where their metadata records can be found. This file is used by just about every other script in the system.
- subroutines.pl – A tiny library of Perl subroutines, mostly to read the contents of libraries.db.

This second set is used to index MARC metadata:

- marc-harvest.pl – Copies (mirrors) remote MARC files locally
- marc-add-code.pl – Validates and updates the values of MARC 001 fields making sure they exist and are unique
- marc-index.pl – Slurps up a Solr marc.properties template (template.txt), makes the appropriate substitutions, and indexes the MARC records associated with a given library
- marc-build.sh – A shell script used to run all of the MARC-based scripts. One ring to rule them all.

The third is used to index EAD files:

- ead-harvest.pl – Copies (mirrors) remote XML files locally
- ead-validate.pl – Makes sure the mirrored XML files are well-formed, conform to the EAD DTD, and include an eadid url attribute (done with a stupid stylesheet called geturl.xsl)
- ead-transform.pl – Makes sure each EAD container-level element includes a unitid with a unique id attribute, saves the result to a local cache, and transforms these same files into HTML. The first process is done with a stylesheet called addunitid.xsl. The second process is done with another stylesheet called ead2html.xsl.
- ead-index.pl – Indexes all the cached/transformed EAD files by parsing out container-level elements, creating an XML stream of records of my own design, parsing the result, and passing each record on to Solr. The heart of this script is a fourth stylesheet — ead2solr.xsl
- ead-build.sh – A shell script used to run all of the EAD-based scripts. Another ring to rule them all.

The “secret” to indexing EAD files is really no secret. I simply followed Demian Katz’s instructions. In a nutshell, to index non-MARC content the developer needs to:

- Parse the given metadata into records. I do this with ead2solr.xsl.
- Map each of the record’s values to as many of the underlying Solr fields as possible. Presently I only have titles and I do this through ead2solr.xsl as well.

- Create an XML snippet representing each record and map it to the Solr fullrecord field, described below.

- Denote a record type. I call mine ead.

- Save the whole thing to Solr, done with ead-index.pl.

Currently, my XML snippet (Item #3) looks like this:

```
  <record>

    <id>unaead_id2635150</id>

    <title>Catholic Church. Archdiocese of Detroit (Mich.)

      Collection -- Catholic Church. Archdiocese of

      Detroit (Mich.): Manuscripts -- Letters -- Bp.

      Baraga to his sister Amalia

    </title>

    <date>1836/1203</date>

    <url description='View remote, canonical version of EAD'>


http://archives.nd.edu/findaids/ead/xml/det.xml


    </url>

    <url description='View local version of EAD file'>


http://zoia.library.nd.edu/sandbox/crra-data/ead/una-det.html#id2635150


    </url>

  </record>
```

The VUFind application provides seamless access to the index through its search box, but a bit of work needs to be done to display search results. Specifically a “record driver” needs to be written to accomodate new record types (Item #4, above). This driver inherits methods from a parent driver, IndexRecord.php, and the developer needs to override some of the methods found there with methods considering the content of the fullrecord field. Presently, the only thing I have in my record driver (EadRecord.php) is a method to extract URLs. In the future I will need to include methods to extract names of CRRA members, names of their libraries, and additional descriptive metadata.

You can see the fruits of these efforts in the CRRA "sandbox" — something we are affectionately calling "The Green Interface".

Issues

The whole process functions and could be run automatically from cron on a daily basis, but there is plenty of room for improvement. Issues include:

- **speed** – The indexing process is slower than I'd like. I think throwing more hardware thrown at the problem will make things faster.
- **invalid data and stale URLs** – A small percentage of the MARC and EAD files do not include the required metadata values. No unique identifiers. Malformed MARC leaders. Non-validating EAD files and/or eadid url attributes pointing to broken locations. This is where metadata maintenance comes in.
- **character encoding** – This is one of the bigger problems. Trying to figure out whether or not a MARC record has been exported as UTF-8 is difficult. Solr assumes UTF-8 and I don't think it even knows about MARC-8. When MARC data is not encoded as UTF-8, search results look really ugly. Similarly, some of the EAD files, because of similar issues, really display poorly after they have been transformed, indexed, searched, and displayed.

None of these things are insurmountable. They will be addressed.