

Practicum Final Report

Parmit Chilana

Term: Summer 2006

Supervisor: Tim Cole

Location: Grainger Engineering Library

My role in this practicum has been to understand metadata reprocessing and augmentation techniques and how they can be applied to the Institute of Museum and Library Services Digital Collections and Content Collection (IMLS-DCC) project. The preliminary strategies consist of various scripts and XSLT transformations that have been adopted from the UIUC CIC metadata portal (Foulonneau & Cole, 2005). While many of the steps are generic and are applicable to all collections, there is often need to customize the transformations according to the nature of the collection. I have mainly contributed to this project by understanding and documenting the reprocessing steps in detail and carrying out metadata analyses of several of the harvested collections which serves as a pre-step to the processing.

Metadata Reprocessing

The Steps

There are a number of steps which have to be carried out in order to reprocess harvested metadata records. Many different scripts, xml files, xslt files, and directories are involved in the processing in a very specific order. All of these are explained in the following document: <\\libgrtyr\harvests\Documentation\parmit\ProcessingSteps.doc>

The main idea is to first carry out metadata analysis on each collection (as described in the next section) to determine what type of metadata cleaning and normalization is required.

Before the filtering steps using stylesheets are carried out, new directories have to be created and xml files have to be updated with new information about the collection and repository (i.e. `repositories.xml`, `collections.xml`). The `types.xml` also has to be updated based on the `dc:type` values in a collection.

The filtering steps consist of a number of modifications which have to be made to the following stylesheets depending on the format and quality of metadata corresponding to a collection: `cleaning.expand.xml`, `collection.expand.xml`, `standard.expand.xml`. In some cases, the `standard.expand.xml` stylesheet has to be customized according to a specific collection and has to be renamed accordingly (i.e. using the collection's code name). Generally, a custom stylesheet is used in cases where there are multiple identifiers and the correct has to be selected according to specific criteria, or removing occurrences of certain text, such as "Unidentified", or normalizing names which appear in many variations. If thumbnails are produced for a collection, the stylesheet `thumbnail.deliver.expand.xml` has to be included in the processing order as well. Lastly, there are three other stylesheets which usually remain consistent in all of the processing: `topelements.xml`, `normalizens.xml`, and `removeDuplicates.xml`.

Example Collections

I observed the re-processing steps adopted from the UIUC CIC portal as they were used on the Illinois Alive collection. I carried out these steps on two other collections: King County Snapshots (Washington) and American Missionary Association (Tulane).

All of these collections had accompanying thumbnails which were generated using the Thumbgrabber program and techniques used for the UIUC CIC collections. We learned that while metadata processing generally occurs after the thumbnails have been produced, in some

cases, it is necessary to first re-process a collection before Thumbgrabber can accurately generate the thumbnails (these are explained in the Processing Steps document). The whole process is then repeated, which may a bit inefficient and it should be investigated if there is a way to bypass this repetition of steps. Another problem with Thumbgrabber which we encountered was that in some cases thumbnails of the webpage were being produced rather than the image. To resolve this, we learned that we have to set the `img_only` parameter to 1 when the Thumbgrabber jobs are executed.

Using the metadata analyses for these collections, specific changes were made to the cleaning stylesheet and collection-specific customized stylesheets (`content.lib.washington.edu.expand.xsl`, `louisdl.louislibraries.org.expand.xsl`) were produced, using the standard stylesheet as the basis. For example, the Washington collection contained records where the `dc:coverage` fields consisted of multiple values separated by either `
` or `--`. Hence, in `cleaning.expand.xsl`, I used the `split` function specifically for this collection to split by these delimiters. One problem with this was that we could not pass in `
` as a value in the XSL file (because of XML encoding issues) and we were not sure if the `DELIMITER` parameter of the `split` function accepts strings or just characters. To verify, we looked at the `functions.xsl` file (under [\\libgrtyr\harvests\IMLSHarvest\filters\include](#)) and discovered that it does accept string values so we could use `
` as a value in the `DELIMITER` parameter. We could also pass in the second delimiter, `--`, in the `OTHER_DELIMITERS` parameter.

Another example of a problem which we encountered was with identifiers in both the Washington and the Tulane collections. Records in the Washington collection had one URL as an identifier and one non-URL. All records in the Tulane collection also had 2 identifiers,

but both were URLs, and only one was pointing to the record correctly. Based on similar experiences with the UIUC CIC collections, we were recommended to write tests in the collection-specific stylesheets to deal with both situations. For the Washington collection, we wrote a test for identifiers which would select the URL-based identifier (i.e. `<xsl:if test="contains(.,'http:') or contains(., 'https:')">`). For Tulane, we differentiated the correct URL by selecting a consistent property among these URLs in that they all contained `u?` (i.e. `<xsl:if test="contains(.,'u?')">`)

All of the changes and additions to stylesheets were first made locally and tested out using MSXSL, a command-line based program used for running XSL. The specific instructions for testing out these stylesheets (and their order of processing) are explained in this document: <\\libgrtyr\harvests\Documentation\parmit\commandLineXSLT.doc>. After verifying the output locally, the changes were transferred over to the server and the `dispatchManager.xml` had to be updated. Also, a new batch file specific to the collection had to be created and executed (as explained in the processing steps document).

There is a final indexing step for loading the augmented records into the database, which is explained in this document:

<\\libgrtyr\harvests\Documentation\parmit\IndexingSteps.doc>

Metadata Analysis

For the UIUC CIC collections, metadata analysis has been carried out by running SQL queries on the harvested metadata stored in a database. For IMLS-DCC, the same procedure has been adopted. Previously, there had been no documentation on what type of queries are used for this analysis, so I've organized all the queries in this document (some have been slightly modified): \\libgrtyr\harvests\Documentation\parmit\Queries_analysis.doc. These

queries are used to get an overview of what type of metadata fields are used in a collection and to inspect individual metadata fields and values.

Example Collections

I carried out the above analysis queries on the following collections and they are available here: <\\libgrtyr\harvests\IMLSHarvest\analysis>

- SNAD
- KingCounty (Washington)
- Illinois Alive
- Tulane
- PALMM
- KMODDL
- ACNATSCI
- InfoMine
- MSU
- Maine
- UNC
- Perseus
- CF Memory
- Colorado
- Wisconsin

It was very useful and interesting to look at similarities and differences in the types of metadata fields and the level of granularity that is used in these collections. It was also interesting to see the variations in how the corresponding values are encoded and formatted. Hillmann et al. (2004) discuss their experiences in improving harvested metadata quality with the National Science Digital Library project and mention that there are four categories of problems which they have experienced with harvested metadata:

- ***missing data***, where certain metadata elements are missing in the supplied metadata
- ***incorrect data*** – where the values do not conform to standard element use
- ***confusing data*** – where multiple values occur in a single metadata element and include embedded html tags and other confusion
- ***insufficient data*** – where there's no indication of use of any controlled vocabularies

Based on the analysis which I carried out, I can attest that these are some of the most common problems. I've organized the problems which I encountered according to these 4 categories below. Some of these problems have already been dealt with based on the UIUC CIC experience, but attention should be paid to some of the unique metadata challenges in the IMLS-DCC collections.

Missing Data

- In many collections, the query results show that there are values for fields such as source or format, but when we look at these values, they are blank or empty strings "".
- Often words such as None or Unidentified are also used, but not consistently
- In some cases, identifiers were missing from records
 - Appropriate changes were made in the `filtering.vbs` and `filtering.wsf` files to filter out such records

Incorrect data

- In some instances, there was no consistency in the type of information recorded in the format and type fields. Often, information such as dimensions of images was listed under 'type'. In some cases, the description field had information related to the type and formats in random order.
- There was a lot of variation seen in several collections in the ways names were encoded. For example, there were cases of the same creator appearing 3 or 4 different times because of extra suffixes at the end of names or having endings such as periods or semicolons or mistakes in the spelling. The same situation was seen in subject headings, rights statements, contributor names, format fields and language fields. In the date fields

as well, there were extra brackets ‘]’ or commas ‘,’ at the end. *Currently, there is no normalization in place to deal with these situations.*

- For identifiers, there were some instances where URLs were not correct or there were only internal local identifiers used which have no meaning beyond their original context of use

Confusing data

- There was no consistency among the formatting of date values in many collections. Dates were in YYYY or YYYY-MM-DD, or ca. YYYY and several other formats. The existing steps for reprocessing dates used in the UIUC CIC metadata portal were carried over, but attention should be paid to any other peculiar instances which are not covered by this reprocessing.
- In several instances, multiple values appeared in a single line (i.e. for subjects, creators, contributors, publishers, dates, etc.). These are usually separated by one or two delimiters consistently (i.e. --, ;,
, etc.)
- In some cases, embedded html tags, such as
 appeared in metadata fields, which we deleted if they were not being used as delimiters
- There were also instances of inclusion of foreign characters in a couple of the collections which were difficult to translate

Insufficient data

- In some collections, for fields such as subject and format, it was obvious that there was no controlled vocabulary used and there was a great deal of inconsistency, even if the formatting was appropriate. A handful of records within the same collection may have subject headings conforming to controlled vocabulary, but within the same collection

there would be other records which either have no subjects associated or have something very generic.

Conclusion

Overall, this practicum was a useful experience for me. I got a chance to apply the skills which I have developed through coursework related to digital libraries and also got a chance to enhance my knowledge in the area of XSLT transformations. It was a valuable experience in dealing with the challenges of metadata reprocessing and augmentation and in knowing that there are numerous intermediate steps required for dealing with unique cases.

References

Foulonneau, Muriel & Cole, Timothy W. (in press). Strategies for reprocessing aggregated metadata. In *9th European Conference on Digital Libraries, ECDL 2005*, September 18-23, 2005, Vienna, Austria. (Proceedings Series: *Lecture Notes in Computer Science*.) Heidelberg: Springer-Verlag.

http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/1155136_2_26

Hillman, Diane, Dushay, Naomi & Phipps, John (2004). "Improving Metadata Quality: Augmentation and Recombination," in *DC-2004: Proceedings of the International Conference on Dublin Core and Metadata Applications*, Shanghai, China, October, 2004.

<http://www.cs.cornell.edu/naomi/DC2004/MetadataAugmentation--DC2004.pdf>