Livingstone TIFF and XML Data Review – Results

Adrian S. Wisnicki (23 December 2013)

Goal: Between September and December 2013, Ashanka Kumari (my RA at UNL) and I reviewed the TIFF/JPEG and XML data produced by Livingstone Online between 2004 and 2013. Our goal was to identify the master TIFF images (or, if not available, master JPEG images) and XML transcription files that will become the basis for the new LO project site to be created by LEAP.

For our review we used the following data sources:

1) The Clendennen and Cunningham catalogues (1979, 1985), which sequentially list all Livingstone letters known at the time of publication.

2) The original unprocessed, uncropped TIFFs and JPEGS collected by LO and supplied by Chris Lawrence. These were initially on a series of CD-ROMs.

3) The processed JPEG images to be found in the live version of LO, but as converted to canonical number file names via spreadsheets provided by Frank Smutniak (programmer, UNL).

4) The live LO catalogue image and transcription data (http://www.livingstoneonline.ucl.ac.uk/view/list.php).

5) The TEI P4 XML files supplied to me by Caroline Overy (former RA for LO) as converted to canonical number file names via spreadsheets provided by Frank Smutniak.

6) The TEI P5 XML files supplied by Heather Ball (our current Head RA).

7) The LO MySQL "records" and "foliation_code" databases, via spreadsheets provided by Lisa McAulay (librarian, UCLA).

Exclusions. We did not use the live TEI P4 XML files because these are pre-transformed fragments and should only be relied upon as a last resort. We also excluded non-document images ("picture data") found in the unprocessed TIFF/JPEG data set; this data is being handled in a separate workflow.

We summarize our findings in three spreadsheets, which are provided along with this summary:

Canonical Letters. These are the Livingstone letters recorded and/or not recorded in the C&C catalogues about which we have data in some form. We have decisively identified all the letters here. The major discovery is that we have some 900 items as either images and/or transcriptions -- 300 more than we suspected. The other types of items incoming

from the DLC and NLS (in scope) as well as the 100 additional letters incoming through Jared's South African letters project (out of scope) are not recorded here.

Stray Items. Just under 40 letters that have not yet decisively been identified. Some of these should be easy to identify once I get to them. Others may require more work or, in the case of the SOAS letters, an in-person visit.

Other Items. Images of Livingstone items that are not letters; documentation files; images of non-Livingstone items. Some of the Livingstone items come from the DLC and will be digitized by Glasgow in better quality; other Livingstone items should be taken in via LEAP. The non-Livingstone items fall outside of the scope of LEAP.

Each spreadsheet includes the following column categories:

XML/TIFF Directory File Name. The TEI P5 XML files use canonical C&C numbers. The TEI P4 XML and TIFF image files use a bespoke, non-systematic file naming scheme. A good portion of the P4/TIFF files, however, are named using elements from the MySQL database and so can easily be identified and converted to canonical numbers (or our final file naming scheme). The exceptions to this last rule are identified in this column.

Letter. The canonical C&C letter numbers. Only those letters for which we have some form of data are listed here. If a letter is not catalogued, it is described as "Uncat" followed by the recipient (if known) and the date of production.

TIFF/JPEG Core Files. This column records the survey of the unprocessed TIFF/JPEG image data as well as the incoming DLC data. An "x" indicates that the images for a letter are found among the unprocessed data. A red box indicates that image data for the given letter is not found among the unprocessed data; image data from “Live-Images-Remapped (JPEGs)” category (see next) may be used to provided image data missing from this column. A "G" indicates that this image data is incoming from U. Glasgow. A "DLC" indicates that this data (scans of photocopies) is incoming. A few sources, marked "x / G" still require resolution.

Live-Images-Remapped (JPEGs). A record of the letter image sets found in live site.

XML. A record of the available TEI P4 and P5 transcriptions. Missing transcriptions can be recovered as a last resort (see above) from the live site. However, in those cases it may be easier to redo the transcriptions from scratch (if we choose to redo them at all as part of LEAP).

Records/Foliation Code. A record of the data found in the relevant MySQL databases. On the whole, the “records” databases is more full and duplicates the “foliation_code” database. In these cases an "x" is placed in this column. In a few instances, however, the "foliation_code" database contains unique information. In those cases "foliation-code" is placed in this column.

Notes. Additional notes for me.

Results. These three spreadsheets for the first time correlate all the existing LO image and transcription data produced between 2004 and 2013 and will serve as a roadmap for UCLA's intake of these materials. The process will be streamlined by the fact that, using the MySQL "records" database, Frank Smutniak with assistance from Lisa McAulay and me has already produced MODS metadata records for all the C&C catalogued Livingstone letters. The major finding of the spreadsheets, in addition to the 300 additional letters, is that indeed all the data sources are needed because each, in different cases, contains unique and important data about what we have or should have (but don't).

Error Rate. This survey has drawn on seven data sets to produce some 1000 item records. Some of the work could be done in an automated fashion; some had to be done manually. Allowing for a fairly conservative error rate of 1% to 5% suggests that some 10 to 50 of the records in these spreadsheets have errors in them. It is hoped that at least some of these errors will be identified and corrected through the UCLA intake process.