LEAP Interim Data Report

Adrian S. Wisnicki (2 April 2014)

This report continues the work of the "Livingstone TIFF and XML Data Review – Results" report I distributed in December 2013. Since that time, the preparation of the legacy *Livingstone Online* image and transcription data has continued. This new report summarizes the results of that preparation. Our efforts have focused on four areas:

1) further organization of the TIFF manuscript data,
2) review and renaming of the XML TEI P4 and P5 transcription data,
3) review and expansion of the *Livingstone Online* bibliographic records, which have also been mainly converted into the MODS metadata format, and
4) preparation and organization of the illustrative image data.

This report summarizes our results in all these areas. The report concludes with a summary of the remaining steps needed prior to ingest of all our data by UCLA.

---

1) TIFF manuscript data

1.1) Ashanka Kumari and I, with assistance from Kate Simpson and a few archivists at collaborating institutions, have fully and successfully completed the interim preparation of the TIFF manuscript data, which is now available from the UNL Spacely server:

[redacted]

We have detailed the results of our work in the attached Interim TIFF Manuscript Data.xlsx file, in three tabs – Go Items, Review Later, and No-go Items. These three tabs correspond to the three subdirectories of the Spacely master directory.

1.2) The Interim TIFF Manuscript Data.xlsx file does not include items incoming through LEAP from the NLS and DLC (except DLC letters, which are included). The lists also do not include the images for 46 Livingstone items from the Brenthurst Library, Johannesburg that we have recently acquired. These latter images have been created to LEAP standards with appropriate LEAP file names, and will be used later in the project to test the automatic ingest of new data by *Livingstone Online*.

1.3.1) Within the Go Items and Review Later subdirectories, we have organized the files as follows:

CC-number-directory > TIFFs/JPEGs-directory > letter-images-with-original-file-names

Where CC numbers are not available, we have used a naming scheme outlined elsewhere. [1]

1.3.2) The organization and naming in the No-go Items directory is more ad hoc and reflects the fact that we have no further plans to work with this data.

1.4) The Go Items tab records that we have or will have by the end of LEAP publishable images and/or transcriptions of:

a) 837 previously catalogued letters,
b) 35 previously uncatalogued letters,
c) 32 additional catalogued and uncatalogued items, and
d) 3 contextual items.

1.5) The Review Later tab records that we have:

a) 2 "problematic" letters, one from the NLS, one from SOAS, for which we are missing an image. We should be able to obtain these images in due course;
b) 12 items for which we either have never had permission to publish or for which we have images that are not suitable for online publication. If time and resources permit, we may pursue permission and/or improved images of these letters;
c) 58 items (32 of them unacatalogued) that we found through online auction sites for which we have images and additional bibliographic data. We cannot release these items publicly, but will hold them "behind the scenes" to assist future researchers;
d) 5 letters for which we have copies or alternate versions.

1.6) Finally, the No-go Items records that:

a) We have images for a variety of items from the DLC and elsewhere, both by Livingstone and by others, that are either not suitable for online publication, will be redone through LEAP in a better quality, or fall outside the scope of LEAP. The DLC images in this category have been passed onto the DLC for their archival purposes.

1.7) Summary of TIFF Manuscript Data Preparation:

Despite the complexity of the *Livingstone Online* TIFF manuscript data, the overall findings are very good. We have or will have archival TIFF images for nearly all items on the "Go Items" list.

As a result, we will be able to deliver images and/or transcriptions for a total of 907 Livingstone or Livingstone-related legacy items through the new *Livingstone Online*. This is over 300 more items that anticipated and represents a major accomplishment of LEAP. Likewise, the discovery of 58 auction items (32 unacatalogued) for which we now have

[1] The Key to Interim Legacy Data.xlsx file, also attached. In addition, this file explains the shorthand and color coding used in the main Interim Legacy Data.xlsx file.

images, even if we cannot provide access to them other than through external links, also represents a major LEAP accomplishment.

1.8) We are only missing TIFF manuscript data for:

a) the Pyne letters (Chris Lawrence does photocopies of most of these, and we will scan them in due course),
b) one set of NLS letters (which we may be able to obtain in due course),
c) nearly all the letters from the RGS (we have already contacted the RGS about securing new copies), and
d) a handful of scattered letters.

1.9) In preparing the interim data set, Ashanka and I discovered and corrected a variety of errors that had crept into our previous, preliminary data set. As a result, it is probably fair to set the error rate in the interim data set as between 1% and 2%.

---

2) XML TEI P4 and P5 transcription data

2.1) Ashanka and I reviewed and renamed the legacy XML TEI P4 and P5 transcription files, with some automated renaming done by Frank Smutniak. We renamed the files to the appropriate CC number names, then checked these names against the CC number values provided within the XML files themselves.

The results of our review and renaming are detailed in the Interim Legacy Data.xlsx file. The XML transcription files themselves are on the UNL Spacely server:

[redacted]

2.2) Within the Spacely XML master directory there are two subdirectories, one with the P4 files, one with the P5 files. The two subdirectories are divided further as follows:

2.2.1) The files in the P4 directory are within subdirectories that reflect level of finalization. Documentation that explains this organization is here:

[redacted]

2.2.2) The P5 files are split between one subdirectory for renamed letters and one for files from the *Livingstone Spectral Imaging Project* (not renamed and not enumerated in the Interim Legacy Data.xlsx file).

2.2.3) The subdirectory for P5 letters is further split into two original directories, one for completed transcriptions, the other for transcriptions in the penultimate stage of finalization.

2.3) Summary of XML Review and Renaming:

a) 525 TEI P4 files, created between 2005 and 2010;
b) 98 TEI P5 files, created between 2011 and 2013; and
c) 55 TEI P5 files from the *Livingstone Spectral Imaging Project* (2010-2013).

2.4) In other words, we have a grand total of 678 XML transcription files, which will all be converted to the same TEI P5 standard and integrated into the new *Livingstone Online* through the work of James Cummings for LEAP.

2.5) We found no major issues or errors during the review of the XML files and, given the relative simplicity of the review and renaming, it is fair to estimate a 1% to 2% final error rate.

---

3) MODS metadata

3.1) Frank Smutniak, with assistance from Lisa McAulay and me, successfully converted the *Livingstone Online* MySQL database (itself based on the CC Catalogue) into a standalone file.

This file, LO legacy MODS.xml, attached, captures all the important data found in the MySQL database plus, for each item, adds the LEAP file name. The file contains 2,220 individual MODS records.

3.2) After conversion, I reviewed the LO legacy MODS.xml file and added updated bibliographical information for a selection of items. These are marked in the MODS file with <!-- change -->.

3.3) With the assistance of Ashanka Kumari and Megan Ward, I also created a spreadsheet, MODS Additions-Auctions.xlsx, attached, of new data for 110 original Livingstone items, 67 of which had not been catalogued previously. I discovered the new items through review of the legacy image data plus study of a range of online auction catalogues and other sources.

3.4) Summary of MODS review and expansion:

The records in the LO legacy MODS.xml file are ready for ingest by UCLA. The MODS Additions-Auctions.xlsx file is ready for conversion to proper MODS records. The significant correction and expansion of the Livingstone bibliographic record represents another major accomplishment of LEAP.

---

4) Illustrative image data

4.1) Ashanka Kumari and I also prepared and organized the illustrative image data (a.k.a. the "picture" data) in two stages:

4.2) First, we assembled all the illustrative data into a single master directory, which is now available on Spacely:

[redacted]

4.2.1) In organizing the illustrative data, we sought to identify master images (TIFFs or JPEGs) and removed duplicates and lesser quality images. For the most part, we left original file names unchanged. As provided, the illustrative data was mixed with the TIFF manuscript data, so separating one data set from the other became a major component of our work.

4.2.2) The illustrative data, roughly speaking, falls into three categories:

a) Images of manuscripts, illustrations, or objects that somehow contextualize Livingstone and his manuscripts;
b) Images taken in and near repositories that hold Livingstone manuscripts; and
c) Images of the *Livingstone Online* team, often in repositories with Livingstone holdings (There is some overlap between this category and the previous one.)

4.2.3) The master illustrative data directory also contains one subdirectory, Texts, that holds images of small drawings that Livingstone made in his letters. The legacy P4 transcription files reference these drawings, so there may be the opportunity to use the drawings in the new *Livingstone Online* site.

4.3) Second, we created a spreadsheet, Illustrative Data.xls, that captures all the available metadata for the illustrative data in MODS format (this does not include the files in the Texts subdirectory; see 4.2.3), and so that will allow for easy export of the data into proper MODS records. A partial, previous picture metadata file created by *Livingstone Online* plus a MySQL database of picture data significantly facilitated our work in the creation of this new file.

4.3.1) The Illustrative Data.xls file includes the master image file name, a description of the given image, and the image credits. We were able to identify all the illustrative images and to provide the appropriate image credit, excepting a series of pictures from the Wellcome Library for which we do not have the original source information.

4.3.1.1) The Illustrative Data.xls file names highlighted in red are those for which we could not find a master image, and so had to turn to a lesser quality image taken from the live *Livingstone Online*.

4.3.1.2) Illustrative Data.xls file names highlighted in green indicated images that were taken from the "DLC Digital Catalogue," an unfinished PowerPoint project by Gary Li (former *Livingstone Online* photographer), found among the legacy image files.

4.4) Summary of Illustrative Data Organization:

a) 485 illustrative images accumulated during the 2005-2010 phase of the *Livingstone Online*
b) 46 illustrations taken from Livingstone manuscripts during this same phase and held in the Texts subdirectory.

4.5) The above numbers do not include the significant amount of illustrative data produced and collected by me between 2010 and the present, both for the *Livingstone Spectral Imaging Project* and for *Livingstone Online*.

4.6) Our illustrative data is a major project asset and will allow us to add a significant visual dimension to both our new site and the our blog (http://livingstoneonline.wordpress.com) where, indeed, we have already begun to use some of the illustrative images.

---

5) Final steps prior to data ingest by UCLA

5.1) Although the legacy data and metadata is now in a well-developed state, we have to take a few *final steps* to prepare this data prior to its ingest by UCLA.

5.2) TIFF manuscript data.

5.2.1) The following two steps should be done in sequence, although this needs to be confirmed.

a) Frank Smutniak needs to rename the TIFF manuscript data – both the CC directories and actual image files – to the agreed LEAP file naming scheme. This process can be automated and should be relatively straightforward.
b) After renaming, UCLA colleagues will need to crop a small subset of the TIFF manuscript data so that each image corresponds to one manuscript page (there are several instances where one image contains two manuscript pages). The procedure for this has already been reviewed and agreed upon between me and Lisa McAulay.

5.3) XML TEI P4 and P5 transcription data.

5.3.1) The following two steps should be done in sequence, although this needs to be confirmed.

a) Frank Smutniak also needs to rename the XML data to the agreed LEAP file naming scheme. Again, this process can be automated and should be relatively straightforward.
b) James Cummings needs to convert the P4 and P5 files to one common P5 standard. This process will be carried out through a separate LEAP workflow.

5.4) MODS metadata records.

5.4.1) The LO legacy MODS.xml file is ready for ingest by UCLA.

5.4.2) However, Frank Smutniak needs to convert the data in the MODS Additions-Auctions.xlsx file into proper MODS records prior to ingest of the images.

5.5) Illustrative data.

5.5.1) The illustrative data is ready for ingest by UCLA.

5.5.2) However, Frank Smutniak needs to convert the data in the Illustrative Data.xls file into proper MODS records prior to ingest of the images.