Livingstone Online – Site Notes, Data Summary, and LEAP Data Mgmt Task List

Created by Adrian Wisnicki, 6 Sept. 2013

1) LO Site Notes

Livingstone Online (live version, http://www.livingstoneonline.ucl.ac.uk/) builds on a combination of HTML and PHP. The project was designed and implemented with no input from me. Access to the letters is through this page, which is powered by a MySQL database (http://www.livingstoneonline.ucl.ac.uk/catalogue/view.php).

This database provided access to and correlates pretransformed XML transcriptions of Livingstone letters (I also have all the original XML TEI P4 files, see below) and JPEG images (I also have TIFF versions of most of the images, see below).

There is no documentation for the site, other than what the previous programmer sent to me in emails (see below). The TEI guidelines provided on the site were not fully implemented in practice.

Livingstone Online (Drupal). Over the last year, a graduate student and I converted the LO site to Drupal 7.5, excepting the MySQL database, conversion of which was beyond our skill level. The main Drupal landing page, just so you can see it, is: [Removed] The page displays errors in the Livingstone Online banner unless you view it in Mozilla, preferably from a PC machine.

One of the problems with the Drupal site is that currently there's no way to navigate between the pages (Karin and I will be working on creating header and sidebar menus), so you need the URL for every given page. Here are a few samples (all static HTML) just so you can see them and get a feel for the new site:

[Removed]

They key page for the site that I would like to develop over the lifespan of the grant is the access page:

[Removed]

This page, as I explained in our meeting this week (and to Lisa last week), would allow users to manage and organize the data in a variety of ways and so maximize the opportunities for different kinds of users to take advantage of the site.

To start, I'd like to implement the Drupal site locally at UNL (I have all the files for it) and then work on it from there with Karin, rather than the textmend site. In time, the site would be transferred to UCLA for further development.

Separate from Livingstone Online is the **Livingstone Spectral Imaging Project** (http://livingstone.library.ucla.edu/), which we now plan to integrate with Livingstone Online through this grant. The underlying data for the multispectral critical editions can be found in the spectral image archive (http://livingstone.library.ucla.edu/livingstone_metadata/). You can also view http://livingstone.library.ucla.edu/livingstone_archive/, which is the correct URL for the

archive but currently is not working through the UCLA site for some reason. Once you drill down, the archive contains a series of subdirectories, each of which corresponds to one manuscript folio of a given Livingstone diary and contains all the relevant spectral images, plus metadata, checksums, and, where they exist, XML transcription files, as in this example:

http://livingstone.library.ucla.edu/livingstone_archive/Data/DLC/DLC297b_151-144_011r/

The spectral image archive, as I mentioned, has a single directory for each manuscript folio, but for the Livingstone Online data we'll want to group all the folia of each letter in single separate directory. Also, for LO data we'll have only one TIFF image per folio rather than several as we have for the diary, and one XML file for one set of letter folia.

Our goal, then, with the new project will be to organize and develop the LO data into an archive that can be integrated with the spectral image archive.

2) LO Data Summary

Keith has now loaded all the relevant Livingstone Online files to Spacely, to the following directory: [Removed] When giving subdirectories below, I presume you're starting from this main directory.

What follows is a rough directory by directory overview of what's in the main Spacely Livingstone directory:

01c LO TIFF data

This directory contains 25 subdirectories, most of which correspond to a single institution which has provided LO with manuscript images or other images linked to Livingstone objects, etc. Scattered in here are also images of the LO Team 2005-2010 at work. There is no order to the images in these subdirectories and no file naming in the majority of cases that gives any clue to the content of the images (I think there are somewhere on the range of 10,000 images). Sorting this out is probably the major data management challenge of the project, particularly as a good chunk of the work will have to be done manually. Brian and I need to explore this image data in detail in order to develop a plan for sorting this out and for determining what can be automated.

01 Livingstone Online

This directory contains all the other files I have associated with the site.

01 Livingstone Online\00 LO WT Preliminary

Ignore this directory and all subdirectories.

01 Livingstone Online\01a Site Files

The subdirectories in this directory contain the different site iterations of Livingstone Online over time. The subdirectories with the 03 prefix contain the current Drupal site. We may also want to

review 01a (the original iteration of the live site when I started working on it in 2010) to see what's of use to us there. The files ucgaw3l_content and ucgaw3l_locat may also be of interest because they're the MySQL dumps, but we can likewise go to the original MySQL database.

01 Livingstone Online\01a Site Files\03 LOtest Backup\FTP_clone

The Drupal site.

01 Livingstone Online\01a Site Files\03 LOtest Backup\LO_Catalogue

This subdirectory contains all the jpeg images used to power the live site and two versions of the preformatted XML files used to power the live site. Everything is organized by individual letter directories, the names of which directories are linked to the bibliographical metadata through the MySQL database (see above). The file names in these letter directories do not have any kind of consistent naming scheme, but the file names, in particular when URIs, may allow us to correlate the original TIFF images with the relevant database entries.

01 Livingstone Online\01b LO Tech Specs 2005-10

The three word files in this directory are all the documentation that there is (taken from emails to me) produced by the original developer of the site.

01 Livingstone Online\01c Permissions

Ignore. Publication permissions.

01 Livingstone Online\02 TEI P4 Transcriptions

The TEI P4 transcriptions produced by the LO team before I came into the picture (2005-2010) as well as various notes. We will use the XML files in this directory rather than those that are live.

01 Livingstone Online\02 TEI P4 Transcriptions\01 Livie

Inside this directory, there are five subdirectories, which contain the actual XML P5 transcriptions in various stage of completion. The subdirectory names make it obvious which stage any given set of transcriptions is in.

01 Livingstone Online\03 TEI P5 Transcriptions

This directory contains all the TEI P5 transcriptions produced under my guidance between 2010 and the present.

01 Livingstone Online\03 TEI P5 Transcriptions\02 H.Ball Transcriptions\LOTrans_2013-06-24

This subdirectory contains the most recent version of all the TEI P5 transcriptions created under my guidance and will be the basis of our project, theoretically. I say theoretically, as I think my lead research assistant is about to hand me a more recent version of this directory and all its contents. The numbered files in this subdirectory are arranged like the TEI P4 subdirectories, with directory names indicating stage of completion.

01 Livingstone Online\04a Images

This goes with the 01c LO TIFF data directory described above and contains all the images acquired by the site since I started working on it. Again, organization is by institution. Inside the subdirectories, there are various notes about the images here and those previously gathered that will be useful to us in sorting out all the images.

01 Livingstone Online\04b Image Sorting Notes

More notes that will be useful to us in sorting out all the images.

01 Livingstone Online\04c South Africa Project

Ignore. This relates to a new project we're developing on the Livingstone letters in South Africa.

01 Livingstone Online\05 The New LO

Ignore most of the directory, except for the one subdirectory I describe directly below.

01 Livingstone Online\05 The New LO\04 Databases in XML

These are all the MySQL databases, converted into XML. As I've already discovered with Brian in looking at these, there are some issues here and we'll probably need to go back to the MySQL database for the database information.

01 Livingstone Online\06 XSLT

Ignore. Various XSLT created as part of my (not so successful) attempt to learn XSLT.

01 Livingstone Online\07 Research Articles

Ignore.

01 Livingstone Online\08 Analytics

Ignore. Google analytics for the site. These numbers were much higher when the site was actively developing in 2005-10.

01 Livingstone Online\09 Misc Livingstone Items

Ignore.

01 Livingstone Online\10 topicClouds

Topic clouds created by Matt Jockers based on the TEI P5 (or maybe P4--I can't remember now) transcription files. I'll be calling on Matt to help with the subject word part of the access page (see above).

01 Livingstone Online\11 Undergraduate Fellows Reports

Ignore.

01 Livingstone Online\12 IUP LO Photos

Photos of me and students at my previous university working on the site. We'll want to integrate these with the project team photos found in 01c LO TIFF data (see above).

01 Livingstone Online\Special action files

Ignore.

01 Livingstone Online\Copy of LO Codes

Ignore. Obsolete file.

01 Livingstone Online\Copy of LO Codes

These are the passwords needed to access various parts of the UCL (not UCLA, take note) Livingstone Online site. It's probably best to review all the parts of the site with me at your side.

You can ignore the rest of the "loose" files in the main 01 Livingstone Online directory

3) Data Management Task List

So that brings us to what I think are the main data management tasks for the project (most of which will be carried out by UNL and James Cummings) based on the above:

1. Develop and implement a plan for sorting out the manuscript and picture image data in 01c LO TIFF data as well as the image data I have acquired subsequently.

2. Integrate the TEI P4 and TEI P5 Livingstone Online files. These files also need to be integrated with the Spectral Imaging Project TEI P5 files. All this work, as well as creating an ODD, will fall to James Cummings (see the project narrative), but direction should come from UNL.

3. Correlate the TIFF manuscript images and TEI transcriptions with the relevant metadata found in the MySQL database.

4. Develop a robust file naming scheme and directory structure for organizing and linking the TIFF image data and XML transcriptions.

5. Create appropriate and consistently structured metadata for the TIFF images and XML transcriptions.

6. Assemble the LO TIFF images, XML transcriptions, and metadata into an archive, that itself can eventually be integrated into the spectral imaging archive.

7. Redevelop the MySQL database (or create a new database based on that one) for Fedora that, through the Drupal interface, will provide access to the new JPEG versions of the TIFFs (UCLA will need to generate new JPEGs to a consistent standard) and dynamically transformed versions of the XML files (UCLA will need to develop the XSLT).

8. Harvest the XML transcriptions so as to allow users to sort and manage the data through the Drupal interface access page (I imagine starting this work with Karin, then carrying it on with UCLA, with assistance also from James Cummings).

9. Document in a thorough manner everything we have done to develop the data.

Finally, note that the above only takes account of the existing LO data and does not yet address the new data that we'll be getting in from the NLS and the DLC.