Subject:	Re: LEAP: Conversions
From:	Adrian S. Wisnicki [redacted]
To:	[redacted]
Cc:	[redacted]
Date:	Friday, June 20, 2014 3:36 PM

Hi James,

Thanks very much for this and sorry for the delay in following up. Let me start by apologizing for the length that this email will take. My delay in following up with you has had to do with the fact that I'm not sure how to handle one aspect of the conversion. I'm copying Frank Smutniak from UNL on this email (who's providing data management on LEAP), who'll be part of at least this process and may be able to provide useful advice on the questions I set out below. Frank, James is converting our legacy XML TEI P4 and P5 files to one common TEI standard.

I think it might be good for the three of us to talk through this via email so as to be clear on the best way to proceed. In and of itself, the issue I'm about to outline isn't that complex, I think, but I'm just not sure how to handle it. Here goes:

1) Currently, all our TEI P4 and P5 files are ready for further processing. I'm attaching them to this email so that you can see them. As you'll see they are distributed among various directories and subdirectories, and I'd like this directory structure kept. There are two sorts of files here: A) in a subdirectory of the P5 files, there is a spectral imaging directory and B) all the other files. What I write below only applies to #B.

2) The XML files are currently named either by a four digit Clendennan and Cunningham number (a Livingstone thing) or by a somewhat adhoc name (though internally standardized) that I've assigned. So, for instance, "0358.xml" or "Uncat to Russell 28 Nov 1865.xml". If you dig through the files (other than the spetral imaging files), you'll see what I mean.

3) Eventually, all the file names (except for the spectral imaging files) need to be converted to the LEAP file name standard. This is regular and takes the following base file name form:

liv_000001
liv_000002
liv_000003, etc.

"liv", the first segment, means Livingstone. The second six digit segment is the unique item number. So, for instance, an XML transcription of the second item would be liv_000002.xml. Each of our current XML files, therefore, has a unique LEAP file name that has been assigned to it (and I know what it is and can provide that data in a spreadsheet that matches the current name to the LEAP name). So for XML file names it's a matter of going from A to B. So far so good.

4) However, here's where it gets tricky. For TEI P4 files, the base file name must also replace the current <TEI.2> value inside of those files. So, making up an example, **0358.xml** might currently have <TEI.2 id="**LETT0358**">. 0358.xml will now become **liv_000459.xml** so the <TEI.2> must become <TEI.2 id="**liv_000459**">. For TEI P5 files, a similar change must be made: something like <TEI xmlns="http://www.tei-c.org/ns/1.0"

xml:id="**LETT0358**"> would need to go to <TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="**liv_000459**">

5) And we're not done yet. Each TEI P4 or P5 file, will contain a series of <pb/> elements with an @n attribute and a value. For instance, the current 0358.xml might have:

<pb n="1r"/>
<pb n="1v"/>
<pb n="2r"/>
<pb n="2v"/>, etc

In each file, to each successive <pb/> element, must be added @facs attribute that builds on the LEAP file name, but that incrementally increases with each <pb/>. So, to take a made-up TEI P5 example, we would have the following when done:

liv_000459.xml (LEAP file name)

<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="**liv_000459**"> (corresponding internal <TEI> element)

<pb n="1r" facs="**liv_000459_0001**"/>
<pb n="1v" facs="**liv_000459_0002**"/>
<pb n="2r" facs="**liv_000459_0003**"/>
<pb n="2v" facs="**liv_000459_0004**"/>, etc.

The @facs then, correspond to the given images of the item and will allow us to link image and text etc, while the use of the LEAP file name elsewhere will allow us to link files with metadata, etc.

Does this all make sense? My explanation is on the long side, but what's being discussed here really, I think, isn't that complex. We have a series of tasks to be done. Each given task would fall to one of you (or other). These tasks need to be done in a sensible sequence. Here, in summary, is a list of the tasks:

1) Convert all XML TEI P4 and P5 file names (except the spectral imaging XML files) to the corresponding LEAP file names. Keep existing directory structure intact. [Frank to do this?]
2) Change the appropriate values of the <TEI.2> and <TEI> elements to the relevant base LEAP file names. [Not sure who would do this]
3) Add incrementally increased @facs values based on the LEAP file name to the successive <pb/> elements in each file. [Not sure who would do this]
4) Convert all XML TEI P4 and P5 files to the LEAP standard. [This will be done by James].

So, what's the order in which we do these things and who will do what? Let me know your thoughts. I'm happy to arrange a telecon if you think it'd be easier to talk this through.

Thanks for reading this far!

Adrian