TEI and Images Correspondence

Use Case: Outsourcing Digitization and Text Encoding

Lead developer: David Jiao
E-text project manager: Michelle Dalmau
Projects in critical need of check: Indiana Magazine of History, Indiana Authors and Their Books
Script due: 11/3/2007
Documentation: DLP Etext QC Tool

When projects are outsourced to vendors for both digitization and text encoding, the vendors follow strict file naming and ID designation conventions according to our contractual agreement and specialized encoding guidelines.

We need to make sure the number of page breaks <pb> with id attributes (contain file name identifier) in the TEI document correspond to the same number and file name of facsimile TIFF images. This check needs to be run as part of the automatic quality control process conducted by the Digital Media Specialist when files are downloaded from vendor drop boxes.

We also need the ability to re-generate identification numbers should vendors:

- Exclude TIFF images (probably means id references in the TEI document are inconsistent)
- Mis-name TIFF images (incorrect ids referenced in TEI document)
- Add extra page breaks in the TEI document
- Mis-identify files in the encoding <pb id=""> (incorrect id referenced in TEI document)

The Indiana Magazine of History (IMH) project is in critical need of this check. We hope to develop a process that is also applicable to other e-text projects that fit under this use case scenario. However, we will first need to address the specific needs of the IMH.

- See: Page ID Issues with the TEI (compiled by Annette Richmond)

Approaches for version 1 of this tool discussed by David Jiao and Michelle Dalmau include:

- Comparing counts of page breaks in the TEI with corresponding facsimile images. Record count for each. If count is off, then human intervention is required.
    - We could do both count and identifier matching (id content in <pb> with filename of TIFFs)
- If re-numbering of sequential identification numbers is required, the person running the check will be prompted for a paramter (e.g., the first page sequence number). The tool will then re-generate new identification numbers starting from the first identification number (e.g., VAA4025-001-1-001).
    - (?) Support the ability to re-number sequnetial identifiers from the middle of a document (e.g., after page 76, n=76). Not sure if this is necessary or we just renumber from start to finish regardless where the problem occurs.

Use Case: Outsourcing|In-House Digitization and In-House Text Encoding

For this case, sequential identification numbers of the id attribute of the page break <pb id=""> tag should be generated automatically after facsimili TIFF page images have been created.