Workflow for TEI Encoding Projects

This page is needs to be updated. Certain links to specific project information are restricted to staff.

Visit the E-Text Readings and Resources page for more information on text encoding. It links to standards, guidelines, element references, and tools, as well as information on TEI Lite, the TEI Header, document analysis and Schematron.

Step One: Document Analysis

To get started

- Doc Analysis at a Glance
- For more resources, go to the Document Analysis section of E-Text Readings and Resources

Analyze

- Decide how many items you need to review to get a sense of the collection (e.g. 1 volume every 10 years)
- Look through the materials and make a list of all of the 'features' present
    - Example: Indiana Magazine of History Document Analysis (restricted access)
- Determine basic structure of documents and decisions that need to be made prior to compiling the tag list.
    - Example: TEI Encoding Issues for Indiana Magazine of History (restricted access)

Step Two: Conduct Needs Assessment

To get started

- User Needs Assessment Study conducted by the Institute of Museum and Library Services (pdf)

Conduct Assessment

- Understand how current form of documents are being used. If they are to be "born digital," investigate use of similar resources. Do this by conducting:
    - interviews with end-users and content manager
    - content analysis of reference queries

Step Three: Compile TEI Tag Set

To get started

- TEI P4 Element Reference
- For more resources, go to the General Guidelines and Element References section of E-Text Readings and Resources

Compile

- Compare text features with project requirements to determine structural, semantic and referencing needs
    - Structural
        - How closely should the text file replicate the original page image?
        - Will the text view be the default (or only) view, or is it an adjuct to the page image?
    - Semantic
        - What are the access points?
        - What semantic features need to be marked up? (e.g. Do you need access to personal names in the text?)
        - How granular should the markup be? (e.g. Is it necessary to separate out name parts?)
        - Which features require authority control? Which thesauri or controlled vocabularies will be used and referenced as part of the encoding?
        - Which features need to be cross-referenced in the text?
            - Example: Indiana Authors project (restricted access)

To get started

- Make sure the latest version of Oxygen is installed on your computer
    - Site license for IU (restricted access)
- Setup an XML Catalog
    - If you do not have a catalog set up, download the appropriate TEI DTDs to the directory where you will be working. A bundled set of TEI DTDs can be found on SourceForge.net, or use the DTD or Schema generated for the project. See the Tools section of E-Text Readings and Resources for ways to generate DTDs and Schema.

Sample encoding

- Scan and OCR representative pages from the text and perform sample markup
- Add to your tag set as necessary
- Keep track of any unresolved encoding issues
    - Example: TEI Encoding Issues for Indiana Magazine of History (restricted access)

Step Four: Create Encoding Guidelines

To get started

- **Example:** Indiana Authors Vendor Encoding Guidelines
- **Example:** IMH TEI Encoding Guidelines
- For more resources, go to the E-Text Readings and Resources page
- Begin work in parallel on the functional requirements for delivery, since they will impact the encoding specifications

Include in guidelines

- What's being encoded (e.g. 102 volumes of IN Magazine of History)
- What files are being created from what sources (e.g. OCRed and encoded text from scanned page image; Scanned page image and keyed and encoded text from original documents.)
- What standard is to be used (e.g. Text Encoding Initiative (TEI) guidelines, version P4)
- File naming scheme
- Document structure
- Character encoding
- Normalization, cross-referencing and authority control
- Structural and semantic features that need additional explanation
- Tag list
- Attachments
    - DTD or Schema
    - XML/TEI Sample

Step Five: Create Schematron Validator

This is only necessary at this point if you are encoding in-house and want to perform quality control as you encode. Otherwise, the rules can be generated once the encoding is finished and the files run as a batch. If you create the validator before the encoding, you may need to update it when unexpected encoding issues arise (because they will).

To get started

- See the Schematron page for more detailed instructions on creating and using Schematron documents, and necessary files
- For more information, see the Schematron section of E-Text Readings and Resources

Step Six: Encode

Outsourced

- Sending
    - Gather and inventory materials to be digitized. Include any special handling instructions and provenance if necessary.
    - Insure materials if appropriate and ship to vendor (or FTP if performing text encoding on materials scanned in-house)
- Receiving (to be repeated for each batch of files)
    - Vendor notifies via email when files have been uploaded to FTP server
    - The Digitial Media Specialist retrieves files and documents it (including any problems with the files) on the project page on the wiki
        - DMIC QC Workflow and Template
    - The Digital Media Specialist and the Digital Imaging Specialist perform automated and manual QC on the image files and document the results on the project page
        - Automated QC includes checking filenaming in the TEI files
    - The project staff performs automated (Schematron and DTD validation) and/or manual QC on XML files as appropriate and documents the results on the project page on the wiki
- Test batch
    - Receive test batch (scope to be identified in advance with the vendor)
    - Document any problems with the images or the encoding on the project page and share the results with the vendor. Determine whether or not another iteration of the test batch is necessary.
    - Repeat if necessary
- Meet with entire project team to compare results of test batch with contract
- Vendor uploads remaining files according to schedule specified in contract
- Vendor returns originals once all files have been received and all QC has been performed

In-house

- Gather and inventory materials (Digital Media Specialist oversees digitization of originals and OCR)
- Check OCR files if necessary for appropriate structure and acceptable level of error
- Hire and train encoders if necessary
- Encoder
    - Retrieves text files according to project workflow
    - Performs encoding according to project guidelines, documenting any unresolved issues on the designated wiki page
    - Performs DTD and Schematron validation (if appropriate) on the completed file, and makes any necessary changes
    - Uploads encoded file to the designated folder and marks the file as encoded
- Project staff
    - Performs additional quality control on completed files as necessary
        - DMIC staff runs auto QC to verify TEI file names are correct
    - Updates guidelines to reflect encoding issues and decisions made
    - Uploads files to server or submit to repository (Fedora)