UNIVERSITY of VIRGINIA LIBRARY

Early American Fiction

- **EAF Home**
- **About EAF**
- **Restricted Collection**
- **Public Collection**
  - **E-Text Center**
  - **Questions?**
  - **VIRGO**

Early American Fiction Project Workflow

Book Handling

Prior to scanning, selected volumes are pulled from the stacks, inspected, and relocated to the digital lab. The books remain in the lab until they have been digitized, have a TEI header form, and the jpeg derivatives have been checked.

Workflow Database

Each book is given a record in a FileMaker Pro database.

With the book in hand, the EAF staff records information into the FileMaker Pro database:

- bibliographical information
- EAF project number
- call number
- notes on form and condition
- digitization dates
- camera operators

Upon volume imaging completion, the FileMaker Pro record is filtered to a TEI header.

Parsing, tiff header integration and quality assurancewill be done by the Electronic Text Center staff. AACR-2 compliance and MARC record generation will be conducted by the UVa Special Collections Cataloging Department.

Digital Image Creation

File-naming convention: [xxx-001]

- For images, the suffix will always be .tif-- supplied by PhotoShop (.jpg after the conversion) and for texts, .xml -- supplied by our web forms.

- The first three numbers will be a project number assigned to the book, followed by a dash.

- In the event that a volume has more than 1000 pages the next four slots will be free for sequential digital image numbers. This means that the image number does not reflect the pagination scheme but it overcomes the need to deal with unnumbered pages, preliminary numbering, repeated numbers due to printer error, etc.

- The eighth character is to remain blank as a safeguard against a missed image that needs to be numbered after the fact.

Stay within 8.3 DOS limits for all files and directories. Do not use spaces in either file names or directory names. The 8.3 file limit is essential for ISO 9660 conformance, to accomodate DOS, Windows, and CD production (all of our CDs will adhere to ISO 9660).

See the EAF Digital Image Scanning Procedures for a detailed description of camera operation, software settings, imaging, batching, and database tracking.

Conversion to JPEG

Run the batch-processing scripts in PhotoShop to produce a large JPEG file. From these we will generate gif thumbnails and two other levels of jpeg files.*.*

From the large JPEG version:

- GIF: mogrify -format gif -interlace plane -geometry 5% *.jpg
- MEDIUM JPEG: mogrify -geometry 75% -quality 75% *.jpg
- SMALL JPEG: mogrify -geometry 50% -quality 75% *.jpg

The aim is to keep the jpegs a known and predictable percentage of the original, so that they maintain relative size differences (e.g. an image of a small book looks smaller than an image of a large one.)

For examples of dual-quality jpegs see the following:

- Creating Online Materials for Teaching United States History, University of Virginia, Charlottesville, Virginia, June 16-22, 1996.

- Books at Virginia: Rare Book School (RBS): Introduction to Electronic Texts. July 1996. Two sessions: Early American manuscripts and typescripts

- The Jefferson/Church letters in our Modern English collection.

Text Processing

JPG files are uploaded to vendor's FTP site for processing according to a "Data Conversion Design Document" (currently Revision 1.3). The goal of the vendor is to reproduce the source in every aspect, including capturing line breaks and page breaks at the exact location as in the source.

Every <divx> has a <head> in the C-H scheme as in TEI, but the head is numbered along with the <divx> -- a <div0> takes a <comhd0>, a <div1> takes a <comhd1>, etc. At present, we think we will use the n= attribute to record this information : <head n="comhd1">. This will be easy to change to <comhd1> for C-H purposes.

The <text> tag in TEI cannot take a <head> itself, but its C-H equivalent needs a <head> and an <attrib> field. One solution is to add a <div1 type="chad"> at the top of every <front> before the real <front> matter, and move it up before teh <front> for the C-H format. Its <head> -- <head n=comhd0> -- contains a <bibl> containing the full, inverted author name (<author>) and the volume short title (<title>), including the date of publication in parentheses.

We still need to decide the precise form of the tags in the <text> that correspond to the C-H <attribs> group: <attauth>, <attgend>, <attgenre>, <attdate>, and <attbal> for full author name, author sex, genre of work, date of publication, and *Bibliography of American Literature* number. A <ref type="attribs"> containing a <bibl> is possible as a container for this information, within the <div1 type="chad">.

The end result needs to be a parsed TEI document that can be automatically re-shaped in a couple of details to form a C-H document.

Vendor Guidelines for tagging

Contract out to bid.

Guide for image description : <figDesc>

Book illustrations and other figurative content will be described as to its content, for searching purposes, using the

TEI <u><figure></u> tag.

Procedures for parsing, indexing, and testing completed texts when returned from the keyboarders

Will follow usual ETC practices. In particular, we will be checking for unintentinally minimized tags during parsing. The TEI.DTD allows minimization, but we do not want it. to guard against this, run the parser as:

```
nsgmls -s -w unclosed -w min-tag FILENAME
```

Electronic Text Center
University of Virginia Library • PO Box 400148
Charlottesville VA 22904-4148
phone: 434.924.3230 • fax: 434.924.1431

Maintained by: Etext Center
Last Modified: