



Historic Pittsburgh Full-Text: Proofing Production Notes and SGML Headers

Revision: 3.0

Last Update: 2002-12-18

Table of Contents:

1. General Notes

- 1.1 Background
- 1.2 Location of Files
- 1.3 Special Instructions for Multipart Works and Volumes of Serials

2. Proofing the Production Notes

3. Proofing the SGML Headers

- 3.1 The "Quick" Way
- 3.2 The "Detailed" Way

4. Special Instructions for Headers Created from Minimal MARC Records

5. Creating an SGML Header from Scratch

- 5.1 Open Template
- 5.2 Create the new file
- 5.3 Number of Images, <extent>
- 5.4 Keywords; Personal name; Subject Heading; Geographic Term
- 5.5 Empty fields from template
- 5.6 Shipment Date & Number
- 5.7 Save and Validate

1. General Notes

1.1 Background

After a shipment has been assembled and the spreadsheets created, SGML headers must be generated for each book. This step takes place *before* the shipment is sent to the scanning vendor.

Headers are generated either automatically (with a script), or by hand, although hand-creating headers is almost entirely deprecated at this point, as we insist that some kind of bibliographic data in electronic format accompany full-text items when we acquire them.

The process of automatically generating the headers is discussed elsewhere, in the documentation about preparing a full-text shipment. This document concentrates on how to review headers that have already been created. It also includes some instructions for creating headers from scratch that are now should be of mainly historical interest.

1.2 Location of Files

If you've used a script to auto-generate the production notes and SGML headers, they should be located on the networked drive for the text collection, under a subfolder called `headergeneration`, and then within a folder named for your shipment.

Examples:

```
P:\Histpitt\headergeneration\shipment13 (For Historic Pittsburgh)
N:\headergeneration\shipment3 (For 19th Century Schoolbooks)
```

Within this folder, each source document will have the following 2 files:

- Header: `voyagerid.toc`
- Production Note: `voyagerid.rtf`

If you need to generate a SGML header from scratch, because there is no MARC record for the item (rare), the header template with which to create headers from scratch is located here:

```
P:\bookheaders\headertemplate.sgm
```

1.3 Special Instructions for Multipart Works And Volumes of Serials

Multipart works and volumes of serials are not like most of the monographs found in Historic Pittsburgh. Here are tips to help you handle such cases:

For all record types, check volume number or year of coverage that appears in two places: `<filedesc>` (describes digital file) AND `<sourcedesc>` (describes

book). Without it, the user will not know which particular piece of a multipart work we have digitized.

For serials, the automatically supplied volume number may not be appropriate since it represents our own calculation of the probable volume number gathered from the cataloging and holdings records.

Because many serials do not have a volume number or an author main entry (1xx; <author>), it is not helpful to put the volume number ahead of the title as we do for multi-part works, which often do have an author main entry. In addition, we can sort titles alphabetically if the title precedes the volume number.

2. Proofing the Production Notes

For each item in the shipment, a production note is derived from the MARC record. These are saved in Rich Text Format (file extension .rtf). The title of the work, a volume number for multi-part works, and an author (if there is one) that are prominently featured on the title page are automatically retrieved and incorporated.

Edit the production note to more closely match the title page data as follows:

- Keep the full title of the book as it appears in the MARC record.
- Make sure the long titles are properly centered on the page. If the title stretches across the page, it may not fit within the borders of the reprinted production note. It is best that we center the title here so that NMT will not have to charge for extra time needed to adjust the page.
- Capitalize title words according to standard practice - that is, capitalize the first letter of the first word and the first letter of each word thereafter, except prepositions, articles, and coordinate conjunctions.
- If necessary, change the author's name order to read: first, middle, and last name. If there is only a corporate author, add it to the production note.
- Volume or serial information should also be included.
- Do not provide any author information for books with multiple authors.
- When in doubt about what to add to or edit from a production note, check out the previous shipment of reprinted books on the DRL shelves.

3. Proofing the SGML Headers

The "Quick" Way of proofing described below will probably be sufficient for checking the SGML headers, but the Digital Production Librarian may decide that the "Detailed" Way, also described below, is sometimes necessary.

3.1 The "Quick" Way

There is a batch script that will ensure each header is SGML-valid, and will check in a rough way for any ampersands that have not been turned into the SGML entity `&`. If the batch script can be run with no errors, that is usually a sufficient degree of proofing, although a detailed, field-by-field method of proofing is covered in the next section.

3.1.1 Running the Batch Script to Proof Headers

The batch file will use the program "nsgmls" to validate the headers. It will also do a simple check for un-escaped ampersands (using "grep") that might cause problems.

HOW TO USE:

1. The batch file is called `dos_batch_validate.bat`, and should be located in the same place as the `headergen.pl` script for the text collection. **Copy** the batch script file into your headergeneration shipment directory
2. Open a dos command shell (*Start Menu -> Run -> type "cmd" in box*).
3. Change directories to your current shipment, e.g.

```
N: [hit "return"]  
cd headergeneration\shipment3
```

4. Run the batch file by just writing its file name at the prompt, i.e.

```
dos_batch_validate.bat
```

5. The script should run, and any problems with either validation or un-escaped ampersands will be printed to the screen. These problems can be fixed by opening the SGML header in a text file, correcting the problem, saving the file, and re-running the batch script.

3.2 The "Detailed" Way

Double check every field, but be especially careful about the following fields:

<author encodinganalog="100">

For works with no 1xx field, substitute a **blank line** for the `<author></author>` line at the top of the `<filedesc>`

We must include this step because this information is a requirement of our contract with the scanning vendor, NMT (Northern Micrographics). NMT incorporates the header into data and returns it to the DRL as scandata.txt.

<title encodinganalog="245" nfchar="###">

Look for typographic errors in the transcription of the title.

Non-filing characters; nfchar="###"

Non-filing characters indicate the number of characters to be ignored in filing if the title begins with an article. Enter the number of characters in the article, plus spaces, punctuation, and diacritics that precede the first significant word (See AACR2 for explanation).

Example #1:

```
<title encodinganalog="245"
nfchars="0">Occupational changes by the
Civil Works Administration.> </title>
```

Example #2:

```
<title encodinganalog="245" nfchars="6">--
the world we live in. </title>
```

<title encodinganalog="246">

246s (Varying form of title) do not have non-filing characters. Some 740 Added Entries (Uncontrolled Related/Analytical title) may need to be changed to 246s.

Check OCLC's *Bibliographic Formats and Standards Manual* for cases where obsolete cataloging practices need to be updated. If you create this field, be sure to place it in the <sourceDesc> area as well.

Number of Images

Insert the number of images into the <extent> field, if it has not already been inserted. This information cannot be obtained until the Digital Production Librarian has proofed all spreadsheets for the shipment.

Example:

```
<extent>66 Group IV Compressed 600 dpi TIFF  
Images</extent>
```

Here, **66** represents the number of images of a scanned item.

The number of images is NOT the same as pagination. Pagination of the original (source) document will be recorded in the

```
<extent encodinganalog="300"> field in <sourcedesc>.
```

<sourcedesc>

In this field, the titlestatement should be the same as the titlestatement in <filedesc> with one exception. If there is no 1xx field, delete the full tag:

```
<author encodinganalog="100"></author>
```

Do not leave the line blank as instructed in step 1 above.

Voyagerid number (unique identifier)

Verify the correct voyagerid number. This is how the field should appear:

```
<IDNO encodinganalog="voyagerid">#####</IDNO>
```

Everything else that is part of the publicationstmt, should be correct and needn't be altered.

Imprint

Check 260 field for odd information and spaces.

Example:

```
<imprint encodinganalog="260">[Pittsburgh] :  
Pittsburgh Personnel Association, 1934.  
</imprint>
```

Extent

Check 300 field for odd information and spaces. Because the description is often obsolete (i.e., not according to AACR2r standards), you may need to add or delete punctuation.

Well-formed Example:

```
<extent encodinganalog="300">vi, 58p. : ill.  
; 23 cm.</extent>
```

<notesstmt><note encodinganalog="500">

Delete any-copy specific notes (usually 590s) that are not relevant to the electronic edition.

<keywords encodinganalog="6xx" source="lcsch"> or <keywords encodinganalog="7xx" >

Delete unnecessary subfields associated with authority controlled name headings (obsolete subfields that had some meaning before the adoption of *AACR2r*).

Delete 690s or other non-LCSH data disguised as subject descriptors.

4. Special Instructions for Headers Created from Minimal MARC Records

Double check everything, but be especially careful about:

<extent encodinganalog="300">

Note that many of these records lack a size designation, that is, no \$c (15p. : ill. ; 20 cm.) Be sure to supply this information by measuring the book. This information is important for reconstructing books faithfully to the original and to distinguish editions.

Non-filing characters; nfchar="###"

<keywords encodinganalog="650" source="lcsch">

Delete unnecessary subfields associated with authority controlled name headings (obsolete subfields that had some meaning before the adoption of *AACR2*).

Delete 690s or other non-LCSH data disguised as subject descriptor fields.

5. Creating an SGML Header from Scratch

Content for the record may be copied from either a *Voyager* record or an *OCLC* record that shows all the MARC fields (do not use an OPAC view or *WorldCat*)

5.1 Open Template

Open the template file in a text editor. Again, *Emacs* is recommended, and again the template file is located here:

```
P:\bookheaders\headertemplate.sgm
```

5.2 Create the new file

So that you don't alter the template, you must first rename and save the file with the NOTISID number. Then, enter information for ALL fields. Do not forget to enter information for the following fields:

5.3 Number of Images, <extent>

Insert the number of images into the <extent> field, if it has not already been inserted. This information can not be obtained until the Digital Production Librarian has proofed all spreadsheets for the shipment.

Example:

```
<extent>66 Group IV Compressed 600 dpi TIFF  
Images</extent>
```

The number of images is NOT the same as pagination. Pagination will be recorded in the <extent encodinganalog="300"> field.

5.4 Keywords (6xx); 600 Personal name; 650 Subject Heading; 651 Geographic Term

Each term must be inserted between <term></term> tags

Example:

```
<keywords encodinganalog="650" source="lcsch">  
<term>Occupations</term><term>United  
States.</term></keywords>
```

5.5 Empty fields from template

Delete both tags and the information in between when:

- There are no notes in the source document

```
<notesstmt><note  
encodinganalog="500">xxx</note></notesstmt>
```

- You have fewer subject headings/keywords than the template

```
<keywords  
encodinganalog="650" source="lcs" ><term>xxx</term>
```

5.6 Shipment Date & Number

Insert shipment date and number.

Example:

```
<date>March 2000</date> Shipment  
#9.</projectdesc></encodingdesc> </header>
```

5.7 Save and Validate

When the record is complete, save it and validate as described above.