# CHAPBOOK PROJECT - RARE BOOKS AND SPECIAL COLLECTIONS

## TEI REFERENCE GUIDE FOR THE PROJECT

#### TEI CODING

The workflow in this document is based upon **Best Practices for TEI in Libraries** (http://purl.oclc.org/NET/teiinlibraries) and **TEI P5:Guidelines for Electronic Text Encoding and Interchange** by the TEI Consortium. edited by Lou Burnard and Syd Bauman 1.9.1. Last updated on March 5th 2011.

It is necessary to select a level of encoding that is appropriate for the project and the resources.

**Best Practices for TEI in Libraries** defines five levels of encoding, from Level 1 being the minimal and Level 5 being the maximum with significant subject knowledge needed by the encoders.

#### LEVEL 4

Level 4 is most suitable for projects with the following characteristics, which match the goals of the Library's Chapbook Digitization Project.:

- sophisticated search and retrieval capabilities are desired
- the texts will be used for textual analysis
- extensibility is desired; that is, one desires to keep open the option for a higher level of encoding to be added by the scholarly community at a later date<sup>1</sup>

Level 4 texts contain elements and attributes that <u>describe content</u>, not just appearance, of the text.

Texts encoded at Level 4 are able to stand alone without page images in order for them to be read by students, scholars, and general readers.

<sup>&</sup>lt;sup>1</sup>Best Practices for TEI in Libraries, page 37.

#### ELEMENTS RECOMMENDED FOR LEVEL 4

```
<front>
<titlePage>
<back>
<div>
<head> (headings)
<body>
(paragraph breaks)
< lg > and < l > (lines)
<figure> (illustrations, woodcuts)
<note> (handwritten notes)
<hi> (changes in typeface).
<l> lines of verse
list> (lists)
<lb/>(line breaks)
<pb>or <facsimile> (link to the image file)
<sic>,<corr>, or <choice> (errors or typos)
<unclear> (illegible)
```

#### THE STRUCTURE OF THE CHAPBOOK TEI RECORD

# **TEI Structural Markup**

Every TEI file begins with a technical declaration that states that it is an XML-encoded document, and that it uses the Unicode character set according to its "UTF-8" encoding.

```
<?xml version="1.0" encoding="UTF-8"?>
```

Next, it specifies that the document uses a subset of TEI P5 elements that are declared in a document type definition called "mcgillchapbook.dtd" which combines the TEI elements and rules with a set of special character entities, and allows for any future modifications:

<?xml-model href="file:/U:\RBD\Common\Chapbooks\Chapbook Project
current schema/oddRBSC26.rnc" type="application/relax-ng-compact-syntax"?>

After these technical declarations, the TEI file proper begins.

#### <TEI>

Contains a single TEI-conformant document, comprising a TEI header and a text

The TEI file's contents are enclosed in an outer wrapper: <TEI> </TEI>. The opening tag also includes some name space attributes:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"> </TEI>
```

The TEI file is comprised of three main sections inside this <TEI> outer wrapper – the <teiHeader>, the <facsimile>, and the <text>.

#### <teiHeader>

Supplies the descriptive and declarative information making up an electronic title page prefixed to every TEI-conformant text.

Contains metadata about the specific letter the TEI file contains, and the project to which it belongs. The <teiHeader> is comprised of four main sections:

<fileDesc> Contains a full bibliographic description of an electronic file. </fileDesc>
<encodingDesc> Documents the relationship between an electronic text and the source or
sources from which it was derived.</encodingDesc>

<revisionDesc> Summarizes the revision history for a file. </revisionDesc>

For the McGill Library Chapbook Project, a final header will be produced by an automated procedure using the MARC record from the Library catalogue. The encoders will use a sample header copied from the chapbooktemplatecurrentversion.xml file.

The encoders and reviewers will add their name, date and notes into the header file in the <change> element.

### <facsimile>

Contains a representation of some written source in the form of a set of images rather than as transcribed or encoded text.

The facsimile block contains a <graphic/> element for each page image, which specifies whether the page is a verso or recto, its page number, and the unique image ID for each page image.

The unique McGill image IDs are LC call number, followed by the Aleph system number and the image sequence number (001, 002, 003 etc.).

This image below shows the image files of a chapbook digitized by LTS staff.

```
PN970_D4_D4_no44_1857_003483191-001.tif

PN970_D4_D4_no44_1857_003483191-002.tif

PN970_D4_D4_no44_1857_003483191-003.tif

PN970_D4_D4_no44_1857_003483191-004.tif

PN970_D4_D4_no44_1857_003483191-005.tif
```

So, a five page chapbook would look like this:

### SUMMARY TEI STRUCTURE

# Summary: The TEI structural "skeleton" for a chapbook.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="file:/U:\RBD\Chapbooks\Chapbook Project current schema/oddRBSC26.rnc"
type="application/relax-ng-compact-syntax"?>
<teiHeader>
        <fileDesc></fileDesc>
        <encodingDesc></encodingDesc>
        profileDesc>
        <revisionDesc></revisionDesc>
</teiHeader>
<facsimile>
    <graphic rend="recto" n="1r" url="file image name goes here.tif"/>
</facsimile>
<text>
        <front>
            <titlePage
            </titlePage>
        </front>
        <body>
                [lines of text are entered here]</lb>
        </body>
         <back>
                [back matter on the text goes here, if any]
</text>
```

### TEI HEADER

Copy and paste the xml lines contained in the header from the TEI chapbook template, from the beginning to <facsimile>. The template file is called - **chapbooktemplatecurrentversion.xml**. The TEI header will be updated with cataloguing information from the MARC record as a later step in the project.

</TEI>

#### TEXT

The OCR process and review by LTS staff should have created an html file that accurately represents the characters in the work that you will be reading. In our review of ten sample works, RBSC and LTS have decided the following:

- Only leave misspellings if seen in the original.
- Verify the letter case.
- Extra paragraph breaks added by the Abbyy software will be removed.
- Extra spaces will not be removed unless they appear in the middle of a word.
- Quotes will be left as detected, either curly or straight..
- No need to replace 'abbyy' dash glyphs with hyphens as the TEI encoders will change to the appropriate entity..
- Latin glyphs (e.g. long form s) will be kept.

Tightly bound volumes where the text is missing will be left until the end of the project and highlighted in red on the spreadsheet.

It is also important to let LTS know when a significant difference if found between the reading of the work and the html file, so that the PDF file that LTS has created can be changed. Please have Sharon email Greg about any of these discrepancies that turn up.

# SUBJECT CLASSIFICATIONS IN THE <TEXT.

Select one of the following terms to categorize the chapbook work. Enter this term in the <text> type= from the dropdown choices.

Religious

Household

HistPolitBiograph

GeographHistory

TravelAdventure

OddStrangeEvents

ProseFiction

RomancesFairyFolk

NurservRhymes

Dramatic

MetricalTales - note: this includes verse

SongBooks

**JestBooks** 

Occult

Prophecies

Crimes

Instruction Miscellaneous

The opening <text> tag must have a classification.

### <DIV>

We are using unnumbered <div> elements for any of the following parts of a chapbook work.

cover
frontispiece
preface
alphabet
imprint
advertisement
epistle
colophon
chapter
section
bookplate
epilogue
epigraph

Code the <div> type as follows:

<div type="preface">

## PAGE NUMBERS

We record the idealized page number on the n= of pb>:

# <pb n="401">

The idealized number is important, because the actual number in the <fw> may be in error, and so could not be relied upon for referencing.

Remember that each <pb> starts a page. Thus the n= attribute with the value "6" goes on the <pb> before the content of the page that has the number "6" printed on it, even if the number is printed at the bottom.

Coding a word break over two pages - in a page break

It is be–
<pb n="4" type="inWord"/>
cause we could better spare any other.

### FORM WORK - HEADERS AND FOOTERS

We record the page number as printed on the page in <fw>- the form work element.

<fw type="pagenumber">401</fw>

Other types of <fw> characters are:

catchword signature header footer altpagenumber enumeration

The location in the header or footer is specified with a rend=.

<fw type="pagenumber" rend="align(right)">401</fw>

Place the <fw> statement in the text where the characters appear in the text.

#### **ITALICS & BOLD**

Italics and bold can be coded in two ways.

If an entire line is in italics, code as a rend= in the element, for example

<pubPlace rend="italic">London:

If one work within the paragraph is in italics, use the <hi> element with a rend=, for

example:

Oh! what a pity it is, that such a pleasant<lb/>household should be spoiled by a discontented face amidst the bright s<hi rend="italic">smiling</hi>) ones!

<hi rend="bold">&ldquo;It is never too late to mend,&rdquo;</hi>

#### **FRONT**

The front matter is any of the following parts of the work - the cover, frontispiece, epistle, titlepage, imprint, preface. Check the alignment. If it is the same for the entire front matter enter **rend=** in the **<front>** tag.

Within the <front> tag, create a <div> for each of the types of front matter present in the work before you.

<div type="cover" rend="yellow coloured pictorial wrappers">

Add the page numbers after the <front> tag.

```
<front rend=align(centre)>
<pb n="1"/>
```

### **COVER**

If the work has a cover, add a <div>. Then code the text that is on the cover. Often it might repeat the text that you will find printed on the title page.

Use <bibl>, <title>, <publisher>, <series>, <biblScope> to code bibliographic text:

```
<front>
```

The level = "s" is the series title and level = "m" is the title of the work (monograph) Note that the series number should be coded with the type of number.

#### PRICE OF THE WORK

Sometimes the original price of the work is on the cover or imprint. The value should be coded as a <measure type="price">:

PRICE <measure type="price"> TWO CENTS</measure>

#### FRONTISPIECE

The frontispiece is an illustration facing the title-page. If the frontispiece is a foldout use rend="foldout"

A frontispiece is coded as a type of <div>.

```
<div type="frontispiece">
<figure>
<ic:iconClass>
<name/>
<desc/>
</ic:iconClass>
</figure>
<div>
```

### **FIGURES**

Add the <figure> group of tags for every image (woodcut) in the work. Be sure to code for the positioning of the image and whether it has colour.

In this example, the woodcut is positioned sideways. Reverse means the image is facing away from the stitching.

```
<figure rend="vertical reverse">
<ic:iconClass>
<name></name>
<desc></desc>
</ic:iconClass>
</figure>
If a figure has a title caption, code it in the <head> tag.
<figure rend="coloured">
<ic:iconClass>
<name></name>
<desc></desc>
</ic:iconClass>
<head>CRUSOE SAVING FRIDAY'S LIFE.</head>
</figure>
<figure>
<ic:iconClass>
<name></name>
<desc></desc>
</ic:iconClass>
<head>The Fox is <hi rend="printed sideways">n </hi>oted for his
cunning.</head>
<figure>
```

Note: <name> and <desc> are being coded by the Reviewer and will be a two part ICON class code from <a href="http://www.iconclass.nl/iconclass-2100-browser">http://www.iconclass.nl/iconclass-2100-browser</a>.

An example of the ICONCLASS code entered:

```
<figure rend="coloured">
<ic:iconClass>
<name>46A182</name>
<desc>master ~ slave</desc>
</ic:iconClass>
</figure>
```

#### LINE BREAKS

## From the TEI P5 guidelines -

By convention, <u>lb</u> elements should appear at the point in the text where a new line starts.

This element is intended to be used for marking **actual line breaks** on a manuscript or printed page, at the point where they occur; it should not be used to tag structural units such as lines of verse (for which the <u>l</u> element is available) except in circumstances where structural units cannot otherwise be marked.

```
<titlePart>
<lb/>L'auteur susdict supplie les Lecteurs
<lb/>benevoles, soy reserver à rire aus me livre.
</titlePart>
```

#### ENTITIES

The TEI header contains entity definitions for characters, rules, etc that are not represented by the ASCII character set

Entities are defined in the header and then used to code for the appearance of the character in the text. All of the definitions are in the header, even if they are not used in the XML file.

If you encounter a character that is not defined, please see Sharon. It will need to be defined and added to the TEI header in the chapbook template.

Entities are entered with a preceding & and the defined values will appear in the dropdown in the editor.

For example:

<head rend="align(centre)">&orn;<lb/>LADY'S PALFREY.</head>

```
CREWE HALL<lb/>LIBRARY.
&rule;
Juvenile Books.
```

These are the entities defined in the TEI header as of February 2013.

```
<!ENTITY mdash "&#x2014;">
<!ENTITY ndash "&#x2013;">
<!ENTITY slong "&#x17f;"> - a long s
<!ENTITY pound "&#xa3;">
<!ENTITY owithdm "&#555;">
<!ENTITY ae "&#xe6;">
<!ENTITY AE "&#xc6;">
<!ENTITY oe "&#x153;">
<!ENTITY OE "&#x276;">
<!ENTITY mdot "&#xb7;"> (middle dot)
<!ENTITY Isquo "&#8216;"> <!-- left single quotation mark,
U+2018 ISOnum -->
<!ENTITY rsquo "&#8217;"> <!-- right single quotation mark,
U+2019 ISOnum -->
<!ENTITY sbquo "&#8218;"><!-- single low-9 quotation mark,
U+201A NEW -->
<!ENTITY ldquo "amp;""> <!-- left double quotation mark,
U+201C ISOnum -->
<!ENTITY rdquo "&#8221;"><!-- right double quotation mark,
U+201D ISOnum -->
<!ENTITY orn "ornament">
<!ENTITY rule "single rule">
<!ENTITY drule "double rule">
<!ENTITY thrule "thick rule">
<!ENTITY swell "swell rule">
<!ENTITY crule "curved rule">
<!ENTITY vrule "vertical rule">
```

#### **BANNERS**

A banner is a type of figure.

```
<figure type="banner"/>

rend="within banner align(centre)">DEAN & amp;. SON'S PENNY
BOOKS.
</figure>
```

#### TITLE PAGE

```
The title page has several possible types (listed below)
<valItem ident="titleblock"/>
<valItem ident="halftitle"/>
<valItem ident="colophon"/>
An example:
<titlePage >
<docTitle>
<titlePart type="main">BREAD,</titlePart>
<titlePart type="sub">THE<lb/>STAFF OF LIFE.</titlePart>
</docTitle>
The imprint is the place, publisher, date, etc describing the work that follows the title.
<docImprint>
<pubPlace>NEWARK, (N.J.)
<publisher rend="align(centre)">BENJAMIN OLDS.&rule;/publisher>
<hi rend="align(centre)">&orn;</hi>
</docImprint>
<titlePage type="page"><docEdition>No.<num
type="cardinal">139.</num></docEdition>
        <docTitle>
          <titlePart type="main">CONTENTMENT<lb/>IN<lb/>HUMBLE
LIFE.</titlePart>
          <titlePart type="sub">A<lb/>MEMOIR OF THOMAS HOGG.</titlePart>
        </docTitle>
<docImprint>
<pubPlace>LONDON:
<publisher>Printed by T. Richardson, 245, High Hold<lb/>Ib/>JOHN BYSH, 8, CLOTH
FAIR,
WF<lb/>&rule;<lb/></publisher>PRICE<measure type="price">Sixpence</measure>
</docImprint>
```

It is not possible to code rules as separate elements in the <titlePage> tag. Add the rule within one of the <titlePage> tags.

Coding for an editor or author

#### **EPIGRAPH**

If the title contains a short verse, it should be coded as an <epigraph>.

```
<epigraph>
<lg>
<l>Poor children! how they must have cried,</l>
<l>When in that horrid wood</l>
<l>How keen their sufferings too, have been,</l>
<l>To die for want of food!</l>
</lg>
</epigraph>
```

#### BODY

Within the <body> tag, create a <div> for each of the chapters or sections in the work. Most chapbooks will not have <div>'s created as they are simple works with the text all in one block.

If the work has distinct sections, sometimes with a heading, then unnumbered <div>'s should be created for each section and use <div type="section">.

Use <head> type ="work" of the title of the work is repeated in the body.

Other headings do not need to be specified with a type.

```
<div type="section">
<head type="work"
rend="align(centre)">THE<lb/>SUN&ndash;FLOWER.<lb/>&rule;</head>
<head rend="align(centre)">ANN AND THE CAT.</head>
```

This example is for text that has chapters and chapter headings and in this case only, one would code a main and sub title of heading for the chapter if it existed.

### **DROPPED CAPITAL - LETTRINE**

A common style of typesetting is to start a paragraph (in books, usually the first of a chapter) with its first letter set large enough to span several lines.

This style is known as "dropped capitals", or (in French) «lettrines»

Use the <c> - character element to enclose the initial and rend="lettrine">

# <c rend="lettrine">O</c>nce upon a time

#### **BACK**

The back matter is any of the following parts of the work - advertisement, colophon, cover, epistle. It is not necessarily an entire page, for

Each of these types will appear as a back type and the choices can be selected from the dropdown menu.

For example:

```
<br/>
```

This is back matter - a list of publishers' other books

```
<back>
```

```
<div type="advertisement">
  <head rend="align(centre)">HISTORY BOOKS,<lb/> PUBLISHED
```

```
BY S. & WHOLESALE AND RETAIL.<br/>
KEYS, WHOLESALE AND RETAIL.<br/>
PENNY EACH.<br/>
/head>
<item>The History of Fair Rosamond</item>
<item>The Life and Adventures of Richard Turpin</item>
<item>The History of Tom Thumb</item>
<item>The surprising Adventures of Puss in Boots</item>
<item>Rhymes for the Nursery</item>

</back>
```

## **NOTES**

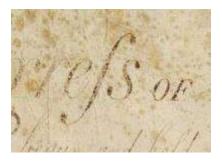
If there are handwritten notes on any pages, we are coding them in a <note> element. We are not recording illegible scribbles, damaged pages, and Library accession numbers and price information.

<note type="handwritten" rend="align(right)">Love from Mom on your 8th birthday!</note>

If any text is unclear, and you cannot make out a portion of the note, use the **<unclear>**tag to surround the visible text.

<note type="handwritten">Mrs. T<unclear reason="illegible">.</unclear></note><lb/>
<unclear reason="missing"/>

### LONG S



(This is an image of a page with text that includes a long s.)

The liquor is in the &slong; afe.

# **QUOTATION MARKS**

Coding for left and right quotations >

'WHAT say you, Jessy my dear, to
spending<lb/>a few weeks with your Aunt Weldon?&rsquo; said<lb/>Mrs. Crawford, as
she entered the room where<lb/>her little daughter was sitting learning her<lb/>morning
lesson.

# HYPHEN, NDASH, MDASH

From http://www.punctuationmatters.com/the-hyphen-dash-n-dash-and-m-dash/ (January 29, 2013)

Some definitions:

# hyphen

- Indicates breaks within words that wrap at the end of a line.
- Connects compounded words like "mass-produced".
- Connects grouped numbers, like a phone number 555-860-5086 (but not used for a range of numbers, like a date range).

### ndash

- Joins numbers in a range, such as "1993–99" or "1200–1400 B.C." or "pages 32–37" or open-ended ranges, like "1934—".
- Joins words that describe a range, like "July-October 2010".

### mdash

- Works better than commas to set-apart a unique idea from the main clause of a sentence:
- "Sometimes writing for money—rather than for art or pleasure—is really quite enjoyable."
- Separates an inserted thought or clause from the main clause, such as:

"Shows when dialogue has been interrupted:

"I reached in and pulled the spray can out of my pants—" "In front of the police?"

## Examples:

Hyphens are not replaced with an entity.

In this example there is an mdash in the text.

The Seasons. – Winter, the longest season.

The Seasons — Winter, the longest season.

In this example there is an ndash in the text.

The mass-produced chapbooks

The mass – produced chapbooks

In this example there is an ndash breaking the words apart between one line and the next.

I was hoping to be there because it was cold.

I was hoping to be there **be–<lb type="inWord"/>cause** it was cold.

### **INDENTATION**

Indentation should be represented using the rend attribute on the element affected, never with spaces or tabs typed into the text. It should be encoded at the highest relevant level of granularity: on the entire paragraph rather than its individual lines.

In the rendition ladder method of encoding rendition recommended by Brown University there are three distinct keywords for indentation: **indent, first-indent** and **right-indent**.

The **indent** keyword indicates that the entire element designated is shifted to the right by the specified amount

The **first-indent** keyword indicates that the first line of the element is shifted by the expressed amount relative to the other lines in the element. A positive value indicates a shift to the right (i.e. indentation), and a negative value indicates a shift to the left (i.e. an overhanging line). For example:

THERE are a great many little boys<lb/>and girls, who eat
bread every day, but<lb/>they do not know what it is made from,<lb/>or how it is made.
Perhaps they will<lb/>like to read this little book, which will<lb/>tell them about it?
Bread is made from<lb/>prain.

The **first-indent** keyword should not be used for elements which only include a single line, since it is meaningless in this context.

The **right-indent** keyword indicates that the entire element is shifted to the left (away from the right margin) by the specified amount. Strictly speaking this is equivalent to a negative value of indent; it should be used only for elements that are right-justified.

A typical hanging indent:

### RENDITION

**rend=** is an attribute used to describe the layout of the text on the page and some characteristics of the text. We are not coding for types of fonts.

It is not possible to have the XML editor prompt for specific **rend=** syntax and choice. So we will need to enter the information in the following format and try to check the order of the descriptive words as we work, trying not to make spelling typos.

Here are the examples of what we have used to date:

```
<div rend="yellow coloured pictorial wrappers" type="cover">
<div rend="grey coloured pictorial wrappers" type="cover">
```

```
THE Alphabet is arranged in lessons
```

### **CORRECTIONS**

#### **VERSE**

This is how to code verse using the <lg> and <l> tags.

In this example the lines are double spaced and a line break is used to indicate this.

In this example the work states a speaker or voice for the speech, so the all the lines spoken by the speaker are contained in the <sp> tag, and the speaker is identified in the <speaker> tag.

```
<sp>
<speaker rend="align(centre)">MOTHER.</speaker>
<lg>
<l>"I'Twas God, my child, who made</l>
<lr>
<l rend="indent(8)">them all,</l>
<l rend="indent(2)">By his Almighty hand;</l>
<l>He holds them that they do not fall,</l>
<lr>
<l rend="indent(2)">And bids them move or stand :</l>
<l>That glorious God who lives afar,</l>
<l>In heaven beyond the highest star.</l>
</lg>
</sp>
```

Verse multiple poems - be sure to code for poem title in the head.

```
<l rend="indent(6)">Beguiling time with friendly chat;</l>
<l>They touch'd upon the price of meat,</l>
<l rend="indent(2)">So high, a Weaver scarce could eat.</l>
</lg>
</le>
<head rend="align(centre)">I.</head>
<lg rend="align(left)">
        <l>AS at their work two Weavers sat,</l>
      <l rend="indent(6)">Beguiling time with friendly chat;</l>
      <l>They touch'd upon the price of meat,</l>
      <l rend="indent(2)">So high, a Weaver scarce could eat.</l>
</le>
```

### COLUMNS

An example of how to code for column breaks using the <cb> tag.

```
<cb n="1"/>
Well, a Pit to hold Spring-Water.
To Welter, to wallow, or lie groveling.
<!-- remainder of column -->

Wey, the greatest Measure, containing five Chaldron.
Whale, the greatest of Sea-Fishes.
```