Transcription and Annotation Guidelines for the Gatekeepers of Knowledge Project

Ville Marttila

Version 0.5

06.03.2012

Chapter 1

Metadata

This chapter will describe the kinds of metadata included in the *Gatekeepers of Knowledge* text files and the format in which it is presented in the form of a TEI header. For the moment it only covers the bare necessities required for the initial library work, which means those parts of the header that need to be either checked or filled in.

1.1 The file description

The file description contains all of the metadata having to do with this particular digital text and its sources, including the original early printed text. Most of the contents in this part of the header have been automatically filled in from the metadata table collected from ESTC. The following components should be checked and corrected or measured and filled in, as appropriate:

titleStmt/respStmt/persName/ptr

Check that this pointer points to the editor responsible for the annotation of the title page (i.e. you).

sourceDesc/msDesc/msIdentifier

Check that the library is the correct one and add the shelfmark to the <idno> element.

physDesc/objectDesc/supportDesc/extent/measure

Measure the height of 20 lines of running text (in full millimeters) from somewhere in the text to provide a measurement of the overall line density (i.e. the character height/type height proportion). The measurement should be taken from the bottom of the first line (does not need to be the first line on the page) to the bottom of the 21st line, so that it covers the total height (interlineal space included) of 20 lines of type.

physDesc/objectDesc/supportDesc/extent/dimensions

Measure and enter the height and width of both the full page (if the title page is damaged, measure an undamaged page somewhere else in the book) *and* the 'title block'. The title block here refers to the printed area of the title page, or the horizontal and vertical distances between the outermost pieces of type on each side (inluding any border), i.e. the page size minus the margin. These measurements are given in millimeters.

Chapter 2

Annotation

This chapter will provide the guidelines for annotating the various visual, structural and semantic features of the texts that the project needs, using the elements defined by the *TEI Guidelines* (<http://www.tei-c.org/Guidelines/P5/>).

2.1 General annotation

This section will define annotations for various kinds of general features that can occur anywhere in the document, such as names, dates, etc.

2.1.1 Annotating names

The purpose of annotating names in the context of this project is mostly to enable their unambiguous association to specific people, places and institutions. Since names in early printed texts occur in a variety of forms, a unique reference is needed. This is accomplished by using the elements <persName>, <placeName> and <orgName> defined by the TEI guidelines for annotating these three types of names.

When the person (or an organization) referred to by the name is involved in the production of the book in some capacity (other than the author), this is indicated by the @role attribute. The values used for this attribute are:

@role	(on the <persName> element)
author	the person mentioned as the original author of the work, either in English or some other language from which this English version was translated;
contributor	a person who is not the author of the original work, but is mentioned as having contributed some additional content added to a later edition of the work;
translator	a person who is said to have translated the work into English;
source	a person on whose work the author is said to have directly based the work;
authority	a person who is mentioned as an indirect source or authority for the work;
influence	a person who is mentioned as an indirect or implied influence or inspiration for the contents of the work;
patron	a person who is mentioned as having employed the author or some other person involved in the book's production or otherwise supported its production;
opponent	a person who is mentioned as opposing the ideas presented in this book or to whose ideas this text is opposed;
printer	a person who is mentioned as having printed the text;

publisher	a person who is mentioned as having commissioned the printing of the text (i.e. "printed for";
seller	a person who is mentioned as a purveyor of the book.
other	some other person mentioned on the title page, whose role is either not mentioned clearly or does not fit into any of the other roles.

In order to provide an unambiguous referent (and a standard form) for the name, the @ref attribute is used to refer to a standardized form of the name and a description of the named entity in a separate list. This list is currently contained in a file called `List_of_names.xml` in the same place as this file. For each name annotated, check the list of names to see if the name has already been defined. If it has, use its @xml:id value, prefaced with a #, and if not, create an entry for the name (including an @xml:id value, which cannot contain spaces and must begin with a letter).

```
<persName role="translator" ref="#nicholas_culpeper">Nich. Culpeper</persName>
```

2.1.2 Annotating dates

Annotating dates with their standardized forms allows them to be extracted and used for various purposes. All dates, whether they consist of a full day-month-year notation or just a year, are annotated using the <date> element. The standardized form of the date is presented as the value of the @when attribute in the form yyyy, yyyy-mm, or yyyy-mm-dd, depending on the information present in the original date.

```
<date when="1663">1663</date>
```

NB! The publication year of a book, given in the imprint information, usually at the bottom of the title page, is annotated using the specialized <docDate> element (see subsection 2.2.1 on page 5).

2.1.3 Annotating prose paragraphs and lists

For annotating paragraphs of prose text, the <p> element is used. It should be noted that on the title page, most of the components are themselves considered 'paragraph level elements' which contain their textual content directly, without the need for an intervening <p> element; the exception to this rule is the <argument>, which must contain either a <p> element or a <list> element.

Lists can occur in the texts in a variety of guises, either numbered or unnumbered. All list structures are annotated using a plain <list> element with no attribute values. Within the <list> element, individual items of the list—including any list markers such as numbers or letters of the alphabet, which are transcribed normally—are annotated using the <item> element. If the list has a heading of some kind, this can be annotated using the <head> element.

NB! It should be noted that the list structure itself does not indicate line breaks, which should be indicated explicitly using the <lb> element as described in subsection 2.2.2 on page 6.

2.1.4 Annotating quotations

Quotations (which on title pages occur in the <epigraph> element) are annotated using one or more of the elements <quote>, <bibl> (bibliographic reference) and <cit> (citation), depending on whether the quotation is explicitly attributed or not. Quotations that do not contain an attribution (but are understood as quotations based on their context) are annotated using a simple <quote> element (in the case of title pages, within a <epigraph> element). Quotations that are either preceded or followed by an attribution, are annotated using the following kind of structure (for a discussion of the annotation of lineation, see subsection 2.2.2 on page 6):

```
<cit>
  <lb/><bibl>Exod. 1. 21.</bibl>
  <quote>
    <lb/>It came to ſpas, because the Midwives feared the Lord,
    <lb/>that God built them ſHoues.
  </quote>
</cit>
```

2.2 Annotating title pages

This section will outline the TEI XML elements used for modeling title pages of early printed books, along with instructions on how to apply them.

2.2.1 Logical structure of the title page

The logical and structural features of the title page are annotated in order to provide a context for the interpretation of visual features (i.e. where does a visual feature occur or what is it that is given more emphasis than something else). The structure of the title page is modeled using the elements defined in the *TEI Guidelines* under section "4.6 Title Pages". The parent <titlePage> element serves as the root element of the title page, containing within it everything found on the title page.

The <titlePage> element itself does not directly contain any textual content, but is made up of a selection of the following child elements:

<docTitle> (document title) contains the title of a document, including all its constituents. The different parts of the title are annotated by enclosing them within <titlePart> elements and indicating their types using the @type attribute and the following values:

@type	(on the <titlePart> element)
main	main title of the work (mandatory)
sub	(subordinate) subtitle of the work, somehow specifying the main title
alt	(alternate) alternative title of the work (usually begins with ", or")
desc	(descriptive) descriptive paraphrase of the work (often begins with something like ", being a…".
add	(additional) part of the title describing later supplemental parts added to the work; often beginning with "also, …" or "to which is added …").

<argument> A formal list (annotated using a <list> element) or prose description (annotated using a <p> element) of the topics addressed by the subdivisions of a text.

<byline> The primary statement of responsibility for the work, usually mentioning at least the author, and sometimes also the translator, or other people involved in the genesis of the text. The names of people mentioned are annotated using a <persName> element with an appropriate @role value, as described under subsection 2.1.1 on page 3.

<epigraph> Contains a quotation (annotated using the <quote>, <bibl> and <cit> elements, see subsection 2.1.4 on page 4 above), anonymous or attributed, that appears on the title page.

<imprimatur> A formal statement authorizing the publication of a work.

<docEdition> (document edition) An edition statement on a title page of a document (e.g. "The third corrected edition."

<docImprint> (document imprint) Contains the imprint statement (place and date of publication, publisher name), as given (usually) at the foot of a title page; will most commonly contain at least one personal name, a place-name and the document date. Any personal names are annotated using a <persName> element with an appropriate @role value, and any place-names using a <placeName> element. The publication year of the book is annotated using the special <docDate> (document date) element, which uses the @when attribute in the same way as the generic <date> element.

<figure> Groups together a graphical representation of a figure, illustration or other graphical element, and its description, encoded as a <graphic> element and a <figDesc> element. The <graphic> element is at this point used to encode the dimensions of the graphical element in millimeters using the @width and @height attributes (later on it can be used to contain a link to a graphical representation of the figure). The <figDesc> element contains a short prose description of the graphical element. More specific information on annotating graphical elements can be found under subsection 2.2.4 on page 9.

2.2.2 Visual layout of the title page

For the purposes of the *Gatekeepers of Knowledge* project, the visual layout of the title page is annotated explicitly and separately from its logical structure in order to enable the analysis of their interrelationships. The principal components of visual layout on the title page are *lineation* and *horizontal alignment* of text. Other issues to be considered are spans of *blank space* and text in *multiple columns*.

Line breaks

The lineation of text on the title page is annotated by placing an empty <lb> (line break) element at the beginning of each new line of text. Its relationship to any enclosing structural elements immediately following it is determined by the extent of the structural element: if the following element (e.g. a list item) is confined to a single line, the line break is placed outside of it, but if it spans several lines, also the first line break element is placed inside it. Also graphical elements are considered to be 'on the line', i.e. they should be preceded by an <lb> element if they occur on their own with no text following or preceding them on the line.

Horizontal alignment

The horizontal alignment of structural components is indicated using a key-value pair consisting of the key *align* followed by one of the values left, right, center and just on the @rend attribute (e.g. align(center)), indicating the justification of text within the element. The value left is considered the default and does not need to be indicated unless a parent element of the left-aligned element has specified some other value. In the case of spans of text that are not represented by a structural element, the semantically neutral <seg> element with the appropriate @rend value can be used to annotate the span.

Blank space

Since most title pages contain text printed in a number of different type sizes (see subsection 2.2.3 on page 7 below) and with varying line heights, empty lines cannot really be used to indicate blank vertical space (except for multicolumn layouts - see *Multiple columns* on page 7). Therefore the empty <space> element is used to represent blank space on the title page. It is used both for horizontal blank space on the line, mainly in the form of unusual indentation at the beginning of lines, but also blank space left in

the middle of the line, and for vertical space between lines. The attributes used to describe the extent of the blank space are:

@dim (dimension) Indicates whether the space is horizontal or vertical.

@unit Indicates the unit of measurement used for quantifying the space, which for the purposes of this project is mm (millimeters).

@quantity Contains an integer value, representing the extent of the space in the units specified.

NB! It should be noted that this element is used to annotate only blank space that exceeds what can be considered the normal interlineal or inter-word space. The extent of the space should be measured in such a way that it accounts only for the *extra* space. In the case of vertical space this means measuring from half an interlinear space below the upper line to half an interlinear space above the top line of the lower line (with the interlinear spaces estimated on the basis of the type sizes of the respective lines). For line-initial horizontal space, the measurement is made from the edge of the text block to the left edge of the first letter, as there is normally no space at all between them, but for line-medial space, the width of an average inter-word space should be deducted from the total distance between the letters around the space.

Multiple columns

Structural elements on the title page that are laid out in several columns (i.e. any blocks of text or graphical elements that occur next to each other on the page) are annotated using a combination of a @rend attribute value indicating the number of columns and <cb> (column break) milestone elements indicating the locations of column breaks. The number of columns that the structural element spans is indicated by adding the value cols(x), where *x* is the number of columns.In cases where only part of a structural element is laid out in multiple columns, the semantically empty <seg> element can be used to annotate the multicolumn segment and carry the appropriate @rend attribute. The column divisions are indicated by placing a <cb> element at the point where the column changes. The number of the column starting at that point is indicated by a numerical value on the @n attribute of the <cb> element.

NB! In terms of layout, these 'columns' should not be understood as rigid, equal divisions, but rather flexible containers that occur side by side on the page, whose size is determined by their contents. This means that for example two-dimensional bracketed lists—often occurring on title pages—can be represented using the column structure, each block of text and each bracket (or pair of brackets) is seen as a 'column' and annotated as such.

Any vertical space within a column, resulting from the unequal number of lines in the columns should be indicated using the <space> element. Strictly speaking, this is only necessary when the space occurs *above* text in the column, since the vertical length of all the columns in one column group is defined by the content of the longest column, so any trailing space in the other columns is implicit.

2.2.3 Typographical aspects of the title page

Since typography and the highlighting of textual elements using typographic means is a central concern of the *Gatekeepers of Knowledge* project, the typography of the title page is annotated to great detail. The size and typeface of each text segment on the the title page is annotated using the @rend attribute. The attribute accepts several values, separated by a space, which means that the typeface, its height in tenths of millimeters (rounded to the nearest even decimal, i.e. a fifth of a millimeter, to account for minor variation between different imprints of the same type) and possible other renditional attributes (described in subsection 2.2.2 on page 6) can be annotated on a single element.

Typeface and size

The typeface used is indicated by a key-value pair consisting of the key *type* followed by one of the following values in parentheses (e.g. type(roman)) on the @rend attribute of the appropriate element (see *Typographical elements* on page 8 below):

@rend	(on any element containing text)
blackletter	text printed in blackletter, or 'gothic', typeface;
roman	text printed in 'regular' roman type;
italic	text printed in italic or 'cursive' type;
swash	text printed in decorative italic or cursive type with exaggerated flourishes in place of serifs and curved lines frequently replacing what would be straight lines in regular italic.

The size of the typeface is indicated by a separate key-value pair consisting of the key *size* followed by a single-decimal numeric value, rounded to even decimals, to the @rend attribute, separated from the typeface indication and other values by whitespace (e.g. size(2.8)), representing the height of the letter measured from the baseline to the top of the tallest ascender (e.g. *b*, *d*, *l*, *k* or capital letters). In order to mitigate the effects of variation between individual impressions of letters, it is advisable to measure a few letters and average their height.

Special embellishment

In addition to different type faces and sizes, there are also other graphical means of highlighting and embellishing either single letters or longer stretches of text. Like the typeface and size, these are indicated using the @rend attribute on an appropriate element (usually <hi>). The following additional values can be used on the @rend attribute (and more can be defined as new kinds of phenomena are observed):

@rend	(on any element containing text)
smallcaps	text printed using small capitals for minuscule letters (the size of the type is measured from capitals as usual, and the height of the small capitals is not measured separately);
spaced	text printed with the letters spaced noticeably more apart than normally (not just for justification but for emphasis);
dropcap(*x*)	a capital letter that has its top is on the top line but extends downwards across several lines (*x* is the number of lines covered by the drop capital);
underlined	text that has been highlighted by a *printed* underlining (not sure if these occur);
overlined	text that has been marked as a numeral (or possibly otherwise highlighted) by a bar printed on top of it.
sup	text that has been printed in superscript.
sub	text that has been printed in subscript.

Typographical elements

The @rend attribute describing the type can be used on any element containing text, either a structural element like <docTitle>, <byline> or <docImprint>, or a general element like <persName> or <date>. In cases where no structural or other element coincides with the change in type, the semantically neutral <hi> element is used to annotate the text segment in question. The type of each span of text is thus considered to be defined by the *nearest ancestor carrying the @rend attribute with the relevant values*. It should be noted that the two values describing the typeface and its size are inherited independently,

which means that if only the size of the type changes, the typeface need not be respecified on the new element, but can be inherited from a more distant ancestor.

NB! The most important principle with these (and other) attributes is thus that of *hierarchical inheritance*, which means that an attribute value applies to all of the text contained—either directly or within descendant elements—unless it is overridden by a new value of the same typology (i.e. typeface, size, justification) lower down in the hierarchy.

The *default typeface* of the text (i.e. the typeface and size in which most of the regular running text of the book is printed) is established by the @rend attribute value of the main <text> element containing the text of the book. The default or 'baseline' typeface of the title page (if it exists and is different from that of the text) is established by using a @rend attribute with appropriate values on the <titlePage> element.

NB! Care should be taken to ensure that the nesting structure of <hi> elements and other elements describing the highlighting of items through type changes reflects the semantics of the highlighting. Although visually similar, the following two text segments are semantically different, which should be reflected in their encoding:

1. *One thing highlighted in italics* and *a second thing highlighted in italics.*

2. *First segment in a paragraph highlighted in italics followed by a bit* further highlighted in roman *and a second thing in italics.*

Assuming that both examples occur in a context where roman type is the default (i.e. the typeface specified in the <titlePage> or <text> element), the first example should be encoded as:

```
<hi rend="italic">One thing highlighted in italics</hi> and <hi rend="italic">a second
  thing highlighted in italics</hi>.
```

whereas the second should be annotated as:

```
<hi rend="italic">First segment in a paragraph highlighted in italics followed by a
  bit <hi rend="roman">further highlighted in roman</hi> and a second thing in italics.</hi>
```

I.e., the use of the highlighting element (or any of the more semantically specific elements) should indicate what is the thing being highlighted and what is the 'local default' or background typeface. In cases where a component of the title page—for example the title—consists of lines of varying type size with no clear hierarchy of highlighting, they should be annotated using individual <hi> elements for each line, apart from ones that are printed in the 'default' type defined on the <docTitle> or <titlePage> element.

2.2.4 Graphical elements on the title page

Considering the visual focus of the *Gatekeepers of Knowledge* project, it is natural to also annotate—at least to some degree—the graphical elements on the title page as well as the textual ones. At this stage, before any kind of automated pattern recognition and annotation system, graphical elements on the title page will be annotated by recording their dimensions, a brief prose description of them, and their location on the title page. This is accomplished by using the <figure>, <figDesc>, <height> and <width> elements defined in the *TEI Guidelines*.

A graphical element, whether an ornamental block or a real illustration (with few examples discussed below), is annotated by the following structure:

```
<figure rend="center">
  <figDesc><height extent="35mm"></height><width extent="120mm"></width>A brief prose description of the graphical
    figure, e.g. "an ornamental bar decorated with geometric shapes" or "a decorative horizontal
    divider made up of two acanthus-leaves extending to opposite directions".</figDesc>
</figure>
```

Some simple and frequent decorations that do not need a description—like horizontal lines—are annotated using the <graphic> element instead, allowing for their visual representation in HTML renderings of the title pages. The following annotations have been defined for some of the common elements:

horizontal line	`<graphic width="XXmm" url="hori_line.svg"/>`
left curly bracket	`<graphic width="XXmm" url="left_bracket.svg"/>`
right curly bracket	`<graphic width="XXmm" url="right_bracket.svg"/>`
	(More annotations can be defined as required.)

NB! Also graphical elements are considered to be 'on the line', i.e. they should be preceded by an <lb> element if they occur on their own with no text following or preceding them on the line. The height of a line containing a graphical element is determined by its @height attribute value (and it is assumed to be surrounded by some separating space, just like lines of text). *Inline* graphics, i.e. graphical elements that occur on the line among text, are treated just like text and assumed to be aligned to the foot of the line.

Decorative borders surrounding the entire title page are indicated using the @rend attribute on the <titlePage> element itself, with the following values:

@rend	(on the <titlePage> element)
border(line)	a border consisting of a single line surrounding the title page
border(2line)	a border consisting of a double line surrounding the title page
border(pattern)	a border made up of a repeating decorative or ornamental pattern
border(image)	an illustrated border that contains unique, non-repeating figures or images, possibly combined with repeating patterns