Document Analysis for Schema Design

Instructions

This worksheet is intended to guide you through some of the analytic process needed for an effective text encoding plan (and by extension effective schema design). The questions below can be applied both to an individual document, in preparation for encoding it, and also to a collection of documents, in preparation for designing an encoding plan for the entire collection.

There are many different ways to work through this analysis. It can serve simply as a thought experiment or a way of gaining an understanding of what would be involved in planning your data modeling work in detail. More practically, if you are actually working through the questions below, you may find it helpful to create a formal document recording the information you generate. Experimentation will help you learn what works best for you, but here are some suggestions:

- In a table or spreadsheet, create an inventory of document features as you respond to the questions in Section 1: Document Analysis.
- For each feature you've listed, list the elements and attributes needed to represent that feature. (For instance, if you list bibliographic citations as a feature, then you might next list the specific bibliographic elements needed for your representation.) You should also include in this list any new elements you will need to create. You can later refine this inventory based on a clearer sense of project needs (perhaps demoting or removing some features and their elements). Eventually, this inventory can then be used to guide the creation of a test document and a test TEI customization.
- For each element and attribute in the list, include a brief description of how it will be used (what specific features it will be used to encode, how they are to be recognized). These notes can later serve as the basis for your encoding documentation.
- For each element, identify the TEI module where it appears (which will enable you to quickly identify the set of TEI modules you'll need to include in your schema) and the attribute class for each attribute.
- For each element, indicate what special customization will be needed for specific elements (for instance, an added attribute or a change of model class).
- Later, as you develop your schema, you can use this document as a checklist, checking off the elements you've taken care of and keeping track of what remains to be done (and keeping notes on any remaining problems or loose ends).

Document Analysis

Overall structure and genre

[These questions focus on how your markup will represent the overall structural architecture of the document.]

What is the overall structure of the document? Does it consist of a single textual object or an aggregation (e.g. a complete works, a multivolume document, etc.)? Does it have front matter or back matter? What textual unit will be represented in your encoding as an individual <TEI> structure?

Will you be capturing a documentary version of the text (using <sourceDoc>) or a facsimile (using <facsimile>? (The analysis below focuses on a conventional <text> but if you are using one of these additional representations you can extend the analysis to include them.)

What are the major structural components of the document? If you were creating a high-level outline of the document, would it contain subdivisions (sub-subdivisions, etc.)? If so, what are these? (E.g. chapters, poems, acts and scenes, generic sections, entries, etc.)

What kind of classification of these structural components is appropriate for your project? (I.e. what are the values for @type on <div>?) Consider here whether a fine-grained or coarse-grained classification is most likely to be useful, given your audience and the use to which you'll be putting this information (analysis, navigation, formatting…?)

Structural details

[These questions focus on lower-level structural features of the text.]

Within the <div> structure of the text, what features of the text will you need to represent in order to provide for display, searching, analysis, or other processing? (For instance, if you are encoding a diary and want to be able to index and sort the entries by date, you would need to represent those dates in a way that allows them to be reliably identified and processed, for instance as <date> with @when inside <dateline>.) Examples include:

- Structurally important dates (such as dates for log entries, letters, or articles)
- Structurally important names (such as bylines, signatures, or names on title pages)
- Structurally important places (such as place names in letter headings or title pages)
- Bibliographic information (e.g. associated with quotations)
- Material that is rhetorically distinct: quotations, dialogue, asides, editorial notes, speech bubbles or captions in figures

This is just a set of illustrations, not an exhaustive list. Look through your documents and consider what features are structurally important to the way they'll be presented and used. For each feature you identify, consider what level of consistency in the representation you'll need in order to do the anticipated processing or discovery. What components are essential? Where do you need controlled vocabularies? Are there features of your text that are very consistent in their structure, such that you might benefit from a schema constraint to require them? For instance, should a letter be required to always begin with a dateline and end with a signature? List the specific elements and attributes that would be required, and any constraints on the order of elements that you would need to impose or test for.

Transcription

[These questions focus on how you will use markup to represent significant aspects of the transcription process.]

What forms of editorial intervention into the text will you be making as part of the transcription? Will these interventions be represented in the markup or done silently? For instance:

- Regularization or modernization of spelling, typography, or punctuation
- Correction of typographical errors
- Supplying variant readings from multiple witnesses

If the text is not perfectly legible, how will you handle illegible or difficult-to-read passages? What criteria will you apply in determining whether to treat a passage as illegible or simply unclear? (I.e. what level of certainty do you need to have in order to propose a tentative reading?)

Do you need to differentiate between different levels of certainty in your transcription?

Do you need to assign responsibility for specific readings and transcriptional decisions?

Document appearance

[These questions focus on how you will represent the appearance of the source document and the details of its material properties.]

What aspects of the source document's appearance are significant for your project and need to be represented explicitly as part of the markup? For instance: italics, typeface, alignment and justification, type size (absolute or relative), ink color, location on the page (for features like notes and handwritten additions), ornamentation.

What material properties of the source document are significant for your project and need to be represented explicitly (e.g. in the <msDesc>)? When considering what needs to be represented, think carefully about what pieces of information will actually be used in retrieval or analysis, and think about what form you'll need it in, to support those activities.

To what extent can this information be represented using a formal system (such as CSS or rendition ladders)? Are there any aspects of the document's appearance that cannot be formally described but still need to be represented? Would a note on the text suffice for these?

Components not captured

[These questions focus on features that you are explicitly excluding from your representation of the document.]

Are there any components of the document that can be omitted from your transcription altogether, based on your intended usage of the document: for instance, front matter, advertisements, running heads, non-authorial sections, footnotes, etc.?

Will these components simply be silently omitted or do you need to account for their absence in some manner (e.g. with <gap> or in the document metadata)?

Annotation

[These questions focus on editorial commentary and annotation that are represented as part of the markup (not on user annotations that are stored separately from the document).]

What forms of annotation or commentary, if any, will be included in the TEI encoding (whether in the transcription itself or in a linked document)? For instance:

- Commentary on specific words or passages (such as glosses, explanatory notes)
- General notes on the text as a whole
- Biographical information about people mentioned in the text (such as might be represented in a personography)
- Interpretive keywords, qualitative analysis codes

Consider whether these different kinds of information need to be handled differently in your output (for instance, authorial footnotes distinguished from editorial footnotes; biographical information linked from personal names; interpretive keywords visible only as part of the search interface). What encoding mechanism makes sense for each one, given the way it will be used?