Project 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). It assumes that you've already performed an initial document analysis, for instance using the worksheet at http://www.wwp.neu.edu/outreach/seminars/_current/handouts/document_analysis_worksheet_1.pdf

The questions below are aimed both at helping you assess the value of each markup feature you are considering, and also at helping you strategize the markup process. In this section, when we talk about "encoding this feature" we mean both "transcribing the content of this feature" and also "adding the markup necessary to explicitly represent this feature." In many cases, the important question is not whether you will transcribe a given feature (such as personal names) but what level of markup detail you will use, and how best to formalize the information being captured.

Project Analysis

Functional goals

In your document analysis, you've already created an inventory of document features and proposed elements that may be needed to represent them. Now, consider each feature in light of your project's functional goals and respond to the following questions:

- Will encoding this feature directly contribute to the functions you have identified as essential?
- Will encoding this feature contribute to functions that are highly valuable (or may possibly prove important in the future)?
- Will encoding this feature now (rather than later) save time or money, or enable you to take advantage of resources you have now and may not have later on?

Financial resources

Consider each feature in light of your project's financial resources:

- Will encoding this feature add very significantly to the effort required to capture the text? (Will it require any additional steps that take the encoder away from the text to do look-ups? Will it require extra layers of error checking or increase the overall likelihood of error?)
- Can you encode this feature using materials you already have available, or does it require access to additional materials?
- Will encoding this feature add significantly to the effort required for error checking and correction? Does it require specialized validation or data verification?
- Will it require additional programming effort to make functional use of this feature in your interface or output?

Staffing and expertise

Consider each feature in light of your project's available staff and their expertise:

- Will it be possible for your project to identify and encode this feature adequately given the staff expertise you have available? Does this feature require specialized subject knowledge? (If so, would the encoding process require two separate passes through the text by two different people, or would it make sense for one person to do all of the encoding?)
- Will encoding this feature require additional training of your staff?
- Will encoding this feature adequately require information that must be looked up or researched (e.g. checking whether a quotation is exact or not; checking the correct regularization of old-style dates; adding biographical information about people named in the text)?

Timing and project lifecycle

Consider each feature in light of your project's likely horizon of completion or activity:

- If your project has a short horizon of completion (e.g. imposed by a grant or personnel availability), can you identify a set of features that would make sense as a first encoding pass through the text, leaving more advanced features for a subsequent project phase? Features that "make sense" in this context might be features that would support a basic display of the text (to enable you to build a user base with a simple initial publication) or features that would enable you to demonstrate proof of concept for a prototype analytical tool (for instance, basic personography data to support a network diagram of a correspondence network).
- If your project has a phased support model (e.g. a sequence of grant proposals, or a repeated class assignment), can you identify a sensible sequence of encoding activities that would focus on capturing additional specific features over time? What would be the highest-priority features to focus on? What groupings of features would make sense (from the perspective of both training and error-checking)?