Document analysis

Let’s say you have a document and you want to produce a TEI-encoded edition of it.

Don’t read the entire TEI Guidelines to try to learn all the elements and attributes to encode your document.

Why?

The TEI Guidelines are meant to “apply to texts in any natural language, of any date, in any literary genre or text type, without restriction on form or content.” So they cover a lot of things, most of which won’t apply to a document of a particular type.

Even if you knew the whole Guidelines, there’s no point in encoding everything you can identify. Who has the time to identify everything, and what for?

What for, exactly? (1/3)

Let's say you have a single purpose for encoding a document (e.g., to be able to search separately words in the original manuscript and edits made by another hand).

But you may have ambitions to do other things, such as:

- adding your own annotations
- producing an e-book version of the manuscript for non-specialist readers

What for, exactly? (2/3)

And if you're encoding not for personal scholarly interests but for public consumption, you'll need to think about how the encoded text could help serve your users:

- a source for producing a transcription for a non-specialist reader?
- a source for producing an edition with corrections and annotations?
- data for linguistic analysis?

What for, exactly? (3/3)

Still, you can't cater to all possible uses.

If encoding for public consumption, you can encode for common uses and provide a foundation so that others can add more markup to the XML documents in the future to meet their own specialized needs.

Exercise

With a partner, choose one of the sample pages from various printed books. Discuss:

- What are the salient features?
- How would you instruct someone to recognize them?
- How do they relate to each other?
- What would you gain by marking up one feature set over another?
- Are there advantages to adding information? To normalizing information or making it explicit?
- Is there anything anomalous, inexplicable, or simply difficult?
- Are there any concurrent (overlapping or conflicting) organizational hierarchies?

For a challenge, pick a page with text in a language you don't understand. You'll be surprised how many features of a text you can pick out based purely on conventions of print!

for doing document analysis in the future

TIPS

If you have a large set of documents to encode

Be sure to pick a representative sample before you begin to see if there are any surprises in the structure. The documents may not all be structured like the first few you pick up!

And remember

Markup is an act of interpretation. You are making judgments about what you see in the source document. While some features of a source document are quite clear to anyone looking at it, there are other features that are open to interpretation. So there isn't a single correct way to represent a source document.

You'll be pleased to know that the TEI has mechanisms for you to encode multiple interpretations of what you see!

Questions?