2010-11-16

Digital Publishing and Preservation Using XML

Welzenbach, Rebecca; Schaffner, Paul; Hawkins, Kevin

What can you do with encoded texts?

To recap: what's the point?

- As a **publisher**, it allows you to regularize your content structure and appearance and to maintain a single source for your content that will be published in various formats.
- As a **researcher**, it allows you to do fine-grained searching instead of just full-text searching.
- As a **librarian or archivist**, it allows you to store the content in a widely used, open, non-proprietary format.

XML for publishing (1)

XML encoding of the *block-level components* of a document allows you to:

- Enforce a consistent structure on your documents
- Format a document's components consistently (and format a whole collection consistently!) … or globally change the formatting of a particular component
- Easily derive versions of your document in various formats (web, PDF, e-book, etc.)
- Automatically create a list of figures

XML for publishing (2)

XML encoding of *phrase-level components* could allow you to automatically create an index of personal names, place names, or key concepts (assuming these are all tagged).

Once you've tagged these components, you can then allow people to search on them:

"Give me all instances of 'Bush' as a name, not as a common noun."

XML for researchers

XML encoding of structural and especially non-structural components of a document allows you to query a corpus of texts.

- Find all instances of “Bush” as a name
- Find all instances of “rose” in verse, not as a name
- Give me a list of author names in bibliographic citations, regardless of whether these citations are given in footnotes or endnotes

Demos

- A London Provisioner's Chronicle, 1550–1563, by Henry Machyn: search within transcription or modernized text
- Newton Project: single source for both normalized and diplomatic transcriptions
- Oxford English Dictionary *(subscription access required)*: can display any combination of pronunciation, etymology, quotations, and date chart
- Middle English Dictionary: can restrict search to headwords, etymology, definition, quotations, etc.
- Data for Research: can restrict search to title, author, abstract, references, etc., plus limit query based on various metadata fields

XML for librarians and archivists

We'll hear more about this later!

Questions?