Digital Publishing and Preservation Using XML

Welzenbach, Rebecca; Schaffner, Paul; Hawkins, Kevin

http://hdl.handle.net/2027.42/109398
TEI workshop session 9

CHOICES

You can encode (mark up) almost anything.

But what should you encode?

And to what depth?

Your choices will be dictated by ...

- the nature of the material
- the character of your incoming data
- the amount of your funding
- the patience of your funders
- (how much time left till your retirement)
- the scale of the project (how many items)
- the scope and variety of the project
- the purpose of the project
  - desired functionality
    - expected (or guessed-at) audience
    - potential for repurposing
    - potential for sharing/reuse
• your own knowledge or ignorance

• the existence of standards
  o why to avoid them. They are:
    ▪ complex, hard to use
    ▪ not tailored to material
    ▪ not supported by local expertise or compatible with local systems
    ▪ not as good as what you can come up with yourself
    ▪ etc.?
  o why to use them. You can
    ▪ leverage community expertise
    ▪ share data
    ▪ share tools
    ▪ entertain at least a faint hope of "sustainability"
  o library practice summarized in GUIDELINES. (http://www.tei-c.org/SIG/Libraries/teiinlibraries/)
    ▪ a good starting-point
    ▪ provide good suggestions rooted in actual practice
    ▪ do not merely define but apply tags, with examples
    ▪ offer five 'levels' of commitment:
      1. raw OCR marked off into pages, linked to page images
      2. = LEVEL 1 + chapter divisions and headings
      3. = LEVEL 2 + refinements. Text may (?) stand on its own.
      4. Better text (keyed or corrected), tagged enough to stand alone
      5. = LEVEL 4 + considerable manual intervention based on subject knowledge.

Our own projects as examples:

  o (LEVEL 1) Mass-digitization projects Making of America (http://quod.lib.umich.edu/m/moagrp/) and the Google-scanned books going into HathiTrust (http://www.hathitrust.org)
(LEVEL IV) The Text Creation Partnership. See sample TCP file.

(LEVEL IV) Middle English Dictionary (and Compendium. See sample MED file.)

(LEVEL IV) Knight's American Mechanical Dictionary

(LEVEL V) The Faculty CV project

These differ in

- their adherence to standards
- their labor-intensity
- their longevity (the price of success?)
- their scale and scope

But share a common rationale:

- intelligible display
- intelligent navigation
- contextually useful search restrictions
- constraint by method and cost
- susceptibility to incremental improvement