

# BEST PRACTICES FOR TEI IN LIBRARIES

DIGITAL LIBRARY BROWN BAG SERIES  
20 OCTOBER 2010

Michelle Dalmau, Digital Projects & Usability Librarian  
Indiana University Digital Library Program

# Overview: TEI in Libraries



- History of text encoding in libraries
- Very high-level intro to TEI
- Motivations for text encoding
- Type of materials
- Principles governing text encoding
- Challenges with text encoding

# History of Text Encoding in Libraries



- Libraries were early-adopters of text encoding, motivated by access and preservation of digital texts
  - ▣ Release of the TEI standard (draft in 1990; first version, 1994) coincided with growth of the WWW in the early 90s followed by the emergence of digital libraries (as we know them today)
- More than half the membership of the TEI Consortium is made up of libraries

# Introduction to the Text Encoding Initiative (TEI)



- Technically: a standards organization for humanities text encoding
- Organizationally: an international membership consortium
- Socially: a community of people and projects
- Practically: a set of guidelines and XML specifications

# Overview: Motivations for Text Encoding



- Store information
  - Access
  - Preservation
- Share information
  - Searching/Browsing
  - Interoperability & Portability: Harvesting/Repurposing
- Analyze information
  - Linguistic analysis
  - Concordances
- Visualize information
  - Interactive timelines
  - Map-based interfaces

# Overview: Types of Materials



- Monographs
- Journals/serials
- Letters, personal journals, manuscripts
- Ephemeral materials usually held in archives (e.g., pamphlets)
- Administrative documents (e.g., meeting minutes)
- Legislative and judicial documents

# Overview: Principles Governing Text Encoding

- Representing the text (a.k.a. descriptive or document-centric markup)
  - Structural
    - Text divisions (chapters, sections, etc.), paragraphs, lists, tables, line groups, lines, etc.
  - Semantic
    - Metadata for the electronic and for the source document
    - References to people, places, events, organizations, etc. within the text (phrase-level)
  - Stylistic
    - Typographic features like bold, italics, small case, indentations, etc.

# Overview: Challenges with Text Encoding



- Presentation is variable (difficult to predict); structure, however, is constant
- Text encoding is not necessarily simple data entry/capture; interpretation and/or research are often at play
- Text encoding is not neutral or objective (thus the need for specific encoding guidelines to govern encoding projects)
- Text encoding is a strategic representation of the text (made more complicated by level of faithfulness to the source text)
- Often, there's more than one way to encode a particular aspect of the text

# Background: Best Practices for TEI in Libraries

- Version 3 (2008-2010): <http://purl.oclc.org/NET/teiinlibraries>
  - Meant to serve as a planning, training and implementation document
  - Attempt to provide a “true” library customization in response to TEI Lite (de facto “library” standard)
  - Revamped for TEI P5
  - Serve as low-barrier gateway to the official TEI P5 Guidelines
  - Address relationships between other metadata standards: MARC and METS
  - Introduce workflow scenarios for the TEI Header development and various levels (1-5) of encoding for in-house (as opposed to outsourced encoding)
  - Accompanied by tools (e.g., schemas and metadata mapping tools) to ensure conformance to the Best Practices

# Background: Best Practices for TEI in Libraries

- Version 1 (1998-1999): The “TEI Text Encoding in Libraries Guidelines for Best Encoding Practices” created as part of a Digital Library Federation (DLF) task force in response to the official *TEI Guidelines* first published in 1994
- Version 2 (2003-2006): Also created under the auspices of DLF, an updated version in response to the *TEI Guidelines P4 (XML)* release with additional encoding examples, tighter specifications for in-house and outsourced encoding and reframing to aid in training
- Version 3 (2008-2010): The “Best Practices for TEI in Libraries” created by the TEI Libraries SIG in partnership with DLF to be hosted by the TEI-C (as opposed to DLF) with focus on in-house encoding

# In-House v. Outsourced Encoding

- 1999-2006: Best Practices attempted to provide enough of a specification for both in-house and outsourced encoding, but this was problematic:
  - ▣ BP weren't detailed enough so another layer of local encoding "guidelines" were typically required making documentation for vendors unwieldy
  - ▣ Semantic markup (level 4) no matter how well-documented is easily botched by non-experts and non-native speakers
  - ▣ BP reflected instances of TEI tag abuse
  - ▣ Outsourcing encoding was an extremely costly endeavor: ("special application" of the TEI per project, high rate of errors, etc.)

## In-House v. Outsourced Encoding



- New reliable, outsource solution offered to members of the TEI-C: TEI Tite and AccessTEI
- TEI Tite is a TEI customization created as a vendor specification for a range of text commonly outsourced by Libraries (from print materials such as books to manuscripts in various languages and scripts)
- AccessTEI is a partnership with a reputable vendor, Apex Covantage, that supports transcription and basic structural encoding of various formats, language and sized-jobs at competitive costs

# In-House *and* Outsourced Encoding

- TEI Tite supports basic structural markup (between levels 3 and 4 as defined by the *Best Practices*)
- TEI Tite documents are missing the TEI Header and semantic markup
- *Best Practices* are intended to complement TEI Tite encoding
  - ▣ Tite documents can be transformed to level 3 (loss of some granularity) or level 4 (additional yet minimal manual markup required)
  - ▣ The BP can be used to develop the TEI Headers and apply richer markup as described in levels 4 and 5
- Learn more about AccessTEI: <http://www.tei-c.org/AccessTEI/>

# Introducing the Best Practices: Overview

- What?
  - <http://purl.oclc.org/NET/teiinlibraries>
  - TEI Header
  - Levels 1-5, from fully automatic to scholarly encoding
- Why?
  - Accommodate a wide range of *in-house* text encoding practice in Libraries
  - Support interchange and interoperability
- How?
  - Uniform set of best practices (with room for local adoption); meant to be used in conjunction with the TEI P5 Guidelines and local encoding guidelines
  - Mapping metadata tools (MARC => TEI Header)
  - Schemas for levels 1-4

# Introducing the Best Practices: Header

- TEI Header records metadata about the TEI document (e.g., encoding practices) and its source (e.g., bibliographic description) and a few other things (e.g., creation of e-text, revision history, etc.)
- TEI Header creation will vary and can be manually created or automatically generated via mappings/crosswalks
  - ▣ MARC to TEI Header
  - ▣ TEI Header to Dublin Core (DC) or Metadata Object Description Schema (MODS) for Open Archives Initiative (OAI) harvesting
- For manual creation or when enhancing automatically generated headers, a content standard should be used to ensure consistency (e.g., Anglo-American Cataloging Rules)

# Introducing the Best Practices: Header

- TEI Header recommendations and requirements are detailed in the best practices in support of both manual and auto-generated TEI Headers from MARC
  - ▣ Goal: Aid in metadata interchange or browsing/searching heterogeneous TEI documents
- Required for all levels defined in the Best Practices; recommendations/requirements are not informed by the levels

# Introducing the Best Practices: Level 1

- [Level 1: Fully Automated Conversion and Encoding](#)
- Purpose: Support keyword searching; assumes links to page images
- Level 1 considerations:
  - large volume of material is to be made available online quickly
  - digital image of each page is desired
  - no manual intervention will be performed in the text creation process
  - the material is of interest to a large community of users who wish to read texts that allow keyword searching
  - sophisticated search and display capabilities are not necessary
  - extensibility is desired; that is, one desires to keep open the option for a higher level of encoding to be added at a later date
- Workflow: Full text generated by uncorrected OCR (“dirty OCR”); page-level metadata inserted by software (e.g., Perl script)
- Example: [Indiana Academy of Science](#)

# Introducing the Best Practices: Level 2

- Level 2: Minimal Encoding
- Purpose: Support keyword searching; simple structural hierarchy to support document-centric navigation (e.g., chapter headings/ToC); assumes links to page images
- Level 2 considerations:
  - large volume of material is to be made available online quickly
  - digital image of each page is desired
  - the material is of interest to a large community of users who wish to read texts that allow keyword searching
  - rudimentary search and display capabilities based on the large structures of the text are desired
  - each text is checked to ensure that divisions and headers are properly identified
  - extensibility is desired
- Workflow: Full text generated by uncorrected OCR (“dirty OCR”); page-level metadata inserted by software (e.g., Perl script); next level of divisions (e.g., front matter, chapter headings, back matter ) are manually encoded
- Example: Swinburne as I Knew Him

# Introducing the Best Practices: Level 3

- [Level 3: Simple Analysis](#)
- Purpose: Create stand-alone full text display with emphasis on structural and stylistic markup; little to no content analysis; page images not required
- Level 3 considerations:
  - ▣ some sophistication of display, delivery, and searching based on structure of the text is desired
  - ▣ each text will undergo quality control to ensure that encoding decisions have been made appropriately
  - ▣ the creator of the texts has limited or no ability to provide content expertise to analyze, tag, or review texts
  - ▣ extensibility is desired
- Workflow: Full text generated by OCR (with some level of correction anticipated) or by transcription; “TEI shell” from level 1 or level 2 will need additional manual markup
- Example: [Victorian Women Writers Project](#) (phase 1)

# Introducing the Best Practices: Level 4

- [Level 4: Basic Content Analysis](#)
- Purpose: Create stand-alone full text display with emphasis on structural, semantic and stylistic markup and content analysis; page images not required
- Level 4 considerations:
  - sophisticated search and retrieval capabilities are desired
  - the texts will be used for textual analysis (e.g., thematic analysis, rhyme scheme patterns, etc.)
  - extensibility is desired; that is, one desires to keep open the option for a higher level of encoding to be added by the scholarly community at a later date
- Workflow: Full text generated by OCR (corrected) or transcription (double keyed); significant structural, semantic and stylistic markup (some may be automated, but by and large, manual)
- Example: [Indiana Magazine of History](#)

# Introducing the Best Practices: Level 5

- Level 5: Scholarly Encoding
- Purpose: Create deeply encoded texts for research purposes; requires subject expertise (usually partnerships between faculty members, technologists and librarians)
  - Semantic, linguistic, prosodic markup
  - Elements for editorial, critical and analytical markup
  - Manuscript descriptions
  - Translations
- Requires full use of elements as presented by the TEI P5 Guidelines; TEI Header Best Practices recommendations, however, apply
- Example: The Chymistry of Isaac Newton

# What's Next?



- Tools that accompany best practices:
  - ODD/schema specification for levels 1-4
  - Thutmose Project, set of style sheets that map MARCXML data into the TEI Header
- Vetting by the TEI Consortium as a sanctioned TEI “customization”
- Version 4, of course!

# The End. Questions?



Photo by Annie Leibovitz

# Thank You!

- Michelle Dalmau, Digital Projects & Usability Librarian, Indiana University Digital Library Program
  - [mdalmau@indiana.edu](mailto:mdalmau@indiana.edu)
- Bibliography and Resources (see handout)
  - TEI-LIB listserv