

Indiana Magazine of History

Encoding Challenges

Melanie Schlosser and Michelle Dalmau
Digital Library Brown Bag Spring 2007

April 4, 2007



LIBRARIES

INDIANA UNIVERSITY

Bloomington



Overview

- Introduction to the IMH project
- Introduction to the TEI and its use in the IMH
- IMH encoding challenges:
 - Text features
 - Subject encoding
 - TEI and serials
- Ways we considered encoding
- Solution: Independent Headers!
- The Survey
- Interoperability: Customization v. Standardization
- Conclusion



INDIANA UNIVERSITY

Indiana Magazine of History (IMH)

- Scholarly journal focusing on Midwestern and Indiana history
- Continuously published since 1905 in cooperation with the Indiana Historical Society
- Features peer-reviewed historical articles, research notes, annotated primary documents, reviews, and critical essays



INDIANA UNIVERSITY

Indiana Magazine of History (IMH)

- Significant subscription base and wide readership
- Supports nationwide interest in American history, covering the old Northwest, the Midwest and Upland South
- Broad audience: historians, genealogists, public librarians, students (secondary and post-secondary) and the general public



INDIANA UNIVERSITY

Indiana Magazine of History Online

- Collaboration between IMH editorial staff and the IU Digital Library Program
- LSTA funding to digitize and encode a 102-year run (from 1905 to 2006)
- Online version is freely accessible except for the most recent two years



Indiana Magazine of History Online

- Online version will support full-text and bibliographic searching and browsing with access to page images and text
 - Search by article title, author, article type, full-text
 - Browse by issue, places names
- ~41,000 pages to scan and encode (~400 per year) by a vendor
 - Images: Archival and derivative images, includes covers and color/grayscale images when illustrations are included
 - Text: OCR and encoded according to the Text Encoding Initiative (TEI), version P4
 - PDF: Articles converted to PDF for printing



Text Encoding Initiative (TEI)

- Text Encoding Initiative (TEI) / *Guidelines for Electronic Text Encoding and Interchange (TEI)*
- The TEI *Guidelines* "are addressed to anyone who works with any text in electronic form. They provide means of representing those features of a text which need to be identified explicitly in order to facilitate processing of the text by computer programs" (Sperberg-McQueen).
- TEI provides elements, attributes, and other mechanisms for encoding prose, poetry, drama, dictionaries, critical apparatus, linguistic corpora, and other scholarly and non-scholarly texts.



Text Encoding Initiative (TEI)

- **Bibliographic Metadata:** <teiHeader>
- **Structure:** <front>, <body>, <back>, <div>, <p>, <lg>, <l>
- **Content:** <persName>, <placeName>, <orgName>, <date>, <dateRange>, , <add>

[See guidelines](#) for complete list of elements



TEI and DLP Projects

- TEI is used in a wide range of DLP projects with similar wide ranging encoding levels
- *Indiana Authors and Their Books*
 - Basic markup of books using TEI Lite
- *Chymistry of Isaac Newton*
 - Scholarly encoding of manuscripts following locally developed and evolving guidelines
- *Swinburne Project*
 - Scholarly encoding of prose, poetry and critical essays by and about Algernon Charles Swinburne
- *IMH* encoding at intermediate level



IMH Encoding Challenges: Text Features

- Unusual and non-standard text features
 - Historical journal, contains a variety of content types
 - Tabular data
 - Primary source materials like letters and diaries
 - 102 years
 - Changed publishers
 - 1905-1913: Published by the editor
 - 1913 - present: IU History Department and IHS
 - Changed layouts
 - 1909
 - 1991



IMH Encoding challenges: Text Features

- Outsourcing
 - Spontaneous and iterative encoding in-house could deal with unusual features as they arise
 - Since we're outsourcing, have to plan in advance for everything that could come up
 - Have to communicate complicated decisions clearly with vendor across time, distance and communication barriers
 - May not be able to outsource work that requires textual analysis
- Budget
 - Limited budget necessitated selective encoding



IMH Encoding challenges: Text Features

- Solution: Detailed encoding guidelines!
 - Performed text analysis to identify unusual features
 - Pulled one volume each decade and documented features
 - Performed sample encoding and validating
 - Semantic
 - Worked with IMH staff to determine which features and content types were most important to users
 - Diverse readership requires range of encoding to support multiple uses
 - Syntactic
 - Providing page images along with text, so not important to replicate the page layout exactly
 - Chose only the syntactic markup necessary for legibility



IMH Encoding Challenges: Text Features

- Semantic features included:
 - Article types
 - Place names
 - Letters and diaries
 - Bibliographies (<list type="bibliography">)
- Semantic features not included:
 - Front- and back-matter, including table of contents, publishing info, and advertisements
 - Personal names
 - Lack of authority control would have made searching difficult
 - Full-text searching can achieve similar results



IMH Encoding Challenges: Text Features

- Syntactic features included:
 - Basic structural markup (as in TEI Lite)
 - Page breaks
 - Paragraphs
 - Headers and bylines
 - Lists and tables
 - Blockquotes
 - Footnotes (<list type="footnotes">)
- Syntactic features not included:
 - Poetry
 - Columns
 - Citations (<bibl>)
 - Same problem with authority control as names
 - Most citations can be searched in <list type="footnotes"> and <list type="bibliography">



IMH Encoding Challenges: Subject Encoding

- IMH Online Index
 - Contains subject indexing
 - Based on printed index
 - Structurally complex and not machine-readable
 - Writing Perl scripts to parse the data and extract subject terms
- Subjects in the TEI
 - TEI has no standard way to encode subjects. The solution we settled on is the most common.
 - We will add the subject information when we receive the files from the vendor



IMH Encoding Challenges: TEI and Serials

- Bibliographic information in the TEI
 - Bibliographic information about a TEI-encoded text is captured in the <teiHeader>
 - Each TEI.2 file can have only one header, and any given portion of text can only have one header that applies to it
- Bibliographic information in the IMH
 - Like most journals, articles in the IMH have two sets of bibliographic metadata: issue-level and article-level
 - It also contains book review articles, consisting of multiple reviews, each with its own metadata
- How to resolve this conflict?



Ways We Considered Encoding...

- Article-level TEI.2 documents, tied together with METS
 - Pros:
 - Can capture article-level metadata “the TEI way” - in the header of the TEI.2 document
 - Cons:
 - Lose the integrity of the issue
 - No way to include front- and back-matter
 - Still have the problem of book review metadata



Ways We Considered Encoding...

- Article-level and issue-level with <xptr> links
 - Pros:
 - Allows for full description of the issue and the articles within it
 - Allows for inclusion of front- and back-matter
 - Cons:
 - Still does not capture the issue as a text



Ways We Considered Encoding...

- TEI Corpus
 - A way to encode language corpora, which are texts (written or oral) collected for linguistic and other research. We could treat articles as ‘texts’ and issues as ‘corpora’
 - Pros:
 - Allows for the grouping of multiple, discrete TEI documents into a cohesive whole.
 - Considered a legitimate way to encode groups of texts
 - Cons:
 - The IMH isn’t really a corpus
 - Still does not allow for front- and back-matter



Ways We Considered Encoding...

- Issue-level TEI.2 documents with MODS records for article-level metadata
 - Pros:
 - Allows for full description of articles and book reviews
 - MODS is more machine-readable than the TEI, so it would be easier to reuse the metadata and integrate it with other resources
 - Cons:
 - Lose the TEI as the authoritative metadata source



The Solution: Independent Headers!

- What are they?
 - Standalone TEI Headers, enclosed in a document-level <ihs> element
 - Created to “build catalogues, indexes and databases that can be used by people to locate relevant texts at remote locations.” (TEI P4 Guidelines)
- Why are we using them?
 - Allow us to capture all relevant bibliographic information in TEI
 - Article-level
 - Book review (sub-article) level
 - Supported by the standard (no extension required)
 - Our text delivery system (XTF) currently configured to extract metadata from the TEI header



The Solution: Independent Headers!

- Why is this a controversial solution?
 - “Not the TEI way to do this.” - Syd Bauman
 - It creates ‘overlapping’ headers. Unlike stylesheets, TEI has no way to ‘cascade.’
 - There are theoretically other ways to capture this information in the TEI:
 - Corpora and the other approaches we considered
 - Repeating elements in the header
 - Extending the schema to allow for bibliographic metadata within the <body> of the text
 - Not supported in P5



Survey of TEI Community

- Informal survey of text encoding community distributed across a number of listserves
- Asked about
 - Use of the TEI to encode serials
 - Use of Independent Headers
- 16 responses from Digital Libraries, Digital Humanities Centers, and independent faculty members
 - 6 are using Independent Headers in some way
 - 10 are using the TEI to encode print serials (journals and newspapers)



Survey of TEI Community

- Conclusion: We are the only people using the Independent Headers as a way to capture more granular metadata in serials
 - Others are using them to:
 - encapsulate bibliographic metadata for multivolume publications
 - store and exchange records about their text collections
 - Most serials encoding projects are either:
 - Encoding at the article level
 - Encoding at the issue level and not capturing article-level metadata
 - Encoding at the issue level and using MODS to capture article-level metadata



The Goal: Interoperability

- TEI document as authoritative source from which we can derive functionality (METS, page turning application) and descriptive metadata (OAI harvesting)
- Reliance on standards for management, preservation and re-use of digital content
 - Self-documenting
 - Seamless integration with our infrastructure
 - Self-describing; can port and manipulate texts in other online contexts



Customization vs. Standardization

“The TEI's adoption as a model in digital library projects raised some interesting issues about the **whole philosophy of the TEI, which had been designed mostly by scholars who wanted to be as flexible as possible. Any TEI tag can be redefined and tags can be added where appropriate. A rather different philosophy prevails in library and information science where standards are defined** and then followed closely -- this to ensure that readers can find books easily. It was a pity that there was not more input from library and information science at the time that the TEI was being created, but the TEI project was started long before the term " digital library " came into use. A few people made good contributions, but in the library community there was not the widespread range of many years' experience of working with electronic texts as in the scholarly community” (Susan Hockey, *A Companion to Digital Humanities*, 2004).



Customization v. Standardization

“Digital libraries will be critical to future humanities scholarship. Not only will they provide access to a host of source materials that humanists need in order to do their work, but these libraries will also enable new forms of research that were difficult or impossible to undertake before.” (Howard Besser, *A Companion to Digital Humanities*, 2004).



In Conclusion

- On Independent Headers
 - We feel good about our unconventional use of the Independent Header! We learned a lot as we investigated solutions.
 - Alternative and viable options for representing article-or item-level metadata in TEI documents in the future:
 - MODS
 - P5 supports certain declarables (e.g., <biblStruct>) in the TEI Header
- On serials encoding with TEI ...
 - Resurrected the need for TEI to be less monograph-centered and support serials encoding, especially print-born serials that are issue-centric (inherent hierarchy)
- Spreading the word ...
 - Presenting at Digital Library Federation Spring 2007 Forum
 - Reporting to TEI council results of survey (“No one uses the independent headers!”)



References

- TEI P4 Guidelines: <<http://www.tei-c.org/P4X/index.html>>
- TEI P5 Guidelines: <<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/>>
- Besser, H. (2004). The Past, Present, and Future of Digital Libraries. In S. Schreibman, R. Siemens, J. Unsworth (Eds.), *A companion to digital humanities* (pp. 557-575). Oxford: Blackwell.
- Hockey, S. (2004). The history of humanities computing. In S. Schreibman, R. Siemens, J. Unsworth (Eds.), *A companion to digital humanities* (pp. 3-19). Oxford: Blackwell.
- Unsworth, J. (2000). The scholar in the digital library: <<http://digitalhumanities.org/view/Essays/JohnUnsworthScholarLibrary>>



INDIANA UNIVERSITY

Happy Birthday Melanie!

Happy Birthday to you

Happy Birthday to you

Happy Birthday deeeearrr Me-la-nieeee

Happy Birthday to you

(sing in our heads)



INDIANA UNIVERSITY

Questions? Comments!

- Melanie Schlosser: mschloss@indiana.edu
- Michelle Dalmau: mdalmau@indiana.edu

Thanks to Syd Bauman, John Walsh and Jenn Riley for the brainstorm sessions and encoding help.