

[title slide]

Our paper will review and evaluate two encoding standards and how together they can be used to drive functionality for electronic text projects with a specific focus on an open-source page turning system developed by Indiana University known as METS Navigator.

[slide of the various parts of the IMH]

Electronic text projects range in complexity from simple collections of facsimile page images to bundles of facsimile images and transcribed text from multiple versions or editions. Add a layer or two of metadata originating from the text itself or generated as a result of electronic conversion of the text, and the complexity only increases when it comes to the management and discovery of the online texts.

[Highlight the various components of a “digital object” as illustrated by slide; unpack diagram later.]

[slide of the IMH front cover and inside page]

To facilitate navigation and exploration within texts and across texts, an increasing number of digital initiatives and cultural heritage organizations are relying on the complementary

strengths of open standards such as the Text Encoding Initiative (TEI) and the Metadata Encoding and Transmission Standard (METS) for describing, representing, and delivering texts online. At Indiana University we are also leveraging and integrating both these standards in support of streamline access to and preservation of electronic texts. For this talk, to illustrate the use of TEI and METS, we'll reference as case studies two projects: the newly launched electronic version of the *Indiana Magazine of History*, a scholarly historical journal published quarterly since 1905, and the *Swinburne* project, an online collection devoted to Victorian poet and critic, Algernon Charles Swinburne.

[slide of the Swinburne screenshot]

Very little in the way of published literature addresses the relationship between TEI and METS, especially with respect to how the two schemes together can drive advanced functionalities such as page-turning, cross-collection searching and library catalog integration just to name a few. A recent article entitled "Cross-collection Searching: A Pandora's Box or the Holy Grail" by Susan Schreibman and her colleagues appeared in *LLC*, and it addresses the use of METS as a "container" scheme for the various metadata and full text representational formats that comprise the *MacGreevy Archive* in support of searching unique or idiosyncratic metadata within the MacGreevy archive itself and searching the archive along with other text collections.

The article also discusses the use of METS as a "valuable tool when building digital object repositories," and often METS serves as the core scheme for digital object repositories like the open source Fedora framework, which we use at Indiana University.

[slide of the MacGreevy and MBooks]

From a vastly different perspective – one of mass digitization versus thematic online collections – another article appeared in 2008 in *Library Hi Tech* entitled “OPAC Integration in the Era of Mass Digitization: the MBooks Experience” by Christina Powell, which chronicles University of Michigan’s experience with integrating bibliographic metadata for monographs usually in the form of MARC metadata with basic-level TEI-encoding that culminate in a METS file to facilitate discovery through their existing online catalog and page turning of newly digitized books as part of their Google project.

Both projects have to handle either the encoded text as the authoritative source for metadata and description as is the case with MacGreevy or handle an external authoritative form of metadata that is then linked to an encoded text. While we need to support both of these instances as part of our own Digital Library Program, this paper is primarily concerned with the TEI serving as the authoritative source of the bibliographic metadata and description of a document.

[slide of Motivations]

Because of the combined strengths of the TEI and METS, we are motivated to do the following as part of our digital library work:

- Establish a workflow to easily generates METS documents from TEI documents to facilitate text and facsimile correlation and page-turning (specifically for METS Navigator)
- Establish a single master source for e-text projects, i.e., the TEI document, for

descriptive, structural, and other metadata about the digital text object

- *Share* workflow documentation for page turning and digital object management and preservation. [Emphasis on “share”] As a result of an informal survey, we are aware that other organizations are leveraging TEI and METS for page turning as well as for other functionalities, but we found few details of how they are going about this process.
- Embed workflow (sample TEI, METS and XSLT) as part of the METS Navigator open source bundle (currently only sample METS documents are included)

TEI and METS

The Text Encoding Initiative Guidelines and the Metadata Encoding and Transmission Standard are two complementary, XML-based encoding standards that are pervasive in digital humanities and digital library projects and initiatives. As most of you know, TEI has a long history in both digital humanities research and digital libraries. METS, on the other hand, is more widely used in digital library environments, and is an integral component of digital object repositories such as the Fedora framework.

But digital humanities projects such as the Mark Twain project, Books from the Past, the Whitman Archive and the Swinburne project, are increasingly relying on METS to manage the various components of a digital object, reconcile duplication of metadata, or drive page-turning applications.

[slide of TEI-METS Definitions]

A hallmark of the TEI is its great flexibility and extensibility. It is difficult, therefore, to make overly determinate statements regarding the function and purpose of TEI encoding.

Nonetheless, while the TEI provides robust and flexible mechanisms for encoding a variety of metadata about a digital object, it is generally accurate to say that the focus of TEI encoding is the text or document itself.

METS on the other hand, as its name suggests, focuses more exclusively on metadata. While digital objects—such as a text, image, or video—may be embedded within a METS document, METS does not provide guidelines, elements and attributes for representing the digital object itself; rather, the aim of METS is to describe metadata about a digital object and the relationships among an object’s constituent parts.

[slide of TEI-METS Main Sections]

[just review dmdSec, fileSec and structMap]

<**dmdSec**> METS is designed to accommodate referenced or embedded XML metadata from other XML-based standards, such as “DC,” “EAD,” “MARC,” and “TEIHDR.” TEI enjoys a privileged role as one of the pre-defined metadata types for METS.

<**fileSec**>

<**structMap**> is arguably the core of a METS document. In this section the files are organized, grouped, and ordered for presentation (often hierarchical presentation as in the case of a book which has front matter, body, chapters within the body, etc.). It is the only required section of a METS document. METS Navigator uses a “physical” and “logical” structMap; the former is an ungrouped sequential representation of the facsimile image files, and the latter

organizes the files into a hierarchical structure and may be used to generate a table of contents.

It is clear that the TEI is a robust encoding scheme, more so now with P5's new <facsimile> elements and attributes, which provide explicit mechanisms for representing structure of a text and connections between text and images that can drive page turning. Some may even ask, in light of P5's new features, why even use METS at all? But there are compelling reasons why METS continues to fulfill important needs in digital humanities and digital library infrastructures:

- METS is far more prescriptive which supports interchange of metadata and objects in a more reliable, predictable fashion
- METS documents are conceptually simpler (skeletal structure) and potentially smaller in file size (no need to embed the actual text) which makes developing METS-driven software applications less cumbersome
- While TEI is focused on textual documents, whether represented as transcribed texts, facsimile page images or both, METS is able to conceptually represent the entirety multi-part digital objects of diverse media types, including text, audio, video, multimedia objects, and supporting applications.
- Lastly, METS establishes strong relationships between parts of a digital object that facilitate integration for software application-building; management of the various bits of an object; and preservation of the digital object when re-appropriated or for migratory purposes. So you get the benefit functionality and preservation in one.

[Slide Introducing METS Navigator]

In April 2006, we released a beta version of METS Navigator, a METS-based, open

source software solution for the discovery and display of multi-part digital objects. Using the information in the METS structural map, METS Navigator builds a hierarchical menu that allows users to navigate to specific sections of a document, such as title page, specific chapters, illustrations, etc. for a book. One can also navigate to arbitrary pages, by typing a page or image number into a form, or navigate to the next, previous, first, and last page images. METS Navigator also makes use of the descriptive metadata in the METS document to populate the interface with bibliographic and descriptive information about the digital object. The Navigator was initially developed for the online display and navigation of brittle books. However, realizing the need for such a tool across a wide range of digital library projects and applications, we designed the system to be generalizable and configurable.

Accompanying the METS Navigator source code is a METS profile, also expressed in XML, registered with the Library of Congress. The profile provides detailed documentation about the structure of the METS documents required by the METS Navigator application.

[Slide: METS Navigator is a metadata driven interface.]

[METS Nav for IMH: Slides 1 and 2]

The original version of METS Navigator is very book-centric. To support the hierarchical nature of serials, specifically to support the *Indiana Magazine of History* project, METS Navigator, and consequently the METS documents that power METS Navigator, are being completely overhauled.

Both versions are built on a Java Struts framework, and the new version includes technologies such as AJAX for dynamic display of XML encapsulated data and for enhanced user interactions with the online texts. The design of both the back-end and front-end is meant to

be “light weight” for easier adoption and integration of the software with a system-wide infrastructure or on a per project basis. This also makes it more attractive for non-Indiana University users to adopt the Navigator as well.

Back in 2007, a series of user studies were held with paper prototypes that represented various use cases for METS Navigator in terms of multi-part content types: manuscripts, journals, encyclopedias, sheet music, ephemera, etc., and though the new version of METS Navigator does not yet account for all of these scenarios, several upcoming projects under development by our Digital Library Program will launch in the coming year or two, creating enough critical mass to generate various “default” template layouts and METS guidelines catered to the various content types. We plan to bundle these templates as well as TEI-2-METS XSLT transformation style sheets with the next release of METS Navigator.

TEI2METS for the IMH

[slide of the various parts of the IMH: Just for some eye candy]

For the Indiana Magazine of History case study, it is important to clarify that this e-text project is tightly integrated with our Fedora digital object repository unlike our legacy TEI projects or Swinburne, which is in the process of migrating to Fedora. Because of this, much of the file processing and data mapping occurs as files are ingested, usually in batch mode, into the repository.

Bibliographic metadata derived from the TEI Header is mapped to MODS and Dublin Core, the structure of the TEI issue-level documents, based on TEI <div> and <head> elements are mapped to the METS structural maps required by the page turner, and so on. The METS documents powering the *Magazine* project’s implementation of the Navigator, as you will see,

rely on pointers by way of persistent URLs into our repository. Thus, the *Magazine* project does not rely on a single monolithic XSLT mapping file, but on a combination of modular XSLT that pull data from the source TEI file and from metadata already generated for storage with the TEI file in the Fedora repository. The Swinburne case study, however, is an example of a stand-alone XSLT file that generates complete METS documents from single corresponding P5 TEI documents. Thus one scenario is applicable for projects supported by digital library systems and infrastructures and another for projects that have not yet adopted such technologies.

[slide of the various parts of the IMH: Focus on issue-level encoding; article-level divisions that point to article-level headers. *TEI as authoritative source* for document and metadata.]

[XSLT Slides – impromptu]

[*Swinburne: stand-alone TEI P5 project, not using the new <facsimile> section but rather a configuration TEI file to define the facsimile images as required by METS Navigator.*].

Conclusion

[Show goals slide again with additional information]

The work completed to date has still been largely experimental and in the case of the *Indiana Magazine of History* hastily launched because we needed a version of METS Navigator far more sophisticated than our current version. However, the software's conceptual model is evolving based on our experimentation, and our new production version of the Navigator, released with the *Magazine* project in April 2008, has proven to be stable and performant. From the Swinburne experimentation, which we presented at DRHA 2007, we were able to generate for the first time complete METS documents, with full structural information, using only XSLT and the original source TEI documents. Additional e-text projects soon ready to launch, and hopefully feedback from others in the community using METS Navigator, will surely spur

additional development work.

Currently, generalization of the new functionality as seen in the *Magazine* project is underway. In the meantime, we have publicly posted our style sheets and soon we will post more comprehensive documentation surrounding the style sheets including guidelines for TEI markup to foster complete METS generation. Every new “development” even if experimental is posted on our wiki with caveats in the meantime. [*reference wiki*]

Already we have slotted to include several interface and functional enhancements including searching/browsing within the METS Navigator interface (within an item, within a collection and across collections), more seamless toggling from full text to facsimile image, various displays of text and image and the list goes on.

For us, it is not really an issue of the TEI versus METS, but rather we advocate the use of both, and stress the importance of rich TEI encoding. The TEI is so expressive, a METS file can be easily generated from the TEI through the use of XSLT style sheet transformations. In essence the METS files for driving page turning or any other discovery and exploratory functionality are nascent in the TEI.

In conclusion, as we explored the practical implications of generating an authoritative TEI document from which METS, MODS and other schemes could be derived, interesting questions arose about the relationship of the text itself, with all of its contextual and descriptive information intact versus the text solely represented by descriptive and structural metadata. A text intact, or an “authoritative encoding of the text,” especially one conforming to a standard like TEI lends itself to a great many things like self-documentation and self-description, which facilitates portability and re-use of the text in other contexts. But an authoritative text likewise lends itself to a distillation of a text that may be necessary to drive, at the very least, the basic

functions for exploration of the text in its entirety. Intersecting standards, whether TEI, METS or others, when used in tandem, often lead to incredibly engaging and sustainable digital applications that can more easily evolve or adapt in response to technical innovations or constraints.