The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0737-8831.htm

Challenges of serials text encoding in the spirit of scholarly communication

Michelle Dalmau
Indiana University Digital Library Program, Bloomington, Indiana, USA, and
Melanie Schlosser
The Ohio State University Libraries, Columbus, Ohio, USA

Received 26 December 2009
Revised 29 January 2010
Accepted 24 February 2010

Abstract

Purpose – The paper aims to describe an electronic text project involving a scholarly history journal, and to share findings related to the encoding of serials using the text encoding interchange (TEI) guidelines.

Design/methodology/approach – The project was completed using a combination of in-house and outsourced digitization and encoding, employing a variety of methods for quality control and encoding guidelines creation.

Findings – Evidence is provided that certain types of encoding should be done in-house, and describes a variety of mechanisms for capturing granular metadata in serials projects.

Originality/value – The paper covers a number of areas, including serials encoding using the TEI and granular metadata capture, which have not been explored elsewhere in the literature. It also provides guidance for others undertaking similar electronic text projects.

Keywords Serials, Text editing, Data handling, Digital storage, Digital libraries

Paper type Case study

1. Introduction

Published continuously since 1905, the *Indiana Magazine of History* (*IMH*) (www.indiana.edu/~imaghist/) is one of the nation's oldest historical journals. Between 1905 and 1913, the journal was first published independently by its founder, George S. Cottman, and then in conjunction with the Indiana Historical Society and Indiana State Library. Since, 1913, the *IMH* has been edited and published quarterly at Indiana University, Bloomington, while being offered as a benefit to the members of the Indiana Historical Society. Recently, the *IMH* features peer-reviewed historical articles, research notes, annotated primary documents, reviews, and critical essays that contribute to public understanding of the history of Indiana and the Midwest. It is the only scholarly journal that specializes in the history of a state that is plausibly described as the "Crossroads of America." Because of Indiana's pivotal place in American history, the journal has also been a leading venue for scholarship in the history of the Old Northwest, the Midwest, and the Upland South. With a significant subscription base (over 9,000 subscribers) and wide readership that includes historians, genealogists, and secondary and post-secondary students, the editorial staff of the *IMH* is committed to exploring alternative forms of access to the journal for increased exposure.

Currently, a subset of the *IMH*, from 2004 onward, is available online to members of the history cooperative (www.historycooperative.org/), a not-for-profit collaboration

Library Hi Tech
Vol. 28 No. 3, 2010
pp. 345-359

0737-8831
DOI 10.1108/07378831011076611

whose mission is to make available electronic resources pertaining to historical scholarship. Access to the content is made available by browsing the table of contents of a given issue or by searching basic bibliographic metadata and full text. Subscribers are provided the full text, including embedded figures and illustrations, but the actual facsimile page images of the journal are not available. Along with online access through the history cooperative, the *IMH* editorial staff partnered with the Indiana University Digital Library Program (DLP) in 2001 to create a freely available online index (www.letrs.indiana.edu/inmh) that provides author, title, and subject access to the articles' citations. However, the combined online access provided by the history cooperative and the online index was not satisfactory. The editors envisioned a resource that incorporated the strengths of each – full text access provided by the history cooperative and article level metadata access provided by the online index – bundled in one open access resource.

In 2006, the state of Indiana awarded the DLP a Library Services and Technology Act grant to digitize and encode a 102-year run (1905-2006) of the *IMH* and make all but the most recent two years accessible on the web free of charge. By digitizing and encoding the journal, the editorial staff sought to better serve their wide readership base by further expanding the journal's accessibility and utility. The goals guiding electronic conversion of the journal include:

- providing users with full text and facsimile page image access to the journal;
- enhancing article level metadata to provide more granular-level discovery of the journal contents;
- supporting deeper encoding of the text for additional access points such as place names and textual features emphasizing primary resources such as letters and journal entries;
- exploring subscription models that enable free access while not compromising the revenue intake necessary to publish the *IMH* in print form; and
- establishing an ongoing encoding workflow in conjunction with the publishers of the print edition for continual online access.

In April 2008, the DLP launched *IMH* online (www.dlib.indiana.edu/collections/imh/). A primary goal of the project is to cultivate and maintain an ongoing, online publishing model for a journal that still relies on print-based subscription for revenue and readership. In order to maintain a print subscription base, all but the most recent two years of the *IMH* are freely available online. In March 2009, the 2007 issues were freely released, and *IMH* will continue to make another year's worth of issues available every March.

2. Academic libraries and scholarly publishing

Libraries have historically housed, or cultivated strong partnerships with electronic text centers, many of which have evolved to encompass broader digital humanities initiatives, such as Maryland Institute for Technology in the Humanities (http://mith.umd.edu/), Institute for Advanced Technology in the Humanities (www.iath.virginia.edu/), and others. The centers situated in libraries benefit through access to content and expertise in cataloging and bibliography, both aspects crucial to a successful electronic text project (Sukovic, 2002, Para. 2). According to Sukovic (2002, Para. 25),

these partnerships benefit the library as well, since, in her view, "enriched electronic texts also have the potential to showcase the university as an electronic publisher." Beyond collaborations historically forged with electronic text centers, academic libraries are now increasingly engaged in online publishing of scholarly journals (Alexander and Goodyear, 2000; Borgman, 2000; McGann, 1996; Rao, 2001; Thomas, 2006; Waiijers, 2002), and have developed software platforms in support of electronic publishing such as DPubs and the Open Journal System (OJS). DPubs (http://dpubs.org/) was originally developed by Cornell University Library as an open-source content management platform, and was later extended to include a journal management component (Thomas, 2006, p. 570). OJS (http://pkp.sfu.ca/?q=ojs) is a robust journal-publishing framework developed as part of the Public Knowledge Project at the University of British Columbia (Willinsky, 2005).

Indiana University's own electronic journal-publishing platform, *IUScholarWorks Journals* (http://scholarworks.iu.edu/journals/) relies on OJS. Since, it was not yet in production when the *IMH* online project was begun, we were obliged to look for an alternate platform. Following Sukovic's (2002, Para. 25) supposition that text encoding forms the basis for journal publishing, we decided to leverage the electronic text expertise in the DLP by representing the journal in the eXtensible Markup Language (XML). The encoded text would then be able to both support innovative discovery mechanisms for readers and preservation of the content in our digital object repository.

3. Encoding serials: print vs born digital

In the world of online journal publishing, the structure of the journal is no longer bound by the constraints that shaped the conventions of print publishing. The *IMH* is historically a print-based journal, although its authoring, editing, and publishing in recent years have taken place in an electronic environment. Like most print journals, the *IMH* utilizes a traditional journal structure. It is hierarchical in nature (volume and issue) with a relatively predictable sequence of content (e.g. articles followed by book reviews, indices as part of back matter, etc.). The editors of the *IMH* felt strongly that the display of the content (e.g. article layout and footnotes) should reflect the print convention, but the discovery aspects – browsing and searching – should be more flexible. As Borgman (2000, p. 422) states:

> Print scholarly journals typically issue a fixed number of issues per year. These issues usually assemble a fixed number of articles in a fixed number of pages. Issues are printed and mailed as a unit. Without these physical constraints, electronic journal articles can be any length, can contain a mixture of text, images and sound, can be distributed to scholars as available. The unit of distribution could become the individual article rather than the issue or the journal.

In fact, while the *IMH* online can be browsed by issue, the searchable unit is the article.

While it is largely bound by its print roots, the *IMH* online project looked to born-digital journals such as the *Digital Humanities Quarterly* (*DHQ*) that are "experimenting with publication formats and the rhetoric of digital authoring" (http://digitalhumanities.org/dhq/about/about.html). In *DHQ*:

> All articles are given a detailed XML encoding to mark genres, names, citations, and other features that may serve the future scholar interested in the emergence of the digital humanities as a research field. As articles accumulate, the journal's interface will develop to

> exploit this markup through nuanced searching, visualization tools, and other modes of exploration (Flanders *et al.*, 2007, Para. 2).

Experimentation at the level that *DHQ* espouses – including experimentation with "audio-visual elements, executable programs and big datasets" (Waaijers, 2002, p. 169) – was beyond the scope of the initial launch of the *IMH* online, but we still committed to a rich-level of XML encoding that would afford us, at a later date, the prospects of exploring more dynamic interfaces. As Waaijers (2002, p. 169) described, XML:

> [...] facilitates the anatomizing of the internal structure of the document. Paragraphs, quotations, conclusions, etc. inside the article, can be coded separately such that the code represents metadata about the content of the fragment. The concluding step is the interlinking of these fragments, thus using them as building blocks for new "documents."

The *IMH* online, as detailed below, adopts an "anatomizing" approach to encoding to both facilitate more granular-level metadata and searching within important semantic units found within articles. The modular approach to encoding will also facilitate alternate views of the journal articles in the future.

4. The *IMH* and the TEI

The use of XML and XML-related technologies to publish online journals is not new to academic libraries (Cole *et al.*, 2001; Wusteman, 2003). In Wusteman's "XML and e-journals: the state of play," she provides an overview of markup standards and their evolution with an emphasis on XML-derived standards. Owing to the scholarly nature of the *IMH* journal and its wealth of primary sources, the journal was a prime candidate for the XML-based text markup standard, the guidelines for electronic text encoding and interchange (TEI) (www.tei-c.org/index.xml). The TEI clearly upholds Wusteman's goal to generate a discoverable, interoperable text that can then be re-purposed, managed, and preserved in the future.

The TEI provides elements, attributes, and other mechanisms for encoding prose, poetry, drama, dictionaries, critical apparatus, linguistic corpora, and other scholarly texts. The standard is extremely flexible and accommodating to most textual genres despite its bias towards humanities, and more specifically, literary texts. At the DLP, we use the TEI to encode administrative documents, legal documents, and ephemera, along with literary texts and manuscripts. Choosing the TEI as a text encoding standard for the *IMH* project was in keeping with our other e-text projects and digital library infrastructure.

In spite of the imminent release of P5 (www.tei-c.org/Guidelines/P5/), which officially debuted at the TEI Members' Meeting in College Park, Maryland, 31 October – 3 November 2007, the P4 version (www.tei-c.org/Guidelines/P4/) of the *TEI Guidelines* was chosen for this project. Though some amount of documentation for the P5 version had been circulating when we commenced this project in the autumn of 2006, it is our policy to wait for standards to be released and stable before adoption. Our implementation of the TEI independent header (IH), described in more detail below, also dictated use of the *P4 Guidelines*, as the IH is not supported in P5.

In strategizing the optimal encoding strategy for the *IMH* online, we faced a major obstacle: capturing adequate detail (both content and metadata) at the journal, issue, and article level, while maintaining the integrity of the original print document. The TEI *Guidelines* provide little guidance on how to represent journals, arguably because

the content characteristics (table of contents in the front matter, indices in the back matter, etc.) are similar to those of a monograph, whose representation is well documented in the *Guidelines*. Despite the similarities in content, the anatomy of a journal is distinctive in its seriality (Holmes and Romary, 2009, Paras 20-9).

We began our research by looking for other scholarly journal projects that utilize the TEI. The literature turned up little information, but we identified two journal projects that rely on the TEI. One, a born-digital journal called *Belphergor* (http://etc.dal.ca/belphegor/), is the first electronic journal published by the Dalhousie University Electronic Text Centre. The journal is encoded at the issue level following the TEI standard, and searching is at the article level (Hannon *et al.*, 2001, Para. 2). The search page "is designed to take advantage of TEI markup and allow readers to search using various criteria; for example, authors' names, titles of works, dates and places" (Hannon *et al.*, 2001, Para. 4). It is unclear without examining the markup whether the article level metadata is embedded in the article text itself or in some other, more structured format, such as the TEI header. The other example, the American Theological Libraries Association (ATLA) Serials Project, consists of a series of journals encoded as part of a print-to-electronic conversion project. Unlike *Belphergor*, the ATLA Serials Project has encoded at the article level using TEI Lite. While these two projects provided us with examples of encoding serials with the TEI, they did not offer any guidance on our most pressing problem: how to use the TEI in capturing article level metadata while encoding at the issue-level.

To gain a better understanding of current practice, we surveyed the digital library and digital humanities communities during the fall of 2007. We received only 16 responses, perhaps an indication of the relatively few TEI-encoded journals in production. The responses were helpful, however. A total of 60 percent of respondents were affiliated with digital library initiatives, 27 percent with humanities computing centers, and 13 percent were independent scholars or faculty members. All respondents use the TEI for encoding monographs, born-digital materials, and serials; yet only 22 percent of the respondents claim to use the TEI for encoding serials and 6 percent use the TEI for solely capturing bibliographic metadata. Of the 22 percent who use the TEI for encoding serials, half encode digitized print journals and the other half use the TEI to create born digital serial publications. For those encoding serials, most encode at the issue level rather than the article level. A majority captures bibliographic metadata in the TEI header, though a few make use of the metadata object description schema (MODS; www.loc.gov/standards/mods/) and the metadata encoding and transmission standard (METS; www.loc.gov/standards/mets/). A significant majority, 67 percent of the respondents, uses the TEI header as a stand-alone or IH; and to our surprise, 75 percent of those who have implemented IHs do so to describe bibliographic metadata at a higher level (e.g. series level) rather than at the article level. Only 25 percent of those who have implemented the IH do so for bibliographic description and exchange as originally intended by the TEI consortium.

5. *IMH* encoding guidelines

Since, text encoding projects are necessarily resource-intensive and we were working with a limited budget, we considered the encoding options carefully and based our decisions on the perceived needs of the users and the requirements of the text itself. The *IMH* staff does not keep formal records of reference questions asked by the readers of the *IMH*, but they were able to give us a general idea of the types of queries

users were likely to have. These included genealogical searches for personal and place names and regiment numbers, as well as subject searches by researchers and students. Using this information, we created personas representing the potential users of the online journal, including an amateur genealogist, an academic historian, a public librarian, and a student. Using these personas, we derived use cases from which to determine functional and metadata requirements.

IMH itself presented encoding challenges due to the variety of content types included and the changes in structure over its 102-year run. In addition to scholarly articles, tables of contents, indexes, and other text types commonly found in journals, the text of the *IMH* contains reprints of primary source materials from letters and diaries to election results. The presence of tabular data and highly structured text such as poetry posed structural difficulties, while foreign languages and a proliferation of proper names created the need for focused semantic encoding.

To determine what features we were likely to encounter during encoding, we performed intensive document analysis on the journal. Skimming one volume every ten years, we listed unusual structural features and semantic content that warranted closer consideration, and performed sample encoding on representative passages.

Keeping the needs of the users in mind, we used the results of the document analysis to derive a specialized tag set and detailed encoding guidelines. In our guidelines, we attempted to provide enough information for encoders working independently to identify important features in the text and make accurate decisions regarding the level of detail in the encoding. The guidelines remained a fluid document, and we revisited them frequently during the encoding and quality control (QC) process, clarifying points of confusion and making changes where necessary.

6. Semantic vs structural encoding

There is a fundamental distinction in text encoding between semantic and structural markup. This distinction proved to be a central factor in the success of the project. A familiar example of structural encoding is HTML, the tag-based language for publishing documents online. It is often pointed out as a weakness of HTML that it represents how something should look, not what it is. For example, in HTML the large, centered type at the beginning of an article can be encoded at the desired size, location, even color; it cannot, however, distinguish between a title, a byline, and an epigraph. More recent developments in online publishing, including XHTML and CSS, have attempted to shore up this weakness somewhat by separating presentation from content. At its heart, however, most web-based encoding remains structural in nature. There are structural elements in the TEI that capture how a piece of text should look, but they are generally de-emphasized in favor of more semantic markup. A TEI-encoded article would not only distinguish between a title and a byline, but could also identify normalized proper names and dates, include the text of footnotes at the point of reference, and represent bibliographic information in a structured way. These elements would constitute semantic encoding.

Via the process described above, we determined that the semantic categories of most interest to the users were article types, proper names of persons and places, bibliographies, and some categories of primary source materials. The article typing was intended primarily to facilitate searching and identification of relevant articles in result sets. Based on these tasks, we eventually settled on three article types: "scholarly article,"

"book review," and "editorial material." Names of people are arguably some of the most important content in a historical journal, and we struggled for quite a while with how best to capture them. In the end, we determined that since normalizing them was outside the scope of our budget, full-text searching would be the most useful mechanism for person name discovery, making further encoding unnecessary. With place names, however, we saw an opportunity to facilitate discovery via searching and browsing, by encoding all place names and normalizing when possible using machine-readable cataloging (MARC) country codes initially and later switching to a more robust controlled scheme, the Getty thesaurus of geographic names (TGN)[1]. Bibliographies were to be encoded as a type of list to allow more focused citation searching, and letters and diaries were to be captured using "section-level" tags with a type attribute.

These types of semantic encoding, while fairly basic by TEI standards, were at the heart of the discovery methods central to the project. What seemed straightforward to us while writing the guidelines; however, proved difficult to accomplish in the encoding phase. Some tasks that we had performed easily during sample encoding, such as identifying place names and letters, proved to be prohibitively difficult for our overseas vendor. They fluctuated between not identifying them at all, and over-identifying them when presented with similar names or structures. For example, the city of Cleveland was as likely to be missed as Cleveland Street was to be marked as a city. When we attempted to clarify our needs, we realized there were no easy rules or straightforward definitions for the types of markup we wanted. Like most semantic encoding, the features we wanted to capture had to be identified by the encoder's familiarity with the language and the source material. When asked how to identify a diary entry, it is difficult to give more precise instructions than "you just know." After many months of back-and-forth and additional documentation, we finally accepted that we would need to perform an additional round of encoding in-house to achieve the kind of results we wanted.

7. Article level metadata

The semantic encoding; however, was not the only challenge in our serials encoding project. We also faced some fundamental incompatibilities between the structure of the TEI and the multi-level hierarchy that is a print journal. In the TEI, bibliographic metadata is captured in the TEI header, which is a highly structured section at the beginning of every document that captures information about the source item and its digitized representation. While not as controlled, and therefore, not as machine-readable, as library-based standards such as MARC and MODS, the TEI header is a powerful and flexible means of capturing bibliographic metadata at the document level. The difficulty comes in when there is a significant amount of bibliographic metadata at more granular levels of the text – in this case, at the article level and the section level. Since, each TEI document has one and only one header, it is not possible to add another header to the document for each meaningful set of bibliographic data. That leaves the encoder with two options – either ignore the article level metadata altogether, or start looking creatively at the standard for ways to represent it. Since, the article is the fundamental access point for our readers, the first option was not a viable solution; we began exploring alternatives. In fact, we searched for several months, during which time we developed a number of possible solutions. Since, these solutions raise a number of issues for text encoding in libraries, and may be of use to others undertaking similar projects, it is worth describing each one, along with our reasons for rejecting them.

When we recognized the problem facing us, our first solution was the obvious one. Articles needed headers, each TEI document gets one header; we would simply encode in article level TEI documents and tie them together with METS. This solution fit comfortably within the standard and our technical infrastructure, but we quickly became dissatisfied with it. In this article-centric scheme, there is no place for front and back matter, and no way to represent the issue as a cohesive whole. It also left us with another level of orphaned bibliographic metadata – the book review. If we were to encode each article the way it is presented in the text, all book reviews for an issue would be contained within a single article. We were back to our original problem of a TEI document with multiple levels of bibliographic metadata.

Our second solution had us creating TEI documents for the articles and the issues, and then linking them using XML Pointer Language (www.w3.org/TR/2000/CR-xptr-20000607). In this scenario, the issue-level document would be a shell containing only front and back matter and links to articles. This solved the problem of extra-article material, but added a level of complexity to the technical implementation that was somewhat daunting. Our e-text delivery system would have to reckon with two levels of encoding – the issue and the article – dispersed across multiple files. This solution fell short in two other areas. It did nothing for the problem of book-review metadata, which is more granular than article level metadata. It also did not address our desire to represent the issue as a whole, the way it was originally published. Instead of a self-contained issue, this solution would leave us with an incomplete issue document, and a number of independent article documents.

For our next solution, we turned to the standard and found TEI Corpus. TEI Corpus is 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" tying them together. The attraction of this solution was that it provided a way to group multiple, discrete TEI documents into a cohesive whole. It was also a method generally accepted by the TEI community – the first reaction of most TEI implementers when presented with our problem was to suggest we use Corpus. After some sample encoding; however, we realized that Corpus does not allow the inclusion of any material outside of its member texts; we still had not found a home for front and back matter. This also seemed to be the wrong solution because the *IMH* is not a corpus. By representing it as such, we would not only lose the issue-level integrity we had been striving for; we would fundamentally misrepresent the nature of our text.

Our penultimate solution was one that has been implemented by others as reported in our survey of serials encoding projects. Since, issue-level TEI documents seemed to be the only way to capture front and back matter while faithfully representing the text, we decided to encode at the issue level and rely on another standard, MODS, to capture additional bibliographic metadata. This was an attractive solution for a number of reasons. It would have allowed us to capture all relevant bibliographic metadata regardless of the structural level to which it applied. A library standard, MODS is more controlled and machine-readable than the TEI header, so MODS records describing the *IMH* would have been easier to reuse and integrate with other resources. The deciding factor that kept us looking for another solution was the desire to keep the TEI as the authoritative metadata source for the project. In addition to wanting to take advantage of the full capabilities of this very powerful standard for describing texts, we were concerned that splitting the bibliographic metadata between standards would cause difficulties

integrating the project into our existing infrastructure, and in future use and preservation of the files[2]. Issue-level TEI documents with associated MODS records would be fully described, but they would not be fully self-contained and self-describing, which we felt was a goal worth pursuing. This notion of the self-describing text has been adopted by others undertaking TEI encoding projects such as the Modern Language Association of America's 2002 Guidelines for Electronic Scholarly Editions, which state: "the text itself should be essentially self-describing, which means that the computer file which embodies it should contain a header with essential 'metadata' (Faulhaber, 2002, Para. 9)." This self-description allows for easier sharing, aggregation, and repurposing of texts.

8. Independent headers

The solution we finally settled on utilized a little-known part of the standard known as the IH (www.tei-c.org/release/doc/tei-p4-doc/html/SH.html). IHs are standalone TEI headers enclosed in a document-level <ihs> element. The IH was created to "build catalogues, indexes and databases that can be used by people to locate relevant texts at remote locations" (*TEI P4 Guidelines*, Chapter 24). The IH was not originally conceived as a way to record hierarchical bibliographic metadata in serials, but since it was developed as a way to capture bibliographic metadata about text collections, it did not seem like an abuse of the standard. Using the IH for the *IMH* online had a number of advantages. Of primary importance was our ability to use the IH to capture all relevant bibliographic information at any level of the document hierarchy. As an existing part of the standard, it did not require customization or weighty alterations to the DTD, which supported our ongoing goal of interoperability and use of standards. Finally, since the IH takes advantage of the existing Header structure, it fit neatly into our current text delivery infrastructure, which relies on the TEI header for descriptive metadata. It also allowed us to come closer to reaching our goal of an entirely self-contained, self-describing text. Although the IH is not contained within the same file as the text of the article or book review it describes, it uses the same descriptive conventions and is easy to connect meaningfully via a system of identifiers and filenames.

Although the IH met our needs for this project, our use of it was not without a measure of controversy. When we first developed the idea, we contacted a number of knowledgeable people in the TEI community for feedback. We wanted to be sure that we had not missed a more obvious solution to our descriptive difficulties, and were no doubt hoping for the "blessing" of the community in implementing what we knew was a somewhat unorthodox scheme. While the responses we received were in some ways encouraging, they made it clear that our solution was "not the TEI way to do this." The "one text, one header" convention described earlier is not an error or an oversight – it is a fundamental component of the structure of the TEI. Each TEI document has one and only one header because each portion of text is supposed to be described only once. Unlike standards such as cascading style sheets, the TEI has no mechanism for resolving differences between overlapping descriptions or instructions. It is simply not designed to handle concepts like "this set of metadata applies to all text in this document, but when the text falls within a certain element, another set of metadata takes precedence." Our system of IDs and filenames, as well as our agreed-upon convention of one attached header per issue and one IH per article, prevented confusion within the context of the project, but did not resolve the fundamental conflict between our scheme and the structural underpinnings of the TEI.

The encouragement came in the form of suggested alternatives – not because they presented solutions that had not occurred to us, but because they did not. The alternatives suggested included ones that we had previously considered, such as TEI Corpus, and ones that were not feasible within the context of our project, such as repeating bibliographic elements in the header for each article in an issue[3].

9. Quality control

QC of TEI-encoded texts is difficult to do effectively without investing lots of staff time. There are automated forms of QC that can be used to an extent with the TEI, but in the end, encoding depends so heavily on the natural structure of language and semantic meaning that it is difficult to write machine-readable rules to validate it. We decided the best strategy was a combination of manual and automated QC.

When we received the test batch from the vendors, we performed intensive manual QC to see how well they were following the guidelines. This turned out to be an iterative process: we documented problems, communicated them with the vendor, and then performed the QC again on the revised texts. In fact, the QC process continued well into the "final" encoding as we continued to discover new problems.

As vital as manual QC is, it is not feasible in all situations due to limited time and resources. In fact, it is a rare project that has the luxury of detailed manual QC of all encoded text. It is more likely that some portion of the text will be checked by hand, while the remainder will undergo some form of automated QC. Automated QC is also useful when checking repetitive and predictable structures that would be monotonous for a manual checker. The simplest form is, of course, DTD or Schema validation. An XML editor (such as oxygen; www.oxygenxml.com/) can validate a file with the touch of a button, and scripts can be written to validate a number of files in a batch process. We included DTD validation in two steps of our in-house encoding workflow, and used it to catch errors in the original encoding and those introduced by the in-house encoders.

DTD validation is useless; however, when it comes to enforcing the decisions you have made about how to implement the TEI in your particular project. For that purpose we implemented Schematron (www.schematron.com/), which we had already been using to validate EAD-encoded finding aids. Schematron is an XML language that expresses rules through data patterns and assertions that can be applied to other XML documents. For the *IMH* project, we were able to enforce certain encoding practices governed by our encoding guidelines using Schematron.

10. Outsourcing vs in-house encoding: lessons learned

Outsourcing in this project was originally intended as a cost-effective way to digitize and encode. While the *IMH* was not one of our larger digitization projects, it would have encumbered a significant investment of resources and staff time to keep all of the work in-house. We have worked with vendors successfully in the past, including a recent experience with a very similar text encoding project, and we were confident that outsourcing the *IMH* encoding would produce satisfactory results. This was not the case. In fact, mid-way through the *IMH* project, we revisited the other in-depth text encoding project that we had outsourced and found many of the same kinds of errors. A long, slow process of trial and error with our vendor lead to some useful lessons that may be of help to others undertaking text encoding projects.

Lesson one: Keep semantic and difficult structural encoding in-house whenever possible. Sukovic (2002, Para. 19) argues that libraries are better situated to handle this level of semantic markup:

> Libraries have traditionally dealt with recognizing and naming various references to people, places, organizations, objects, events, etc. Semantic interpretation has been a regular library practice in assigning subject headings, choosing regularized forms of names, identifying languages used in a publication, and so on. A huge apparatus of codes and rules, thesauri, authority files and labeling systems, has been developed to support tasks of recognizing important information in the document and putting it in an accessible standardized form. The scholarly community depends on the library's interpretation of authorship and the content of whole documents, even corpora.

The increased cost of in-house semantic markup is justified by the quality of the work and the opportunity to develop staff competencies. If you do not have the resources in-house to encode, be careful and thorough in the vendor searching and contracting process. Just because the vendor says they can deliver what you are asking for does not mean they will do so in an accurate or timely manner. Be sure to provide them with a sample document to encode and check that their encoding meets your specifications before signing a contract.

Lesson two: QC is vital. It is not possible to be too detailed in early manual QC, and time invested in it can avoid headaches as the project progresses, and for other projects in the future.

Lesson three: Even if your vendor is doing a wonderful job, make sure to document your interactions with them. Important decisions, problems, and changes to your encoding instructions or guidelines are all worth tracking. At the beginning of the project, identify the people in your organization and the vendor's that should be copied on messages or updated regularly and keep them in the loop.

Lesson four: Your encoding guidelines should continually evolve in response to difficulties and discoveries made by encoders. If outsourcing, it may prove helpful to "test" the guidelines by encoding a subset of the text in-house. This should be done by someone other than the author of the guidelines. Once the project is outsourced it is important to establish a reliable form of communication with the vendors that allows for modification and updating of the encoding guidelines.

A more recent alternative to "customized guidelines" is a TEI vendor specification known as TEI Tite (www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_tite.doc.html), which includes structural markup only. TEI Tite was designed as a uniform specification for the library community to capture clear-cut, structural elements of print-based texts commonly found in prose, verse, and drama. Any semantic markup, including the completion of the TEI header and phrase level markup like personal names, place names, etc. is intended to be completed in-house. Perhaps, TEI Tite will prove to be the normalizing agent for outsourced encoding, and will serve as part of a balanced model that takes advantage of the savings usually associated with outsourcing and the domain expertise in semantic markup readily found in libraries.

11. TEI and serials encoding: alternative approaches

Since, the *IMH* online launched, the TEI community has been actively exploring issues surrounding the encoding of journals and their unique bibliographic metadata needs (e.g. author affiliations, author email, and other contact information). A recent article

by Holmes and Romary (2009) proposes a TEI customization known as "teiJournal" that would support the particulars of scholarly journal encoding while still maintaining full compliance, and thus interoperability, with the TEI schema. The authors of the article propose a born-digital framework for scholarly journal publication, but many of the principles apply to a print-based journal like the *IMH*. The TEI consortium has not yet approved the teiJournal schema customization, but it may prove to be a valuable resource for XML-based scholarly publications since it also provides a model ("publication engine") for web-based publication.

Another approach to capturing metadata for journals is a tighter integration between the Header and other bibliographic metadata standards such as MARC and MODS. If the TEI header had provided a mechanism by which we could explicitly point to other metadata records, perhaps we would have opted to capture the article level metadata in the more structured MODS format, which has advantages over the TEI header in terms of less ambiguous bibliographic structure, widespread adoption in the library world, and interoperability (i.e. shareable metadata via the open archives initiative – protocol for metadata harvesting).

The issue of integrating TEI headers with other metadata standards is being taken up by the TEI in Libraries Special Interest Group (SIG) (http://wiki.tei-c.org/index.php/SIG:Libraries), which is tasked with providing guidance for libraries engaged in text encoding. A significant function of the Libraries SIG is to recommend customizations and changes to the standard, which serve this large and growing base of TEI adopters. The SIG has spent the last year revising and updating the best practices for TEI in Libraries document (http://purl.oclc.org/NET/teiinlibraries). The guidelines for creating TEI headers underwent particularly intense scrutiny. While "there has always been communication between the TEI and MARC communities, often facilitated by e-text centers or text creation groups," libraries utilize a growing number of metadata standards, whose relationship to the TEI header is not always clear (Marko and Powell, 2001, p. 118). The recent revisions to the header recommendations focused mainly on appropriate use of header elements, with significant attention given to mapping between header and MARC records to facilitate automatic generation of headers from MARC catalog records. The issue of explicit linking from the header to an outside metadata record was discussed, but was determined to be too complex to be adequately explored within the existing time frame. It will, however, be a focus in the second round of revisions beginning in 2010.

12. Moving forward: leveraging markup for richer user interactions

One of the advantages to marking up a document in XML is the ability to transform that document in ways that can enhance the user interaction. As McGann (1996, p. 384) states:

> [. . .] when current scholarly journals publish their work online and/or in electronic form, they open their materials to integration within a scholarly network whose range and power outstrip current paper-based publication. Furthermore, electronic publishing permits scholars to present their work in far greater depth and diversity. Essays can present all their documentary evidence as part of their argument [. . .]

Borgman (2000, pp. 424-5) takes this idea further by suggesting that digital libraries should go beyond standard discovery mechanisms (e.g. browsing/searching) for our scholarly publications. Our tools should also allow scholars to interact with and visualize the text in new ways:

> One area of interest is the capability of digital libraries to support the cycle of information seeking, using and creating. By studying the ways that scholars perform these activities, such as how they disaggregate documents and then reaggregate them in different ways to construct new products, we can design better tools and services for digital libraries. Another promising topic is the "social life" of documents. Scholarly documents embody social processes that reflect how and why research was conducted. These processes are also reflected in scholars' perceptions of the literatures of their fields. By studying how scholars utilise documents, such as how individual documents are valued or annotated, we design annotation and retrieval tools that reflect more of the social life of documents. A third area in which these topics converge is electronic publishing. Most aspects of scholarly publishing, from document creation to editing and production, take place electronically, yet a substantial proportion of scholarly publications appears in print form. By studying the criteria scholars employ to select publication outlets, we can design better digital libraries in which to publish scholarly work, and better electronic and print information services to support publishing.

Following Borgman's and McGann's suggestions for richer, scholarly user interactions, future enhancements to the *IMH* online include browsing capabilities based on geographic places and features included in the Getty TGN. The TGN allows for more reliable and accurate discovery via place names in spite of variations in spelling – a common artifact in a journal that spans more than a century. Beyond that we plan to investigate geographic information systems and visualization technologies that will add dynamic content to the existing web site. This content could take the form of interactive maps and intertextual linking to maps from place name references in the articles themselves. Mapping the *IMH* is critical since it is deeply rooted in place – Indiana and the greater Midwest – and will serve the array of readers, from scholars to genealogists alike. The *IMH* editorial team is also in the process of normalizing over a century's worth of subject headings so that we can integrate the journal with the online index mentioned in the introduction for richer topical access. By integrating the subject headings with the TEI encoding, we will be better positioned to provide more advanced discovery and pedagogic functionalities as afforded by the XML-based topic maps standard (www.topicmaps.org/xtm/). Lastly, reader interactions by way of annotations and repurposing of texts are of key interest to the *IMH* editorial staff and something the DLP is currently investigating in the context of two other grant-funded scholarly encoding projects[4]. The tools developed as a result will be modular and, therefore, fully interoperable with the *IMH* online.

In the end, our decision to use the TEI standard, despite certain limitations, fully supports our commitment to open access scholarly communication. Borgman (2000) notes the duality of the publication workflow today; one that is largely electronic in terms of editorial and assemblage practices yet yields a print artifact. One of the primary goals of the *IMH* editorial staff is to support an online publication while maintaining a subscription-based print model and the DLP has treaded new ground: forging a partnership with the *IMH* print publishers so that we can leverage the existing electronic output of the print publication workflow in support of an ongoing, online publication workflow. For now the *IMH* editorial staff is able to maintain a small revenue base to maintain editorial operations, but exploring the impact of shifting to an exclusive online publication is imminent. Whether maintaining the XML-expressed TEI standard is the best solution for electronic publishing remains to be determined. It will depend on how journal publication frameworks like DPubs and OJS evolve to support more dynamic ways for readers to interact with the content. Until then, the DLP is committed to

providing access to the *IMH* online, preserving the content and investigating ways in which to provide tools for richer scholarly interactions.

Notes

1. The original encoding was outsourced and we felt a simpler place name controlled vocabulary mechanism such as MARC country codes would be more manageable, but after needing to essentially re-encode the entire run in-house, we took the opportunity to implement a more robust controlled vocabulary structure.
2. At that time, our e-text projects were exclusively stored in an alternative repository we dubbed Xubmit that stores exclusively TEI and EAD documents. However, later in the web development stages of the *IMH* Online, we opted to ingest this particular TEI project into our Fedora repository. In so doing, our earlier questions about integration with our infrastructure were moot, but the web development decision was made long after our decision to generate the bibliographic metadata using the TEI. In the end, the article level bibliographic metadata captured in the TEI IHs are mapped to the MODS scheme, which is then used by the delivery application, extensible text framework developed by the California Digital Library.
3. With issues containing up to 20 articles, the mega-headers now permitted in the P5 version of the TEI would have been difficult to encode, difficult to read, and painful to parse. When we realized that there was no alternative superior in form or function, we decided to stand by our unorthodox use of the IH. We experienced momentary hesitation when we realized that the IH was not supported in P5, the newest version of the TEI standard, but we eventually concluded that our best option was to continue to use the IH, which, after all, will always be valid, even when obsolete.
4. The Chymistry of Isaac Newton (www.dlib.indiana.edu/collections/newton) and The Swinburne Project (http://swinburneproject.org).

References

Alexander, A. and Goodyear, M. (2000), "The development of BioOne: changing the role of research libraries in scholarly communication", *Journal of Electronic Publishing*, Vol. 5 No. 3, available at: http://dx.doi.org/10.3998/3336451.0005.302 (accessed 19 August 2009).

Borgman, C. (2000), "Digital libraries and the continuum of scholarly communication", *Journal of Documentation*, Vol. 56 No. 4, pp. 412-30.

Cole, T., Mischo, W., Habing, T. and Ferrer, R. (2001), "Using XML and XSLT to process and render online journals", *Library Hi Tech*, Vol. 19 No. 3, pp. 210-22.

Faulhaber, C. (2002), "Preliminary guidelines for electronic scholarly editions", *Modern Language Association of America Committee on Scholarly Editions*, available at: http://sunsite.berkeley.edu/MLA/guidelines.html (accessed 14 December 2009).

Flanders, J., Piez, W. and Terras, M. (2007), "Welcome to digital humanities quarterly", *Digital Humanities Quarterly*, Vol. 1 No. 1, available at: http://digitalhumanities.org/dhq/vol/1/1/000007.html (accessed 25 August 2009).

Hannon, V., Roy, B., Frigerio, V., Barnstead, J. and MacLennan, O. (2001), "Building Belphegor: a multilingual electronic journal using the TEI", *Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, 2001*, New York, NY, poster abstract from ACH/ALLC 2001, available at: www.nyu.edu/its/humanities/ach_allc2001/posters/hannon/ (accessed 19 August 2009).

Holmes, M. and Romary, L. (2009), "Encoding models for scholarly literature", available at: http://hal.archives-ouvertes.fr/hal-00390966/fr/ (accessed 25 August 2009).

McGann, J. (1996), "Radiant textuality", *Victorian Studies*, Vol. 39 No. 3, pp. 379-90.

Marko, L. and Powell, C. (2001), "Descriptive metadata strategy for TEI headers: a University of Michigan Library case study", *OCLC Systems & Services*, Vol. 17 No. 3, pp. 117-21.

Rao, M. (2001), "Scholarly communication and electronic journals: issues and prospects for academic and research libraries", *Library Review*, Vol. 50 No. 4, pp. 169-75.

Sukovic, S. (2002), "Beyond the scriptorium: the role of the library in text encoding", *D-Lib Magazine*, Vol. 8 No. 1, available at: www.dlib.org/dlib/january02/sukovic/01sukovic.html (accessed 19 August 2009).

Thomas, S. (2006), "Publishing solutions for contemporary scholars: the library as innovator and partner", *Library Hi Tech*, Vol. 24 No. 4, pp. 563-73.

Waiijers, L. (2002), "Stratum continuum of information: scholarly communication and the role of university libraries", *New Library World*, Vol. 103 Nos 4/5, pp. 165-71.

Willinsky, J. (2005), "Open Journal Systems: an example of open source software for journal management and publishing", *Library Hi Tech*, Vol. 23 No. 4, pp. 504-19.

Wusteman, J. (2003), "XML and e-journals: the state of play", *Library Hi Tech*, Vol. 21 No. 1, pp. 21-2.

Further reading

Kapidakis, S. (Ed.) (2009), *Publishing and Digital Libraries: Legal and Organizational Issues*, available at: http://hal.archives-ouvertes.fr/hal-00390966/fr/ (accessed 14 December 2009).

About the authors

Michelle Dalmau is the Digital Projects and Usability Librarian for the Indiana University DLP, where she is responsible for coordinating and managing digital library projects with a particular focus on electronic text projects as well as coordinating and leading user studies for the DLP and the greater Indiana University Bloomington Libraries. Her research interests include the integration of complex metadata structures with discovery functionality of online collections as well as pedagogic use of digital resources. Her undergraduate background is in English and Art History, and she holds a Master of Library Science and a Master of Information Science from Indiana University. Michelle Dalmau is the corresponding author and can be contacted at: mdalmau@indiana.edu

Melanie Schlosser is a Metadata Librarian and Assistant Professor at the Ohio State University Libraries. She received a Master's in Library Science from Indiana University School of Library and Information Science in 2007, and previously worked with Indiana University's DLP as a Digital Library Fellow and Resident.