Computers and the Humanities **37:** 57–63, 2003.

Beyond the Web: TEI, the Digital Library, and the Ebook Revolution

MATTHEW GIBSON[1] and CHRISTINE RUOTOLO[2]
Electronic Text Center, University of Virginia, USA
[1]*E-mail: msg2d@virginia.edu*
[2]*E-mail: cjr2q@virginia.edu*

Abstract. Between August 2000 and August 2002, the Electronic Text Center at the University of Virginia distributed over seven million freely-available electronic books to users from more than 100 different countries. Delivered in a variety of formats, including .lit and .pdb, these ebooks have provided proof-of-concept for the adaptive uses of TEI standards beyond the World Wide Web – standards that the Electronic Text Center has employed since its inception in 1992. The first half of this paper discusses the mechanics of ebook production at the Etext Center, the limits of the current technology, and the conversion workflow we hope to implement in the future. The second half discusses user response to our ebook collection, classroom applications of ebook technology, and the advantages and disadvantages that different formats offer to scholars and instructors in the humanities.

Key words: collections, ebooks, library, Microsoft Reader, PDA, TEI, XML

1. The Theory and Practice of TEI-Based Ebook Production

Since 1987, the Text Encoding Initiative (TEI) has served as an international, interdisciplinary open standard, allowing content builders to represent and describe literary and linguistic documents for web-based research and instruction. TEI has remained useful and viable because of three important characteristics. First of all, it is stable. If one builds ASCII content and describes that content with the TEI, that content and its descriptive markup will outlast proprietary formats and will not be affected by the obsolescence of particular computer platforms or software programs. The TEI is also extensible. As technology evolves and users' needs change, TEI can accommodate such changes to remain useful for computing humanists – we have seen this especially with the development of XML versions of the TEI. Lastly, the TEI is malleable. One typically speaks of the TEI in the context of specific project development and the methods of TEI encoding used to structure and describe that project. Because it is so flexible, the TEI lends itself to the equally important task of enabling the reconfiguration and repurposing of data into a variety of different formats, for a variety of potential uses. The TEI's "interchange" capabilities provide the backbone for ebook production.[1]

Since the early 1990s, content producers have demonstrated the interchangeability of TEI tagging by dynamically converting their TEI-encoded documents into HTML for web presentation. While the web environment still provides important advantages for teaching and research, the advent of new ebook technologies, engineered specifically to enhance the readability and portability of digital content, offers an exciting new test of the TEI's adaptive power. With the launch of over 1,500 publicly-accessible ebooks for the Microsoft Reader and Palm devices, the Electronic Text Center at the University of Virginia has taken the promise of data interchange via the TEI into the realm of practice and product.

1.1. "EBOOK": OUR DEFINITION

To distinguish it from traditional web content, the Electronic Text Center defines the "ebook" as any full-text electronic resource designed to be read on a screen, in something other than a web browser. Ebook content can be read on a PC, a laptop, a PDA, or a dedicated reading device, in one or more of a growing number of available formats and software applications. With high-resolution font technologies and layout conventions borrowed from the print world, many current ebook platforms emphasize readability and strive to encourage onscreen reading for an extended period of time. Other implementations, such as handheld PCs and dedicated reading devices, emphasize portability. The most successful ebook solutions will likely offer some combination of enhanced readability and portability.

For the digital library, the ebook is a critical development. Just like a traditional library, a digital library must aggregate, preserve, and maintain information. At the same time, the digital library must make this information readily accessible and useful to a wide and diverse community of patrons, from the general public to scholars with highly esoteric needs. As gatekeepers to vast online collections, digital librarians must do more than respond to market-driven demands for new methods of content presentation and delivery – they must anticipate those demands and investigate the implementation of new technologies as they emerge. The more attuned libraries are to these technological innovations, the more they will be able to influence their development in ways that benefit diverse user communities.

1.2. CONVERSION METHOD

In its first phase of ebook production, the Electronic Text Center repackaged a substantial portion of its TEI collection as .lit files for the Microsoft Reader. Although it is proprietary and encrypted, Microsoft's ebook format is derived from the XML-based Open eBook (OEB) standard, an open specification developed to provide a basic structure for ebooks and extend their utility for users and publishers alike.[2] While OEB markup is much less robust than the TEI, it is quite effective as an XML "holding tank" for digital content, which can then be exported to a number of compatible ebook formats. Using simple Perl search-and-replace routines, the

Etext Center automated the conversion of over 1,500 existing TEI-encoded files into "extended OEB", a hybrid format which allows TEI tags to be carried over and accommodated with stylesheet instructions for aesthetic control. The TEI-to-OEB conversion produced documents that could then be exported into the Reader format with a piece of commercial software.[3] With minor adjustments to the Perl conversion scripts, we were then able to output our OEB files to the .pdb format for Palm systems and the .pdf format for the Adobe Acrobat eBook Reader. With this automated conversion, we have been able put the theory of the TEI's interchangeability – "build once, use many" – into well-documented practice.

1.3. Problems and solutions

Simply using the TEI is not the same thing as using it well. Having first implemented TEI standards back in 1992, the Electronic Text Center has had successes and failures with tagging methodology. A decade of experience has verified what common sense might suggest: consistent methods of encoding are always the wisest. While the TEI can describe complex textual features for the studied bibliographer or linguist, if the method of encoding employed is inconsistent and erratic, the data cannot be easily aggregated and its utility is diminished; repurposing for uses beyond the web becomes laborious and unwieldy.

When thinking of the future uses of TEI core content, it is especially important to consider complex formatting issues and the aesthetic presentation of a text. The principle of separating form from content is fundamental to the theory and practice of XML markup. Yet when one reflects on the role of digital libraries as content providers, and the diverse needs of patrons using library content, we see that engineering content for readability is just as important as engineering it for specialized humanities research. Without careful attention to formatting issues, ebooks derived from TEI content can be difficult to read, especially on small portable devices where screen size is limited. These problems are exacerbated in an automated production workflow. If, for instance, the TEI <table> element does not include attributes describing the number of columns and rows to be included, automating the construction of that table view for other formats becomes difficult and greater manual intervention is necessary. Increasingly, we find it necessary to record both structural *and* presentational features in the TEI markup, and our recent experiments using formatting objects (XSL: FO) to generate print-based output have driven this point home – the printed page is a much less forgiving presentational environment than the reflowing electronic screen.

Another problem we have encountered is the "bottleneck" of ebook production for the Microsoft and Adobe Acrobat eBook Reader formats. After the TEI content is batch-processed into the OEB format, the automated processing comes to an end. From here, we send our XML files one at a time through the encryption and compression software appropriate to each ebook format. While APIs exist that batch process XML into ebooks, at this time most of those programs are Windows

NT, not Unix, based. Thus, at this moment all of the .lit, .pdb, and .pdf files must live statically on the Etext server until we can either purchase hardware to accommodate the software or until ebook batch-processing APIs are developed for Unix systems.

Ultimately, we envision a delivery system where visitors to our website can choose to view and search all of our texts through the traditional web interface, or download them instantly and dynamically in a growing number of ebook formats. The success of this mission will provide the user with a greater amount of freedom and control in the way he or she wishes to access information.

2. Using an Ebook Library

Like similar SGML text repositories, the Electronic Text Center has taken pride in delivering richly encoded data that can be used for sophisticated searching and textual analysis. This focus on data structure, which has enabled us to build highly functional collections, has perhaps come at the expense of attention to design, aesthetics, and user interface. Our recent work with ebooks represents a new focus on the technologies of reading and how they impact our users.

2.1. Usage statistics and user response

So far, the new reading technologies have proven extremely persuasive. From August 2000 to August 2002, the Etext Center has recorded over seven million ebook downloads. Although only a small percentage of our total collection is currently available in proprietary ebook formats, where they are available these ebooks account for a significant portion of our online circulation. For example, we examined the March 2001 statistics for the public portion of our Early American Fiction collection, which presents side-by-side links to SGML, Microsoft Reader, and Palm versions of each text. We found that the SGML texts, delivered as HTML to a web browser, comprised 50% of the total usage; the Reader files accounted for 38% and the Palm files for 12%. In other words, visitors to the site chose a proprietary-format ebook roughly half of the time. As the ebook technologies become more entrenched, and we begin to make additional formats available from our website, we will conduct careful analysis of usage patterns, with a particular eye to how format preference varies among individual titles or content categories. This analysis should prove useful to academic content providers and commercial publishers alike, as no substantial analysis of ebook usage patterns has yet been made public.

2.2. Understanding a new type of user

While we know that our ebooks are being downloaded in great quantity, we still have a relatively poor sense of how they are actually being used. As humanists,

building our TEI collections for other humanists, we've necessarily made some basic assumptions about how these materials would be used. Our early focus on function over formatting probably reflected our anticipation of sophisticated academic users who mine vast quantities of data, extract very specific information, and then print, export, or repurpose that information. When we reformat our collections for the new reading technologies, we foresee a new type of user, about whom we are only beginning to learn. Obviously, we assume that our new ebook users are more focused on the act of *reading*. But many questions remain to be answered: How do readers really interact with ebooks? Do they read "immersively", for extended periods of time, as the hardware and software manufacturers insist? How do the activities that comprise an interactive reading experience – browsing, searching, annotating – differ in the ebook environment?

Clearly, ebooks present trade-offs to content providers and content users. In converting richly encoded SGML documents into ebook formats, we provide readers with the advantages of portability and a user-friendly interface. However, we sacrifice considerable functionality because the current ebook platforms are not SGML/XML-aware and do not support, for example, the kinds of complex hierarchical searching and analysis that fully-functional TEI markup allows. Furthermore, content embedded in proprietary ebook formats is not readily transformed or exported. As we have seen, it is relatively simple to convert marked-up data into these formats, but it is much more difficult to get the data back out again. Annotation features like marginal notes, highlighting, and bookmarking – all highly-touted ebook functions – lose much of their potential value if they cannot be extracted and reused in other contexts.

At the Etext Center, we have tried to minimize the impact of these ebook limitations with delivery systems that combine the functionality of encoded text with the ebook's ease of use. For example, we are using the raw SGML texts to create stand-alone indices of all materials within a particular subject area. Users can then perform cross-collection web searches that take advantage of rich markup and metadata, but have the option of retrieving their results in the reading format of their choice. The hierarchical division structure of our SGML also makes it possible to dynamically deliver discrete subsections of a larger work. A user browsing our Civil War newspaper collection, for example, will eventually be able to request all materials associated with a given day, week, month, or year, and have just those relevant materials packaged as an ebook for instant delivery.

2.3. PEDAGOGICAL CHALLENGES: EBOOKS IN THE CLASSROOM

Although our library serves a global audience – ebooks have been downloaded from about 150 different country domains to date – as a research university, we are primarily interested in how ebooks will impact academic research and teaching. To that end, in Spring 2001 we conducted two classroom pilot projects, in which we provided each student with a hand-held computer containing the textual materials

for the semester in Microsoft Reader format.[4] Superficially, the two courses seemed very different: one was an undergraduate-level course in Religious Studies, taught by a technology enthusiast, while the other was a graduate seminar in English taught by a technologically wary instructor. The courses did, however, have a few very interesting characteristics in common. Both syllabi were extremely text-heavy and relied extensively on rare and out of print materials, making ease of access an unusually important issue for the students. More importantly, both courses incorporated a broad range of text types, encompassing poetry and prose, fiction and non-fiction, with publication dates spanning several centuries; furthermore, both maintained a distinction between "primary" and "secondary" texts. These diverse materials required students to employ a variety of different reading methods, including close reading, selective reading, skimming, and "looking up" specific information.

The pilot classes provided anecdotal confirmation of some of our basic assumptions about ebooks. According to the faculty and students, the ebook's combination of easy access and portability was unquestionably its greatest advantage. Students benefited from immediate, round-the-clock access to all of their materials, some of which were not otherwise readily available. Convenience is of course an important benefit (students were delighted to browse through obscure 16th-century texts at the bus stop at 1AM, rather than in the rare book room) but the pilot participants noted more profound consequences as well. According to the religious studies instructor, ready access to comprehensive primary materials significantly changed the dynamic in the classroom. When the in-class discussion touched upon one author's startling claims about the Salem witch trials, the students instinctively began to search the voluminous court transcripts for proof. The condensation of a semester's worth of readings onto a handheld device also liberated the class from the fixed chronology of the syllabus – if the conversation leapt forward to a text not yet covered, the students could jump to that text and read it on the spot.

While they proved ideal for reading short excerpts and for nimble access to specific information, ebooks – at least this particular implementation of them – were less well suited to other types of reading. For example, students discovered that navigating lengthy Victorian novels was difficult due to limited search capability and the inability to retrieve a discrete region (like a chapter) from a larger work, as fully functional TEI would allow. In addition, the small screen display and automatic text reflow, while adequate for prose works of moderate length, were judged completely unacceptable for reading poetry, where preserving the original visual presentation is essential to preserving the meaning of the work.

In these and other examples, the pilot has underscored the importance of working closely with instructors to determine the optimal format for electronic classroom materials, as different formats facilitate different modes of user interaction. This need for user input applies not just to our recent experiments with ebooks but to all of the electronic collections we provide. While page images in a PDF-based reader might have little utility for a scholar doing linguistic analysis,

an instructor interested in the visual impact of book layout and typography may prefer page images to full-text transcription and encoding. Increasingly we find that we can't allow our own standard practices or assumptions about humanities computing to limit the range of presentation options we offer to our patrons. As we have seen, even within the scope of a narrowly targeted subject area, a variety of textual formats may be needed to meet an instructor's pedagogical or research goals.

3. Conclusion

Since it was established, the Electronic Text Center has pursued a two-fold mission of building SGML- and XML-based content while simultaneously educating and serving the community that will use this content. We see ebook production as an important part of both the research and public service aspects of our mission. We believe that our initial investment in building carefully encoded TEI data will enable us to accommodate new ebook technologies and other methods for delivering content as they evolve, and thus to anticipate the demands of our users. We hope that our high visibility as ebook content creators and distributors will, in some small way, help to foster a commitment to structured data and open standards in this emerging industry.

Notes

[1] The formal title of the TEI specification is, after all, *Guidelines for Electronic Text Encoding and* Interchange (see *http://www.tei-c.org*) – although TEI's rich and robust descriptive capabilities have traditionally received more attention than its potential use as a medium for data exchange.

[2] See OEBF (1999), "About the Open eBook™ Forum", available on the web at http://www.openebook.org/aboutOeBF.htm:

> The Open eBook Forum (OeBF) is an international trade and standards organization. Our members consist of hardware and software companies, publishers, authors, users of electronic books, and related organizations whose common goals are to establish specifications and standards for electronic publishing. The Forum's work will foster the development of applications and products that will benefit creators of content, makers of reading systems and, most importantly, consumers.

It is important to mention that, while the OeBF has a few academic and research institutions as voting members, the majority of interest and involvement comes from hardware, software, and publishing companies – companies whose primary concern is making the ebook a profitable commodity. The Electronic Text Center sees its own ebook development, and that of peer institutions, as an attempt to voice the "humanist's" stake in shaping the future of the OEB and related specifications.

[3] We used ReaderWorks software, developed and distributed by OverDrive, Inc. See http://www.overdrive.com/readerworks/.

[4] For a full discussion of the classroom pilot project, see Marshall and Ruotolo, "Reading-in-the-Small: a study of reading on small form factor devices", Proceedings *of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*, July 2002.