Journal of the Chicago Colloquium on Digital Humanities and Computer Science
Volume 1 Number 3 (2011)

TEI Texts that Play Nicely: Lessons from the MONK Project

Brian L. Pytlik Zillig, Center for Digital Research in the Humanities, University of Nebraska-Lincoln

Abstract

Text curation, like most human endeavors, requires tools. A technique developed for the MONK Project, schema harvesting, provides a useful platform for facilitating the digital conversion and curation of text corpora. The author describes Abbot, an XSLT-based application that has had success in converting various Text Creation Partnership collections, and others, during and after MONK.

Introduction

For centuries, the core activities that libraries performed changed slowly. Libraries mainly collected books and things that looked, more or less, like books. Then they took care of those book-like things and planned to keep on taking care of them. The act of providing for those things can be described as curation. That is, to be specific, an interconnected sequence of activities that includes selection, collection, organization, maintenance, and preservation of objects. Curation, like most human endeavor, requires tools. Unsworth has noted that "We've spent a generation furiously building digital libraries, and I'm sure that we'll now be building tools to use in those libraries…. I'm sure that the texts won't go away while we do our tool-building—but I'm also certain that our tools will put us into new relationships with our texts."[1] While Unsworth was not referring explicitly to digital curation tools, his prediction may fit that domain as well.

Curatorial Attention

Digital technologies produce, among other things, new objects of the sort that libraries have reason to collect and that require new forms of curatorial attention. Such attention differs from analog curation mainly in the details, and these differences are significant. For traditional analog objects, such as books, journals and microforms, access to traditional forms of text typically entailed static instances of those texts. Digital objects, however, have special curatorial needs and libraries must confront them. Intelligent digital libraries, as Crane calls them, can "allow a greater number of users to make more effective use of a wider range of their holdings than was ever feasible in print."[2] This assumes a process where digital holdings are gathered in ways that support combinatorial activities. Besser takes the view that "[i]n moving from dispersed digital collections to interoperable digital libraries, the most important activity developers need to focus on is standards."[3] The present research is concerned with a single aspect of digital curation: conversion of incompatible file formats into a standard form: TEI.

[1] John Unsworth, "Forms of Attention: Digital Humanities Beyond Representation," Canadian Symposium on Text Analysis, McMaster University, November 19-21, 2004, accessed June 3, 2011, http://www3.isrl.illinois.edu/~unsworth/FOA/.

[2] Gregory Crane, "What Do You Do with a Million Books?" *D-Lib Magazine* 12, no. 3 (2006), accessed June 3, 2011, http://www.dlib.org/dlib/march06/crane/03crane.html.

[3] Howard Besser, "The past, present, and future of digital libraries," in *A Companion to Digital Humanities*, eds. S. Schreibman, R. Siemens, and J. Unsworth (Oxford: Blackwell, 2004), accessed June 3, 2011. http://www.digitalhumanities.org/companion.

Source URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

Incompatibility has several causes. Sometimes it is the result of substantive differences in how to construct the data model for texts. But in the case of the large-scale encoding projects located in university libraries over the past two decades, the most common cause has been the result of encoders making choices that were sensible and convenient when considered on their own, but which paid little attention to the resultant divergence and the cost it would impose on future users who might want to manipulate or search texts across different collections. The fact that such "cross-walking" was not technologically feasible when many these projects began undoubtedly contributed to the problem. It is technologically feasible now. Our texts will benefit from procedures to render them interoperable.

Schema-Harvesting

Beginning in early 2007, the Mellon-funded MONK Project sought to develop a procedure for batch-converting dissimilar collections of XML texts into a specialized application of TEI P5 called TEI-Analytics (TEI-A). TEI-A is closely related to TEI-Lite with the addition of linguistic annotation. The effort to develop a conversion procedure yielded a command-line application, dubbed Abbot to fit the MONK theme. Bradley observes that, "for feeding data to many programs, it is likely that XSLT will be used more and more to transform one type of XML markup into another."[4] Abbot's strategy for arriving at a uniform text corpus has been to rely on XSLT. Abbot's key feature is the ability to use a small XSLT stylesheet whose purpose is to read the TEI-A schema file and output a second stylesheet, one which converts source files into a form that validates against the TEI-A schema. The technique was called schema-harvesting.[5] It worked well, and precisely because it worked, it seemed pretty clever for a while (foreshadowing intended).

Abbot's schema-harvesting procedures focus on TEI, but it is worth emphasizing that this approach is extremely flexible and format agnostic; Abbot makes no particular judgment or demand concerning the type of interoperability that is sought, and can transform texts into any arbitrary XML schema.

The current TEI schema (known as proposal five or "P5") is not backwards compatible. This more or less forces institutions at some point to convert their older files. Using conventional methods involving human intervention the possibility for divergence and error is great. A third-party conversion tool like Abbot is more likely to spot and deal with such problems. TEI P5 creates the need to do something and an opportunity to do it properly. It may be useful to see Abbot's work as contributing to the Collections Interoperability effort currently underway within Project Bamboo. That project views "algorithmic operation across textual collections"—which a common form of TEI may facilitate—as highly desirable.[6] Abbot, like Bamboo, sets its course on an ambitious but sensible path—moving toward total interoperability, while at the same time accepting the uniqueness

[4] John Bradley, "Text tools," in *A Companion to Digital Humanities*, eds. S. Schreibman, R. Siemens, and J. Unsworth (Oxford: Blackwell, 2004), accessed June 3, 2011, http://www.digitalhumanities.org/companion.

[5] Brian L Pytlik Zillig, "TEI Analytics: converting documents into a TEI format for cross-collection text analysis," *Literary and Linguistic Computing* 24, no. 2 (2009): 187-192.

[6] Martin Mueller, "Thoughts About Corpora Space and Collections Interoperability," *Project Bamboo*, March 1, 2011, accessed June 9, 2011, https://wiki.projectbamboo.org/display/BTECH/Thoughts+About+Corpora+Space+and+Collections+Interoperability.

of individual text collections. Abbot's method allows for a number of different forms of interoperability from one-off instances to the creation of large, permanent digital libraries.

Abbot was architected in ways that seemed sensible: it used Unix tools (mainly shell scripts) and XSLT. By mid-2009 Abbot had successfully converted 2,585 texts representing seven collections, 806 authors and more than 151 million words. Texts from two Text Creation Partnership (TCP) collections were included: 691 from EEBO and 1077 from ECCO.[7] MONK did not explicitly seek to plan for a ten-fold (or greater) multiplication of Abbot's work, but it was left as a tantalizing possibility.

Gold asserts that a "great challenge of data curation is ensuring that data, once preserved, remains meaningful either within the same research area or ideally across areas or even across domains."[8] It is a substantial challenge of digital curation that distinct but similar collections can be made to interoperate by the lossless (or nearly so) conversion into a common format such as TEI-A. The present paper describes efforts to extend the capabilities of Abbot to convert all 31,000 TCP files into lossless TEI P5.

An unfortunate side-effect of accelerating technological advancement is that achievements that once seemed miraculous quickly begin to appear rather small and wanting. So it was with Abbot. For the present research into converting TCP texts, it quickly became clear that while Abbot development had paused after MONK was finished, the technical environment had not. To be specific, the Relax NG schema that Abbot uses to create the conversion procedure had changed in substantive ways that Abbot could not follow and necessitated some adjustment. Once the program was schooled in the current Relax NG schema format, work began to convert the 31,000 TCP files. Abbot, which presently contains approximately 198 templates and 9,000 lines of XSLT, relies on a pipeline involving more than a dozen shell scripts that replace:

- EEBO header with TEI header
- uppercase elements with camelcase
- numbered divs with unnumbered divs
- pb by changing ref attribute to facs and preserve n attribute
- character entities with standard UTF8 characters
- the &s; with 's'
- the tilde character with macron substitute
- superscript letters represented by '^'preceding each letter
- pipe character to represent line-terminal word break
- plus sign used to represent soft hyphen
- decorated initial character signified by an underscore, with <seg type="decorinit">
- gap with unclear
- letter element with floatingText
- table or text element as last element in a p element (turn the "last child" into "next sibling")
- lang attribute with xml:lang

[7] John Unsworth and Martin Mueller, "The MONK Project Final Report" (2009): 3.

[8] Anna Gold, "Data Curation and Libraries: Short-Term Developments, Long-Term Prospects." Office of the Dean (Library) (2010), accessed June 3, 2011, http://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1027&context=lib_dean.

- q with quote

A randomly-selected set of approximately 6,000 EEBO texts took a bit more than 24 hours to complete, but at the end 96.5% of the text files were valid TEI P5. The full set of 4,000-plus TCP Evans files were processed in roughly 14 hours and ~98% were valid. Work continues on extending the Abbot software and it is anticipated that all TCP files will pass through Abbot in the near future, and that 98% of them will parse.

If this feels a bit like a qualified success, the participants of the Abbot Project (Stephen Ramsay, Martin Mueller, and the author) are well acquainted with the limitations of the current application—mainly speed, portability, and scalability—and are able to imagine a demand for a better instance of the software. Such an application, developed under the auspices of the Center for Digital Research in the Humanities, would likely be a publicly available server-based web application. The chief function of this application would be to convert files, and directories of multiple files, so that texts from various encoded collections could interoperate with those from other collections, projects, and institutions. The type of use that was originally envisioned for deeply encoded texts—enormous libraries of rich content that cut across institutional boundaries—is still unrealized.

Conclusion

Nine thousand lines of code written for corpus conversion will never capture the imaginations of any but a small number of individuals with very particular interests, but Mueller reminds us that "sweeping vistas are made possible by metadata gathered, extracted, and processed through tediously explicit routines."[9] While Abbot's work is indeed tediously explicit, it should move scholars closer to such vistas. Abbot's text-processing pipeline permits new and existing collections to be imagined, combined and re-combined as desired. Such combinations are made possible with markup that is much closer to a lingua franca than the differing practices commonly found in local institutions. Because Abbot normalizes texts, common tools for text mining and linguistic analysis ought to be able to work with them. Tool builders will, with Abbot, have a common target format to focus on.

Abbot's modest success is not yet the million books that Crane writes about, but it is a significant achievement and bodes well for the ability of these particular texts to dwell in an intelligent digital library where scholars can effectively use them. It should soon be possible to seamlessly navigate the centuries of English writing that constitute the TCP texts. Or, if not seamlessly, then with fewer seams than before. Emerging digital technologies will continue to necessitate new curatorial activities that involve gathering, interoperating, and effacing boundaries between collections. Libraries will no doubt remain vital participants in collecting and organizing things, but the shape of those things and the curatorial environment where they exist is changing.

[9] Martin Mueller, "Notes towards a Monk User Manual," (2009), accessed June 3, 2011, http://www.monkproject.org/MONK.wiki/attachments/23776/1724.

Bibliography

Besser, Howard. “The past, present, and future of digital libraries.” In *A Companion to Digital Humanities,* edited by S. Schreibman, R. Siemens, and J. Unsworth**.** Oxford: Blackwell, 2004. Accessed June 3, 2011. http://www.digitalhumanities.org/companion.

Bradley, John. “Text tools.” In *A Companion to Digital Humanities,* edited by S. Schreibman, R. Siemens, and J. Unsworth**.** Oxford: Blackwell, 2004. Accessed June 3, 2011. http://www.digitalhumanities.org/companion.

Crane, Gregory. “What Do You Do with a Million Books?” *D-Lib Magazine* 12, no. 3 (2006): http://www.dlib.org/dlib/march06/crane/03crane.html. Accessed June 3, 2011.

Gold, Anna. “Data Curation and Libraries: Short-Term Developments, Long-Term Prospects.” Office of the Dean (Library) (2010). Accessed June 3, 2011. http://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1027&context=lib_dean.

Mueller, Martin. “Notes towards a Monk User Manual.” Accessed June 3, 2011. http://www.monkproject.org/MONK.wiki/attachments/23776/1724.

—. “Thoughts About Corpora Space and Collections Interoperability.” *Project Bamboo*. Mar 01, 2011. Accessed June 9, 2011. https://wiki.projectbamboo.org/display/BTECH/Thoughts+About+Corpora+Space+and+Collections+Interoperability.

Pytlik Zillig, Brian L. “TEI Analytics: converting documents into a TEI format for cross-collection text analysis.” *Literary and Linguistic Computing* 24, no. 2 (2009): 187-192.

Unsworth, John. “Forms of Attention: Digital Humanities Beyond Representation.” Canadian Symposium on Text Analysis, McMaster University, November 19-21, 2004. Accessed June 3, 2011. http://www3.isrl.illinois.edu/~unsworth/FOA/.

Unsworth, John and Martin Mueller. “The MONK Project Final Report.” September 2, 2009: 3.

Source URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago