Journal of the Text Encoding Initiative

Issue 5 | June 2013
TEI Infrastructures

TEI and Project Bamboo

Quinn Dombrowski and Seth Denbo

revues.org

Publisher
TEI Consortium

Electronic version
URL: http://jtei.revues.org/787
DOI: 10.4000/jtei.787
ISSN: 2162-5603

Electronic reference
Quinn Dombrowski and Seth Denbo, « TEI and Project Bamboo », *Journal of the Text Encoding Initiative* [Online], Issue 5 | June 2013, Online since 25 June 2013, connection on 11 March 2017. URL : http://jtei.revues.org/787 ; DOI : 10.4000/jtei.787

TEI and Project Bamboo

Quinn Dombrowski and Seth Denbo

AUTHOR'S NOTE

In December 2012, two months after this article was submitted to the *Journal of the Text Encoding Initiative*, the Mellon Foundation requested that Project Bamboo draw to a close rather than embark on a second phase of technical development. At the point of its conclusion, Bamboo's infrastructure was not in use by any community of scholars, leaving unanswered questions about the consequences of not specifically leveraging TEI markup in the Bamboo Book Model. The "legacy" of Project Bamboo, as captured through code, documentation, and various notes and documents, will persist via links on the projectbamboo.org website, with the goal of informing future cyberinfrastructure efforts.

1. Introduction

1 Project Bamboo, a cyberinfrastructure initiative supported by the Andrew W. Mellon Foundation, takes as its core mission the enhancement of arts and humanities research through the development of shared technology services. Rather than developing new tools for curating or analyzing data, Project Bamboo aims to provide core infrastructure services including identity and access management, collection interoperability, and scholarly data management. The longer-term goal is for many organizations and projects to leverage those services so as to direct their own resources towards innovative tool or collection development. In addition, Bamboo seeks to model a paradigm for tool integration that focuses on tools as discrete services (such as a morphology annotation service and a geoparser service, instead of a web-based environment that does morphological annotation and geoparsing) that can be applied to texts, individually or in combination with other services, to enable complex curatorial and analytical workflows. This paper covers points of intersection between Project Bamboo and TEI over the course

of Bamboo's development, including the participation of the TEI community in the Bamboo Planning Project, a demonstrator project that aimed to develop services that could process TEI texts from a variety of sources while producing a uniform output, and how TEI-encoded collections have influenced Bamboo's work on collection interoperability and the analysis of large corpora.

2 During its planning phase, Project Bamboo brought together over 600 scholars, librarians, and IT professionals from 115 institutions, through a series of workshops that focused on identifying common and unique scholarly practices in the humanities; the technological and social challenges that impede the uptake of digital research methodologies; and how a multi-institutional, interdisciplinary, and interorganizational community and cyberinfrastructure development project can break down those barriers. Throughout Bamboo's "community design" process, numerous scholar and librarian participants emphasized the importance of facilitating the interchange of TEI-encoded documents, but also the inherent challenges that have become evident through previous attempts at aggregating and processing documents from different sources. These opportunities and challenges have influenced decisions ranging from Bamboo's approach to collection interoperability to the choice of tools to partner with when determining the essential components of a tool set for corpus analysis.

2. Bamboo Planning Project

3 The Bamboo Planning Project was centered on a series of eight workshops held over the course of two years (2008–2010). During this period, workshop participants had the opportunity to propose "demonstrator projects"—scholar/technologist partnerships that would test some of the project's assumptions about the feasibility of different approaches to tool and collection interoperability; how to design, implement, and scale scholarly services; and where scholars' technical roadblocks actually lie.

2.1 Workshops

4 The participants in the Bamboo Planning Project workshops shared a common desire for improving the recognition, support, and opportunities for using digital tools and methodologies in humanities research and pedagogy, but represented a wide range of disciplines and technical backgrounds. While some of the faculty participants had only recently been introduced to TEI, numerous active members of the TEI community were influential in the discussions about the role TEI could play in relation to Bamboo, not only as a set of encoding guidelines, but as a community that shares the same core values. As Martin Mueller noted:

> There is a global and tightly knit community there, and a question that arises in the encoding of a French medieval manuscript may find its answer in a practice developed in the encoding of Buddhist manuscripts in Kyoto. To my mind, the TEI community is a remarkable example of a scholarly group that has global breadth, temporal depth, and is united in a common purpose to use technology to help with philological problems of long standing. (Project Bamboo 2009a)

5 John Unsworth warned against the temptation in creating workflows for scholarly analysis of texts to rely on tools that disregard the metadata provided by TEI-encoded texts:

> Do we need structural markup? Does it matter [for] paragraphs? Paragraphs, sentences, verses, lines, are meaningful units of composition -- meaningful units of analysis... This requires structural markup. [It is] also useful if you're going to ask statistical questions [to] differentiate between core intellectual content and ... table of contents, running headers, index, etc. Those words could throw off your stat results in ways that would obscure information. What about [the] word level? [In] tag cloud visualizations of word frequency in a novel, [the] most frequent words are names of characters. So [one has to] ignore proper names to get at the next level of what's going on in the book, but to ignore them you have to identify them -- tag [them] as proper names. (Project Bamboo 2009b)

6 Project co-director David Greenbaum envisioned Bamboo running a platform that could provide large groups of scholars around the world with more reliable access to scholarly services that can perform particular kinds of analysis on TEI texts from diverse fields:

> If you have TEI encoded set of scenarios, you want to take data and derive social relationships out of it. If there's other people who want to do that, it'd be great for you to run something as a service — send data in, get [a] visualization out. We want to run something like that as a service, but that needs to have reliability we don't have right now. [The Bamboo Shared Services Platform] allows us to build up that infrastructure —exposing services in a more robust way. (Project Bamboo 2009b)

7 The sustainability and scalability issues that would be addressed by a shared services platform do place limits on the potential impact of tools developed to process TEI texts. However, one Bamboo demonstrator project shed light on the difficulty of developing tools that would be sufficiently usable across a broad range of texts to even encounter problems with scalability.

2.2 TEI Bibliography Demonstrator

8 In fall 2008, a team at the University of Chicago including Bamboo program staff member Quinn Dombrowski undertook a demonstrator project proposed by Kent Hooper and Rick Peterson for developing an "XSLT web service engine to transform XML-marked up bibliographic entries into HTML."

> This demonstrator project will be used to create a web service that scholars could use to easily input TEI-tagged bibliographical information which has been validated against the TEI data type description (DTD) and transform it using XSLT into a series of HTML (and, if possible, PDF and other transforms) files. These files would be the primary listing (alpha, by author) and sub-listings of the bibliographical data. . . . A sub-listing transform example would be: "Provide a list of all articles, alpha by title of periodical in which they appear, then by chronological order of publication." (Project Bamboo 2008)

9 In theory, the perceived simplicity and standardization of bibliographic markup practices in TEI—in contrast to the more diverse and complex patterns possible for marking up literary texts—would make this an attractive case study for demonstrating the value of shared scholarly services. The Bamboo program staff assumed that in three weeks, a team with Java programming, XSLT, and design skills could develop "an agnostic service implementation that might serve a very specific scholar's need without sacrificing either the agnostic or the specific."[1] In the proposed timeline, one week would be spent finding existing generic XSLT stylesheets for processing TEI bibliographies. The following two weeks would be dedicated to developing a service for processing XSLT, and to developing a web UI where scholars could either upload their TEI and XSLT stylesheets, or choose a

stylesheet provided by the system (such as a generic stylesheet for processing bibliographies).

10 In practice, it took a year to meet the scholarly requirements of the project. The environmental scan failed to yield generic XSLT stylesheets for TEI bibliographies that were compatible with the specific markup conventions used in Hooper's document (for instance, values of the `@type` attribute on `<biblScope>` elements that include `pages`, `volume`, and `issue`). Instead, Dombrowski developed a basic XSLT stylesheet that transformed Hooper's TEI into a recognizably formatted bibliography. This revealed markup inconsistencies within the TEI-encoded data (such as the ordering of authors' first and last names), which Hooper then revised. The revised TEI, in turn, brought to light legitimate variation in the data that the XSLT stylesheets did not correctly account for. In this way, both the data and the stylesheets underwent iterative development for a number of months.

11 During the summer of 2009, Hooper hired Jacob Jett, a graduate student at the University of Illinois School of Library and Information Science, to take the lead in further work, as Project Bamboo demonstrator project work was winding down. One of the most significant issues related to collation for alphabetical listings. The conventions used by scholars of German literature when alphabetically sorting names transliterated from a variety of languages does not align perfectly with any official standard that can be accessed as part of the XSLT processing. In an alphabetically organized bibliography, an item that is not sorted in accordance with the user's expectations may as well not exist, making near-perfect solutions unacceptable. Despite multiple attempts to solve the problem in the code layer of the project (as described in exchanges on the saxon-help email list),[2] because of constraints on time and funding, Hooper added an `@xml:id` attribute to each `<biblStruct>` element in the TEI, indicating the correct sort order. While this approach would be deeply problematic for an expanding bibliography, it was sufficient for making Hooper's bibliography accessible to scholars in its current state.

12 By 2010, the TEI bibliography demonstrator had successfully transformed Hooper's TEI bibliography into HTML web pages that his colleagues could easily consult online. The amount of custom XSLT work needed to achieve even a minimally useful output from Hooper's data made the program staff reconsider how easy it would be to develop agnostic service implementations that directly serve scholars' specific needs, at least in relation to processing TEI.

3. Bamboo Implementation Project, Phase 1

13 The Bamboo implementation project that followed the planning project focused on exposing collections of texts and services for managing and curating textual material within "Work Spaces" (research environments), while developing the identity and access management (IAM) infrastructure that could mediate access to restricted resources. The implementation project also involved a planning process that would lead to the development of a second application, "Corpora Space" (Project Bamboo 2011a). Corpora Space would enable professional humanists and citizen scholars (including undergraduate students) to work on dispersed digital corpora using an integrated set of sophisticated curatorial, analytic, and visualization tools.

14 Bamboo's work on collection interoperability, with the goal of enabling collections from disparate sources to function similarly within a Work or Corpora Space, focused on defining standard methods for making digital content available to web services. This would be achieved by identifying existing protocols, practices, and ontologies, and extending them where needed (Project Bamboo 2010).

3.1 Collection Interoperability

15 For its first phase of technical development, structural and semantic level interoperability was considered out of scope, as explained in the project's proposal to the Mellon Foundation:

> TEI, for example, addresses structure of textual documents in a manner that enables some algorithmic operation across textual collections or corpora. Yet TEI's diverse variants, which have in many cases evolved to address concerns of real import to adopting scholars, gives rise to much frustration among textual researchers in the arts and humanities—and more so when complicated by uneven mapping of semantic meaning to structural markup. Interoperability hampered by differences in point-of-view is even less tractable (is the Royal Pavillion in Brighton, England situated on Le Terrain Crétacé or in the County of Sussex?). (Project Bamboo 2010a)

16 Instead, the Bamboo collections interoperability work centered on providing Work Space and Corpora Space users with access to a predictable set of data, regardless of the source collection, as defined by the Bamboo Book Model (Project Bamboo 2011b). This model requires:

- [A] source repository id and an identifier for the items within that repository.
- Metadata to supply the required item-level fields: title, creator, and date.
- Full text of the book . . . TEI, HTML, or OCR text files. (Project Bamboo 2011b)

and optionally includes:

- Additional item-level metadata: publisher, publication date(s)
- An identifier for another book if this is a version. If this is a volume in a series, an identifier for that series.
- Structural information about the book, divisions (chapters, acts, etc)
- Scanned images of individual pages (Project Bamboo 2011b)

3.2 Corpora Space

17 The first phase of prototyping work for Bamboo's Corpora Space initiative aimed to make content from HathiTrust, the Text Creation Partnership (TCP)'s transcriptions of Early English Books Online (EEBO) and Eighteenth Century Collections Online (ECCO) texts, and the Perseus Digital Library available for curation (such as OCR correction and identifying different texts that are part of the same work) and analysis. In addition, Corpora Space would leverage the metadata provided through text curation processes (such as annotations to indicate linguistic and literary traits) for the purposes of making novel scholarly claims—both about the text itself, and about the scholarly workflows documented by the provenance metadata associated with any curatorial act.

18 As two of the three collections (TCP and Perseus) use TEI encoding for their content, compatibility with TEI texts was a key trait when selecting candidate tools. These included the Abbot processing scripts that convert XML-like text collections into TEI-A[3];

Juxta, which collates multiple witnesses to a single work and includes specific support for a subset of the TEI elements[4]; and TypeWright,[5] an OCR correction environment that produces lightly marked-up TEI (Project Bamboo 2011c).

19 The Corpora Space workshops, which brought together tool developers and representatives of the above-mentioned collections, included discussions of scholarly workflows that could make use of the TEI markup provided by TCP and Perseus. However, to accommodate time and resource constraints, the prototype development work removed the TEI markup provided by TCP and Perseus to bring those texts into alignment with HathiTrust content, rather than attempting to mark up the HathiTrust texts:

> The functionality that we had decided to implement at the workshop was designed to operate on very simple representations of the texts from our three collections. Because these collections used two very different formats—TEI-A in the case of TCP and Perseus and a simple page-based plain text format for Hathi—we had a choice between attempting to add structure to the Hathi texts, in order to bring them closer to TCP and Perseus, or to remove structure from the TCP and Perseus texts. While preparing for the workshop we ran a series of experiments on the Hathi texts that suggested that the latter would be a more practical approach, and during the first sessions of the workshop we decided to use a JSON format that would include some basic metadata about each document as well as two simple representations of its content: a plain-text version for use in analysis, and an HTML version for display in the drill-down view. (Project Bamboo 2011d)

4. Future Directions

20 The planning and prototyping work for Corpora Space brought to light the need for scholarly data management services that can associate provenance metadata with every curatorial action that modifies a corpus, both automated (such as scripts to correct common OCR errors) and manual (such as manual corrections by student assistants, or annotations that form a scholarly argument as part of the development of a digital critical edition). These services make the history of a text more transparent to scholars who can then assess whether a text is in a "good enough" state to be used for meaningful analysis. Bamboo's IAM services will play a key supporting role in the development of these scholarly data management services, ensuring that individuals and tools alike receive proper attribution for their curation work.

21 In addition to extending its core infrastructure to support scholarly data management, Bamboo plans to build out its Collection Interoperability services to enable scholars to submit their corpus annotations back to the repository from which the text originated. At the discretion of the source repository, those annotations could be incorporated into the publicly available edition, increasing the quality and/or the analytical context of the digital texts available to all. The necessity of providing TEI-encoded data back to source repositories that use TEI encoding will provide Project Bamboo with new opportunities to grapple with the challenges of developing a general-purpose system for enriching TEI-encoded documents that minimizes conflicts with existing markup conventions within those documents.

5. Conclusions

22 Over the course of its development, Project Bamboo's work towards a vision of interoperability for the collections and tools that scholars in different disciplines value has provided many opportunities for testing the practical limits of different approaches to analyzing, augmenting, and processing TEI-encoded texts. As John Unsworth stated in regard to TEI, "[i]nterchange has been part of the goal from the beginning" (Project Bamboo 2009b), and Bamboo has served as a laboratory for grappling with issues that arise when attempting different levels of integration between TEI-encoded texts from different collections. The intellectual work that has gone into enriching texts using TEI in ways that open new avenues of scholarly inquiry is too significant to dismiss in the name of easier interoperability through plain text files. Project Bamboo looks forward to the continued participation of the TEI community in our efforts to define and implement approaches to tool and collection interoperability that are both scalable and capable of facilitating new forms of scholarly inquiry.

BIBLIOGRAPHY

Project Bamboo. 2008. "XSLT Original Proposal." Project Bamboo Wiki. Last modified November 10. http://quinndombrowski.com/projects/project-bamboo/wiki/xslt-original-proposal

———. 2009a. "Analyzing Scholarly Narratives." Project Bamboo Wiki. Last modified March 27. http://quinndombrowski.com/projects/project-bamboo/wiki/analyzing-scholarly-narratives.

———. 2009b. "W3- Perspectives." Project Bamboo Wiki. Last modified January 23. http://quinndombrowski.com/projects/project-bamboo/wiki/w3-perspectives.

———. 2009c. "W4 - Action Plans." Project Bamboo Wiki. Last modified April 20. http://quinndombrowski.com/projects/project-bamboo/wiki/w4-action-plans.

———. 2009d. "NYX." Project Bamboo Wiki. Last modified May 13. http://quinndombrowski.com/projects/project-bamboo/wiki/nyx

———. 2010. "Collections Interoperability - BTP Proposal - 6 July 2010." Project Bamboo Wiki. Last modified July 14. http://quinndombrowski.com/projects/project-bamboo/wiki/collections-interoperability-btp-proposal-6-july-2010.

———. 2011a. "Corpora Space - BTP Proposal - 6 July 2010." Project Bamboo Wiki. Last modified March 1. http://quinndombrowski.com/projects/project-bamboo/wiki/corpora-space-btp-proposal-6-july-2010.

———. 2011b. "Book Model (Draft)." Project Bamboo Wiki. Last modified June 1. http://quinndombrowski.com/projects/project-bamboo/wiki/book-model-draft.

———. 2011c. "TypeWright." Project Bamboo Wiki. Last modified December 19. http://quinndombrowski.com/projects/project-bamboo/wiki/typewright.

———. 2011d. "Corpora Camp Lessons Learned Report." Last modified May 17. http://quinndombrowski.com/projects/project-bamboo/wiki/corpora-camp-lessons-learned-report.

NOTES

1. Kaylea Champion, e-mail message to Bamboo Program Staff mailing list, October 29, 2008. See also Project Bamboo 2009d for anticipated timeline.

2. Jacob Jett, "Problem with xsl:sort and German," email to saxon-help@lists.sourceforge.net dated August 10, 2009, with replies dated August 10–17, 2009, http://old.nabble.com/Problem-with-xsl:sort-and-German-td24904567.html.

3. See http://monkproject.org/docs/abbot.html.

4. See http://www.juxtasoftware.org.

5. See http://www.18thconnect.org/typewright/documents.

ABSTRACTS

Project Bamboo, a cyberinfrastructure initiative supported by the Andrew W. Mellon Foundation, takes as its core mission the enhancement of arts and humanities research through the development of shared technology services. Rather than developing new tools for curating or analyzing data, Project Bamboo aims to provide core infrastructure services including identity and access management, collection interoperability, and scholarly data management. The longer-term goal is for many organizations and projects to leverage those services so as to direct their own resources towards innovative tool or collection development. In addition, Bamboo seeks to model a paradigm for tool integration that focuses on tools as discrete services (such as a morphology annotation service and a geoparser service, instead of a web-based environment that does morphological annotation and geoparsing) that can be applied to texts, individually or in combination with other services, to enable complex curatorial and analytical workflows. This paper addresses points of intersection between Project Bamboo and TEI over the course of Bamboo's development, including the role of TEI in Bamboo's ongoing development work. The paper highlights the significant contributions of the TEI community to the early development of the project through active participation in the Bamboo Planning Project. The paper also addresses the influence of TEI on the Bamboo Technology Project's collection interoperability and corpus curation/analysis initiatives, as well as its role in current (as of October 2012) development work.

INDEX

Keywords: cyberinfrastructure, digital infrastructures, Project Bamboo, interoperability, corpus annotation, scholarly editing, tools

AUTHORS

QUINN DOMBROWSKI

Quinn Dombrowski is the community liaison for Project Bamboo at UC Berkeley, and has been deeply involved with the project since 2008. She holds a BA/MA in Slavic Languages and Literatures from the University of Chicago and an MS in Library and Information Science from the University of Illinois. She has worked as the lead developer for numerous digital humanities projects, including DHCommons, Bamboo DiRT, and Bulgarian Dialectology as a Living Tradition.

SETH DENBO

Seth Denbo is the project coordinator for Project Bamboo at Maryland Institute for Technology in the Humanities. He holds a PhD in history from the University of Warwick in the United Kingdom and is a cultural historian of eighteenth-century England. He has worked on projects in digital history and been Research Associate at King's College London where he was involved in strategic planning for a major European digital research infrastructure. He is also a convenor of a seminar in digital history at the Institute for Historical Research in London.