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

history.state.gov: A case study of Digital Humanities in Government

Joseph Wicentowski, Office of the Historian, U.S. Department of State

Abstract

The field of digital humanities is transforming research and teaching inside academia, but it is also making substantial contributions in government. Government agencies, such as the U.S. Department of State's Office of the Historian, are applying technologies and methodologies developed in the field of digital humanities to government data in order to improve government transparency and the delivery of services, while lowering costs and ensuring better long-term data integrity. Focusing on the decision making processes that led the Office of the Historian to adopt specific technologies (namely, the Text Encoding Initiative and the eXist XML database) for its historical publications, this case study offers broad lessons for other government agencies and digital humanities projects and scholars.

Introduction

The field of digital humanities is transforming research and teaching inside academia, but it is also making substantial contributions in government. Government agencies, such as the U.S. Department of State's Office of the Historian, are applying the technologies and methodologies developed in the field of digital humanities to scholarly projects in government in order to improve government transparency and the delivery of services, while lowering costs and ensuring better long-term data integrity. The Office of the Historian's website[1] is built on technologies that largely grew out of digital humanities research: the Text Encoding Initiative (TEI)[2] and the open source eXist XML database.[3] The combination of a rigorous format for encoding data and a robust database for searching and displaying the data has significantly transformed the way the Office publishes its data and has opened up exciting new possibilities for research. Besides the significant impact these technologies have had on the Office of the Historian, this episode offers important lessons for other organizations and researchers. First, it highlights how important it is to select flexible technologies in order to ensure that a project maximize its short and long term potential. Second, it illustrates how these technologies—each cutting edge achievements in their respective domains in both sophistication and simplicity—enable humanities scholars with little or no programming experience to take control of their own data and how it is presented. This paper, a case study in the intersection of the digital humanities and government, offers lessons not only for other government agencies but also for any digital humanities project or researcher.

---

[1] "Office of the Historian." *Office of the Historian, Bureau of Public Affairs, U.S. Department of State*, accessed June 16, 2011, http://history.state.gov.

[2] The Text Encoding Initiative Consortium is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. The Consortium publishes the Text Encoding Initiative Guidelines for Electronic Text Encoding and Interchange, an international and interdisciplinary standard that is widely used by libraries, museums, publishers, and individual scholars to represent all kinds of textual material for online research and teaching. "TEI: About," *TEI Consortium*, accessed June 16, 2011, http://www.tei-c.org/About/.

[3] eXist-db is an open source database management system built using XML technology. See "eXist-db Open Source Native XML Database," accessed June 16, 2011, http://exist-db.org/.

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

The Backdrop: Government Historians Discover Powerful Digital Humanities Technologies

The Office of the Historian, housed in the Bureau of Public Affairs in the U.S. Department of State, researches and publishes the official documentary history of U.S. foreign relations. Staffed by about 40 professional historians, the Office is mandated by Congress to publish the *Foreign Relations of the United States* series, a venerable documentary edition that reveals the decision-making behind the key policies and actions of the U.S. government in its foreign affairs. (The Office also performs policy-related research for Department principals, holds scholarly conferences, and responds to queries from the public—among its various programs.)

In 2007, the Office of the Historian, faced with a growing archive of historical publications as well as an expanding set of programs, launched a website redesign project. The project began with humble goals: to improve the process of publishing voluminous amounts of historical data, and to make it easier for readers to search and find information on the site. In the process of this redesign, the Office discovered a potent set of new technologies that would take the website far beyond its initial goals and would serve the Office in many more ways than were known at the time. These technologies have been under independent development for many years, but are highly complementary and yield impressive results. The Office of the Historian is grateful for the work of the various communities that developed these technologies and has been contributing back to these communities with code, documentation, and training. But perhaps as important as any one technology or set of technologies, the Office of the Historian has learned a number of general lessons that could help guide any digital humanities scholar or project. This paper will cover the most important lessons and will illustrate each with respect to how the Office of the Historian applied each one. This is the paper that we wish we had at the outset of our project.

Lesson 1: First, Know Thy Data

We are fortunate to be living in such dynamic times, with both hardware and software constantly improving by leaps and bounds. But as exciting (or bewildering) as today's offerings might be, the reality is that there is little guarantee that the software we use today will continue to be supported and maintained long into the future. In contrast, the products of our scholarly endeavors must be supported and maintained. A project spanning any reasonable duration (1-2 years) will likely have to migrate from one database or server to another. Given this, it is critical that new projects focus first on their data and select a format carefully. In other words, when you are considering a new research project or a redesign of an existing project, you should start by thinking about what data format best suits your research goals. Only once you have determined the data format should you select the software, and that selection should be based around what best fits your data—what best enables you to encode, analyze, and publish your results. What factors should you consider when selecting a format for data?

- The format should let you capture all aspects of the material you hope to study.
- The format should be archivally sound.
- The format should be software-neutral; avoid a format tied to any one company, product, or closed standard.
- If no such format exists, you can certainly invent your own, but beware the downsides of reinventing the wheel, so consider adapting an extensible format rather than starting from scratch.
- As a user of the format, you become a stakeholder, and you should have a voice and be able to participate in its continued evolution.

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

These considerations led the Office of the Historian to select the Text Encoding Initiative (TEI) as the format for encoding our data, the *Foreign Relations of the United States* series. Less important than our particular choice was how we went about researching and narrowing the field according to the parameters above.

The *Foreign Relations* series is a documentary edition, meaning that it contains primary source documents, annotated by scholars. We examined all of the features of our volumes to establish a baseline for the editorial features we needed. At the very least, we needed a format that could let us represent all of the elements of documentary editing: sections, headings, datelines, footnotes, page numbers, glossaries, indexes, cross references, and tables. This narrowed our choice to formats designed around textual products, including software-based desktop publishing formats (Word, InDesign, OpenOffice, PDF) or software-neutral XML formats like TEI, NLM, and DocBook.

We were keenly aware that we had to retain fidelity to our original print publications, and our data had to survive well into the future. We kept in mind that the *Foreign Relations* series is an important scholarly resource with a back catalog spanning over 450 volumes published over the course of 150 years, with numerous changes in style over time. Over the course of our 10+ years of experience publishing data on the web, our site had already undergone several overhauls, and our materials were in a dizzying array of inconsistent formats: some volumes were published as text files, some as HTML, some as PDF without HTML. The variation was due to changes in our publishers, changes and limitations in our software, and the advent of our born-digital publications alongside our digitized print publications. We hoped to find a single, unified format that could house both our old and new content, while respecting the integrity of the original publications. We hoped this format could stand the test of time: we knew that we might have to move to a new server in the future (easily once every 5 years), and never again wanted to have to reformat or reprocess our data just to migrate to the server du jour. These considerations led us away from formats tied to any one company and toward open, standards-based formats.

In any discussion of archival, "future-proof" formats, PDF (particularly PDF/A) is a natural format to consider, but it suffers from considerable problems from the standpoint of a digital humanities project; a brief digression on these weaknesses will help illustrate the strength of other formats. PDF's strength is its ability to precisely capture page layout, and if preserving precise page layout is paramount, PDF could be a valid choice. However, the disadvantage is that the data is locked up inside the PDF, severely limiting the usefulness of PDFs for digital humanities research. First, tools for searching PDFs are simply rudimentary. Second, web search engines do not descend deeply into PDFs, meaning that they miss much of the content in long PDFs (and other non-HTML formats too); if a search engine cannot index your content, you and your readers will not find it. Third, since PDFs have fixed dimensions and layout, they are impractical to use on devices whose screens do not match these dimensions; the text can only be shrunken or blown up proportionately, a painful proposition on a mobile phone or e-reader. The growth in mobile devices and e-readers means that fixed-layout formats are far from ideal. Fourth, scanned text cannot be searched unless optical character recognition (OCR) is applied, and OCR could not achieve the levels of accuracy we required; relying solely on OCR for deriving machine-readable text means that search results will be inaccurate, users with visual impairment who rely on screen readers will be frustrated or thwarted entirely. (Federal government websites such as the Office of the Historian's are bound to meet requirements for accessibility under the Americans With Disabilities Act and Section 508.)

For the Office of the Historian, whose publications are official government documents used for historical research, these compromises were unacceptable. We needed a format that could be

proofread, precisely annotated, and relied upon for accurate text. We considered encoding our data all as pure HTML, but this, too, had disadvantages; HTML's small range of tags focused too much on presentation and was not expressive enough for the purposes of scholarly annotating.

Thus, we began investigating formats from the XML family, which provided sophisticated means of annotating our data while retaining the ability to be displayed on our website as HTML. Indeed, the major XML formats (NLM, TEI, DocBook) all provide tools for exporting their data into not only HTML but also PDF, MS Word, Open Office, and so on. Moreover, XML is an open standard, non-proprietary, and plain text-based — making it ideal from an archival standpoint. Of all formats, plain text-based formats like XML have the best chance of being readable by computers long into the future. These two features of XML—archivability and transmutability—appealed to us.

As we examined the different XML formats, we became deeply impressed with TEI. TEI allowed us to maintain fidelity to the original source (the "promise" of PDF) while being editable, archivable, and transformable. Fidelity to the features of the original source, such as page breaks and page numbers, was important to us, since scholarly citations point to page numbers. TEI has an explicit facility for encoding an original publication's page breaks and page numbers. It also has a facility for linking the text to the scanned images of the original document, meaning that we could display the original scanned image for any given page of text. Furthermore, since TEI can be reduced to HTML, search engines can index all of the content, and both current devices (web browsers, screen readers) and emerging devices (e-readers) can display the text readily and flexibly.

TEI's appeal did not stop there. The TEI Guidelines seemed to cover every possible type of annotation we would ever need, and more. It contains mechanisms for capturing all of the structural features of our material (e.g. page numbers, footnotes, font styles) as well as semantic features. In other words, rather than using italics to represent a term from a foreign language, TEI encourages you to explicitly annotate the term with the "<foreign>" tag. Similarly, a person can be marked with the "<person>" tag and further identified with a link to a biographical database. In other words, TEI provides the full repertoire of traditional documentary editorial mechanisms (footnotes, brackets, and prose commentary), and adds a rich new set of rigorous techniques for explicitly annotating documents. Since each annotation is explicit and applied directly to the text as XML, we can use the entire universe of XML-aware tools to search and mine the data we created.

TEI also appealed to us because it is a mature, open standard. TEI is maintained by a consortium of scholarly institutions and has open membership. At over 20 years old, TEI has already released its 5th major revision. The entire standard and its guidelines are freely available, and there are numerous resources for learning TEI. The TEI mailing list is also free and open, and we found the community members of mailing list to be an extraordinarily friendly and generous group.

Before we had decided on TEI, we had considered creating our own XML-based format for our publications, since certain features of our publications seemed unique. However, the TEI Guidelines explicitly allows (and even encouraged) customization of the TEI to fit unique project needs. Had we tried to model our own data format, we would have bypassed the 20 years of collective experience crystallized in the TEI Guidelines. And we would not have been able to make use of the toolset created by the TEI community for validating and transforming TEI documents.

Besides allowing customization, the TEI community also invites participation. We knew that committing to a format meant a significant investment of time, money, and energy, and we wanted to be sure that as a stakeholder we would have a voice in the continued support and evolution of the

format. The TEI Consortium's open membership model allows individuals and organizations to join, participate in interest groups, elect council members, and participate in TEI conferences. Membership also conveys benefits such as discounts on essential software tools and training.

In sum, by focusing on data first rather than software or servers, we were able to select an open, archivally sound, non-proprietary format. Other projects and organizations might well settle on any of the other such formats, but TEI is uniquely well suited to digital humanities projects.

Lesson 2: Sophisticated Digital Tools are Increasingly Accessible to Humanities Scholars (Don't Look at the World as if it's Still 1995 or 2005)

The tools for "doing digital humanities" are now accessible to scholars with a humanities background. This is a very new development, and its significance cannot be overstated.

Many people are surprised to learn that the Office of the Historian's website was created and programmed entirely in house by historians—whose only degrees were doctorates in history, not computer science. We completed it and launched ahead of schedule and under budget. Even more surprising, we did not intend to create the website ourselves at the outset. We had originally planned to craft a set of requirements for the website and hire an IT firm to create it. But two factors changed these plans: First, the IT firms we met with had never created a scholarly resource such as the one we had in mind, nor did they have experience with large amounts of XML, nor had they heard of TEI. We began to fear that the product would not meet our aspirations. Second, the federal acquisitions process is very lengthy, and bidding out and awarding contracts can take many months. So we used the time to learn about the software and technology that could be used for a digital humanities project such as ours. Having selected TEI as our format, we looked for tools that could handle large amounts of XML, and we discovered an open source software package called the "eXist native XML database". In contrast to a traditional relational database, which stores data in spreadsheet-like tables, eXist natively and efficiently ingests and searches entire XML documents. Using eXist, we developed a prototype of the website (which we originally intended to provide to the IT firm as the basis for the website once a contract was awarded). The prototype was so promising (and the contracting process was so far from being complete) that the Office decided to develop the website in house. The Office redirected the funds it had originally planned on using for the programmers toward migrating its existing digital publications to TEI.

By storing our TEI publications in eXist and developing our website in XQuery, we were able to streamline our publishing process, improve the experience of reading and using our content, and apply data mining techniques to expose new historical insights. We can upload a huge TEI file containing an entire 1,200-page book, and eXist ingests the file in seconds. Once ingested, the search engine is instantly updated with the new data. Publishing a book on the web, which had taken 3 weeks of painstaking uploading and RSI-inducing typing and clicking, now takes minutes. Whereas our old website had no good way of displaying footnotes, we can now instruct eXist to automatically duplicate the footnotes—displaying one set as inline pop-ups and another set at the bottom of the page. Whereas the table of contents in our print publications only contained chapter titles (printing a list of all of the 300-odd documents in a volume would occupy 50 valuable pages that could be used for more documents), our website generates full document lists on the fly. Whereas our print publications and old website contained a glossary of terms and a biographical list at the beginning of every volume (requiring the reader to flip from a document to those sections to see if a term or person was explained), the new website displays a list of just those terms and people that appear in that document; eXist filters through the tags and displayed the full entry, saving readers the trouble.

Each document and section of the book has its own bookmarkable URL, and we also give readers the option to view the original page images (or scanned documents) for any given document. Since each archival document has a date, we can display a dynamic timeline of a volume, or show all documents across a single presidential administration's 50 volumes, or show the results of a keyword search as a timeline. (eXist's search engine is highly customizable; since eXist indexes the structure of the document, it can, for example, boost the ranking of a search result hit on a document title over a hit on the body text of a document.) The flexibility is tremendous, and there are still many features we hope to add to the site, simply by taking advantage of eXist's built-in functions and capabilities.

In sharing our experience with our peers in government, industry, and academia, we have encountered individuals who had evaluated TEI or eXist in their earlier days and had dismissed these technologies. We emphasize that these technologies have improved considerably and markedly in the past several years. Indeed, the confluence of four key events at the start of our project in 2007-08 radically changed these technologies in ways that made them much more powerful for digital humanities applications such as ours:

First, the 5th major revision of the TEI Guidelines was released in 2007. This revision contained a brand-new framework for customizing the TEI format, a very attractive feature given some unique aspects of our publication. Coincidentally, in October 2007, the TEI Conference was held nearby our Washington D.C. offices at the University of Maryland College Park, so we were able to attend training sessions and ask questions. Everything we learned about TEI confirmed that we had made the right choice for us; its scholarly rigor and flexibility would serve us well. Moreover, in one key conversation, James Cummings of Oxford University Computer Services encouraged us to look into native XML databases as opposed to traditional relational databases. His confidence in XML databases gave us confidence to investigate them further.[4] Similarly, we attended a presentation at the January 2008 American Historical Association meeting by the team from Rotunda, the University of Virginia Press's digital imprint, about their Founding Fathers Papers digital edition. Their website did all that we hoped our site would do, and in conversations after the presentation, Holly Shulman, Mark Saunders, and David Sewell kindly shared with us the elements of their site: they encoded their texts with TEI and stored them in a native XML database. With reinforced confidence in the applicability of TEI and native XML technologies to drive a scholarly resource such as ours, we moved ahead with our prototype using eXist.[5]

Second, at this same time, the eXist development team was just adding new features to the software that drastically increased its speed and reliability. eXist replaced its original indexing (think "search engine") routines with a pluggable indexing platform, allowing new indexes to be added and customized. For example, the eXist team integrated the industry-standard Lucene engine as the basis for eXist's "full text" search capabilities (including wildcard, boolean, and fuzzy options), added "spatial" indexes for geospatial data, and added fast "range" indexes for ordered searching and sorting of date-based or number-based data. We were able to ask the developers about their plans and try out these features on the eXist website. The demonstrations on the eXist homepage convinced us that eXist could rapidly search large amounts of XML and meet the needs of our

[4] A helpful article by Cummings describes his early work with eXist. James Cummings, "Exploring TEI XML Documents with XQuery," *Proceedings of TEI Day in Kyoto* (2006): 99-115, accessed June 16, 2011, http://coe21.zinbun.kyoto-u.ac.jp/tei-day/tei-day2006.html.

[5] A paper by the Rotunda team provided also provided helpful lessons about outsourcing digitization. John Carlson, Mary Ann Lugo, and David Sewell, "Outsourcing Complex Digitization: Lessons Learned," 2007 TEI Annual Meeting, accessed June 16, 2011, http://rotunda.upress.virginia.edu/docs/research/TEI2007Poster.pdf.

website. The fact that we could download the software and easily install it on our PCs and Macs was very convenient and and gave us additional confidence in the software. Since joining the eXist community, we have even contributed code and documentation to the eXist project; we have found that being able to engage with our tools and their developers on this level is a great asset to ambitious digital humanities projects such as ours.

Third, the programming language used by native XML databases like eXist, called XQuery, had just been deemed an official "1.0" standard by the World Wide Web Consortium. This meant that books like Priscilla Walmsley's *XQuery* were published in 2007 and in bookstores.[6] We were able to read this book and begin to teach ourselves XQuery. As a very new programming language, XQuery was designed around doing one thing extremely well: searching and working with XML. XQuery hides much of the traditional complexity of programming languages and lets you focus on searching through your data to answer your research questions. At the same time, it has all of the capabilities of other programming languages, so it is not limited to simple searches. This makes it ideal as a first programming language for those with humanities backgrounds. Within hours you can learn enough to write simple (one line) but powerful queries; but you can also create an entire website with sophisticated analytical tools using just XQuery. Indeed, XQuery could even be your first and last programming language. XQuery's simplicity and accessibility meant that we did not have to learn a host of languages and databases. (Consider that most websites rely on at least two languages: a server language like PHP and a database query language like MySQL. XQuery combines these functions in a single language.) We could focus on mastering XQuery, and with that we could harness the full power of the eXist server to make our TEI data shine.

This combination of TEI, eXist, and XQuery—and the helpful communities supporting each of them—was the recipe for our success. Indeed, our experience in both learning these tools and teaching them to others has reinforced our belief that these tools are essential tools for the digital humanist. (Of course you may find others, depending on your project and its particular needs.)

Thanks to these technologies, the barrier to entry for digital humanities research is now significantly lower than it was even a few years ago. This is not to say that a great deal of learning and effort and is not required. Rather, these technologies give students and scholars of the humanities the tools to conceive and perform sophisticated digital research themselves, without relying on dedicated programmers. Indeed, the distinction between scholar and programmer may be disappearing. Because the encoding guidelines and programming languages are so purpose-driven and domain-specific, humanities scholars can acquire the skills to query and master their data. Of course, collaboration between humanities researchers and dedicated programmers can be extremely fruitful (and in certain domains may always be necessary), but scholarly intentions can be "lost in translation" between the scholar who dreams up ideas and the programmer who implements them. Now, increasingly, scholars can take control of the tools and fully realize their goals without losing anything in translation.

Lesson 3: Plan Now for Inevitable Change Ahead

When we began planning history.state.gov in late 2007, our immediate goal was to find a format and software that would let us launch a website within 18 months. We met that goal, launching in early 2009. But it was not long before we were faced with adding capabilities to the site that we had never

[6] Priscilla Walmsley, *XQuery: Search Across a Variety of XML Data* (Sebastopol, CA: O'Reilly Media, 2007). See also the community-driven "XQuery Wikibook," accessed June 16, 2011, http://en.wikibooks.org/wiki/XQuery.

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

planned. The following three challenges certainly won't be the last we have to face. Such challenges illustrate how the uses of a set of data or a research project can change in significant and unforeseen ways during its lifetime. The best way to prepare for these challenges is to keep the data in flexible formats and containers from the start.

The first such challenge for the Office of the Historian came in the form of the November 2009 Open Government Directive, an Obama administration order requiring that all federal agencies contribute to a central clearinghouse website for government data, data.gov. Each agency had 45 days to contribute three high value datasets, and most had to scramble to meet the deadline. But within days, the Office of the Historian was able to help the State Department meet its obligations. We were able to do this because XML was one of the preferred data.gov formats, and so complying with the order was as simple as writing a new XQuery file for eXist. (A Washington Post article about the quality of various agencies' responses to the Open Government Directive was largely critical, but the State Department was praised for the Office of the Historian's contribution.[7] Also, the Office of the Historian's dataset reached number 3 on the top 5 list of datasets on data.gov, as reported by Wired Magazine.[8]) Indeed, the ability to participate readily in open government data initiatives alone may ensure a bright future for XML databases like eXist in government agencies at the local, state, and national level, both in the U.S. and abroad.

The second challenge came with the emergence and growing popularity of e-readers, such as the Kindle and iPad. The promise and challenge of e-readers was the major topic of discussion at the O'Reilly Tools of Change conference (a publishing industry conference) where the Office of the Historian presented in New York in March 2010. Entire sessions at the conference were devoted to the challenge of adapting existing content to the new "ePub" format used by the iPad and the "Mobipocket" format used by the Kindle. Most publishers struggle to perform this conversion, because the data for each book is stored in a closed format such as InDesign—which, somewhat like PDF, is designed around giving publishers fine-grained control over each page's layout. In contrast, the e-reader formats are much more like webpages, allowing text to flow freely depending upon the size of the screen and font. Publishers were finding that converting data from PDF or InDesign to these e-reader formats required costly, time consuming, and error-prone conversion procedures. In contrast, the Office of the Historian was able to present a single XQuery script for eXist that could convert each of our existing *Foreign Relations* volumes from TEI into the ePub format within minutes.

When we set out to create history.state.gov, neither the Open Government Directive nor data.gov had been conceived, and these e-reader devices and formats did not exist. These capabilities were never part of our requirements. However, because of the non-proprietary and presentation-neutral qualities of TEI because of the flexibility of eXist and XQuery, we were able to transform our data, and provide it to users and readers in new ways.

Besides these two challenges, we have been able to use the combination of eXist, XQuery, and TEI to support research in ways we had never imagined at the outset. Encoding our texts in TEI and

---

[7] Ed O'Keefe, "Info released under Obama transparency order is of little value, critics say," *Washington Post* (2010), accessed June 16, 2011, http://www.washingtonpost.com/wp-dyn/content/article/2010/01/27/AR2010012704589.html.

[8] Eliot Van Buskirk, "Sneak Peek: Obama Administration's Redesigned Data.gov," *Wired Magazine* (2010), accessed June 16, 2011, http://www.wired.com/epicenter/2010/05/sneak-peek-the-obama-administrations-redesigned-datagov/.

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

storing them in a database like eXist turn our books into a database. The power of an encoded book (or corpus of books) as an integrated database goes far beyond keyword search. For example, by tagging people with unique identifiers, we can quickly find every instance of a given person in our volumes, regardless of the spelling of their name or title. Moreover, because we tag each person where they appear in each document, we can analyze where two people co-occur, and how often such pairings occur. These co-occurrences can expose interpersonal connections that would otherwise be impractical to find. We may even augment these findings by using data visualization frameworks to create diagrams of relationships. Similarly, since we did not discard the original back-of-book subject indexes but instead preserved and coded them with TEI, we have assembled an inverted index of sorts: a comprehensive list of subject headings that refer to each document in a volume. We have also examined the extensive cross-references in our footnotes, so that for a given document we can display a list of all other documents that refer to it. The more cross references that point to a given document, the more important it is likely to be. Or, conversely, an instance of a cross reference to a distant volume could expose an under-appreciated research possibility. We have been able to analyze the relationship between the number of footnotes in a volume and how long it took our researchers to complete the volume. The possibilities for mining the structural and semantic features of a TEI-encoded corpus of books stored in an XML database are staggering.

In other words, a digital humanities project may start out with specific goals and requirements, but new challenges and opportunities for using the data will almost certainly arise. The project may prove useful to another group of researchers, or the data may need to be displayed in new ways, or aggregating two datasets together may promise to yield valuable new results. Depending on the choice of technologies, the project may be locked into fulfilling only its original goals, since the cost of adding new capabilities or reformatting data can be so high. As the Office of the Historian has discovered, the digital humanities and computer science communities have developed excellent formats and technologies for maximizing flexibility and minimizing friction when new capabilities are needed.

Conclusion

As the Office of the Historian's case study demonstrates, the tools for digital humanities are reaching new levels of maturity, sophistication, and accessibility. The availability of formats like TEI and databases like eXist (with the elegant, high level programming language XQuery) means that digital humanities scholars—in academia and government—have excellent choices for encoding, analyzing, and publishing their data today, and preserving and adapting their work long into the future.

Bibliography

Cummings, James. "Exploring TEI XML Documents with XQuery." *Proceedings of TEI Day in Kyoto* (2006): 99-115. Accessed June 16, 2011, http://coe21.zinbun.kyoto-u.ac.jp/tei-day/tei-day2006.html.

Carlson, John, Mary Ann Lugo, and David Sewell. "Outsourcing Complex Digitization: Lessons Learned." 2007 TEI Annual Meeting. Accessed June 16, 2011. http://rotunda.upress.virginia.edu/docs/research/TEI2007Poster.pdf.

O'Keefe, Ed. "Info released under Obama transparency order is of little value, critics say." *Washington Post*. January 28, 2010. Accessed June 16, 2011. http://www.washingtonpost.com/wp-dyn/content/article/2010/01/27/AR2010012704589.html.

"Office of the Historian." *Office of the Historian, Bureau of Public Affairs, U.S. Department of State*. Accessed June 16, 2011. http://history.state.gov.

Van Buskirk, Eliot. "Sneak Peek: Obama Administration's Redesigned Data.gov." *Wired Magazine* (2010). Accessed June 16, 2011. http://www.wired.com/epicenter/2010/05/sneak-peek-the-obama-administrations-redesigned-datagov/.

Walmsley, Priscilla. *XQuery: Search Across a Variety of XML Data* (Sebastopol, CA: O'Reilly Media, 2007). Accessed June 16, 2011. http://en.wikibooks.org/wiki/XQuery.

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