Unleashing TEI and Plain Text Data

Open Access Week Call to DH Community

Motivated by a recent mock keynote debate, "A Matter of Scale (http://digitalcommons.unl.edu/englishfacpubs/106/)," presented by Matt Jockers and Julia Flanders as part of the Boston Area Days of Digital Humanities Conference (http://nulab.neu.edu/events/dhdays2013/) and the imperative that librarians involved with many things "digital" learn not only how to build tools, in this case for textual analysis, but leverage existing tools to support teaching and research endeavors rooted in the text. Coming from the tool-building perspective and tradition, I seldom have time to explore existing tools for textual analysis. This is partly because at IU we are so vested in textual markup following the TEI Guidelines for which few external tools exist that act on the markup (thus our focus on building). But as is the case with many academic libraries attempting to balance scale of digital production, we are not always in the position to build boutique interfaces, tools and functions for hand-crafted markup. Further, often early research inquiries can be better defined if not answered by initially playing and experimenting with raw data sets before embarking on markup. Finally, after many years of leading e-text initiatives and championing the TEI, I would love to sit around with folks and compare and contrast, not just the possibilities, but also the outcomes of real research inquiries that formed the basis for many of the TEI collections I am offering up to the community for experimentation. In other words, what can we ascertain without/beyond the markup and can those very queries yield answers regardless of the markup?

The other motivator for this call is two-fold. At IU we've always exposed the TEI/XML, but at the most atomic level. I am exploring workflows moving forward in which we batch not only the TEI but other versions of the data, primarily plain text, for easier harvesting and re-purposing. One reason for doing this – there are many good ones – is that we want to demonstrate to our faculty partners the possibilities of sharing data in this way. The content can and should be analyzed, parsed, and remixed outside of the context of its collection web site for broader impact and exposure. I am hoping, with your help, to figure out how to best push versions of this data into the flow, around a more formal call, initially, to the digital humanities community-at-large as part of Open Access Week 2013 so I can track the various morphings and instantiations of this data to share back with the IU community, especially my faculty partners.

I recently blogged about this very concern on Day of DH 2013: <http://dayofdh2013.matrix.msu.edu/mdalmau/2013/04/08/oh-the-one-fun-thing/>. So this is the first step of a multi-step process that I would like to see culminate in a greater unleashing of XML and plain text data (later summer / early Fall?).

This session is by no means limited to the following e-text data I will provide (data access details forthcoming):

- Indiana Magazine of History (http://dlib.indiana.edu/collections/imh/, one the nation's oldest scholarly historical journal, 1905-2011)
- Victorian Women Writers Project (http://dlib.indiana.edu/collections/vwwp/, 1830-1929)
- Indiana Authors and Their Books (http://www.dlib.indiana.edu/collections/inauthors/, 1850-1929)
- Brevier Legislative Reports (http://www.dlib.indiana.edu/collections/law/brevier/, transcripts from Indiana Legislature, 1858-1887)
- Wright American Fiction (undergoing web site migration, 1851-1875)

Serve up or use up even if snippets of your own data of interest.
Nor is it limited by the following tools I have identified for starters:

- TXM (http://sourceforge.net/projects/txm/, http://wiki.tei-c.org/index.php/TXM) for TEI files; installation required
- PhiloLogic/PhiloMine (http://code.google.com/p/philomine/) also for TEI files but maybe too much overhead to get started; installation required
- Mallet (http://mallet.cs.umass.edu) for topic-modeling; see also recent article by Elijah Meeks and Scott B. Weingart:http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/
- VUE (https://vue.tufts.edu/) for visualization and other analyses; registration and installation required
- Voyant (http://voyant-tools.org) for textual analysis; web-based
- MONK (https://monk.library.illinois.edu/cic/public/) for web-based textual analysis of pre-defined data sets (MONK collections w/include Wright American Fiction or CIC collections)

In fact, it would be best to partner up with folks who are a familiar with a particular tool. Vote for this session and come to this session, claim a tool!.

PS All data will be posted on this public-facing wiki page: https://wiki.dlib.indiana.edu/x/WYK2Hg.

TEI and Plain Text Data from IU Libraries

- https://iu.box.com/s/wn2ef6ne8dnu1wuty9y8: Download TXT and XML/TEI files