

Marrying Literature and History: Web-Based Digital Humanities Portals for Research

Kalev Leetaru
February 18, 2007

At their core, a digital literature and a digital history project both involve the management of digital materials and of making those materials available, either as standalone objects for direct viewing, or as data sources for computation. However, the digital history project will often use its digital objects largely as access surrogates to enable access to them to a global audience, while the digital literature project will often use its digital texts as computational surrogates, using their digital form merely as a method to expose them to machine processing techniques. Within the digital literature community, self-contained suites have been developed such as NITLE's Literary Toolkit,¹ but the text and literary focus of these toolkits mean that content developed for them and research conducted within them may be hard to transfer to other domains. By the same token, the vast amount of historical text offered by digital history projects like Valley of the Shadow are largely static virtual counterparts of their physical brethren, typically without the supporting analysis frameworks of digital literature projects. (While there have been numerous research projects applying various data mining and other technologies to large digital history projects, including a collaboration between NITLE and Valley of the Shadow, they have rarely become permanent staples of the sites – though there are examples, such as Tuft's London collection – http://gis.nitle.org/resources/articles/article_03.htm).

This strong divide between the two disciplines leads to a collection of isolated, siloed projects, each developing their own critical infrastructures for managing the digital material at the heart of their projects. These infrastructures are largely incompatible and often reinvent or duplicate work of each other. In an attempt to begin taking the first steps towards reconciling these differing needs, I-CHASS is developing a set of centralized portal environments that will offer a common platform that may be built upon for individual projects, regardless of discipline. Rather than turning to different systems to meet the access needs of a digital history project or the analytic needs of a digital literature project, this common platform will allow both to be met by a shared infrastructure, eliminating the duplication of core resources, and allowing the disciplines to share each others' developments. Digital objects uploaded to this platform are stored with their associated metadata and access rights set to either make them globally available for browsing and searching (meeting the access needs of digital history projects), or restricted for local analysis (meeting the needs of some digital literature projects that use copyrighted texts). Regardless of access privileges, each digital file uploaded to the system instantly becomes available to the large number of analysis tools built into the platform, and orchestrated by workflow processes. The output of any

¹ <http://web.media.mit.edu/~raphael/papers/IterativeVis/paper.html>

analysis procedure becomes immediately available as its own digital file, which may be subjected to further analytic processes, or made available to others through a browseable/searchable digital library interface. In this way, a digitized version of a manuscript that is uploaded for attribution analysis can also simultaneously be made available to the global community to access, while a historic text uploaded as part of a digital history project could be subjected to attribution analysis later on without any additional work.

As an example of this process, a digital history project digitizes a collection of booklets on the history of a particular organization. These are all uploaded to the system and made available to the public. Several of the booklets are missing their cover and front pages that list the author information, and, in the course of analyzing the content of one, it becomes important to know its author. However, given the sheer number of booklets and authors, it would take a considerable amount of time to examine the records for each author to try and find mention of something that might be the booklet in question. One of the researchers logs into the portal and clicks on the “Analyze a Text” link and selects the booklet as the data file, and “Attribution Analysis” as the analysis tool, and selects a set of the other booklets where authorship information is known as the training set. Another mouseclick later, and the request is dispatched off to a remote supercomputing system, where the data files are pulled from the central repository, processed, and the researcher is emailed when the results are available. The system suggests that the work shares strong similarities with those of Author A, which the historian uses as a starting point and searches that author’s records, and sure enough, finds a reference to what appears to be the booklet, as having been written by that author in 1846. In the traditional disjointed model, this kind of cross-discipline technique usage would be unheard of, as the historian would not likely have expertise in using digital literature attribution systems, and would not have the time or resources to convert all of his materials over just for the purpose of a single attribution check. However, in the platform being developed by I-CHASS model, such crossdiscipline and exploratory interaction is facilitated with point-and-click ease.

Similarly, a digital literature project, having digitized a complete collection of manuscripts, can make that collection available to colleagues around the world with a click of a button, instead of having to maintain a second copy of the material in a separate digital library for download. Additionally, such a collection of manuscripts might be sourced from a large geographic region, and for the purposes of analysis, it might be useful to subdivide them by the region discussed in each. In a traditional digital literature system, this would have to be performed by humans or the data would have to be exported to external GIS tools that have been developed for the digital history field (such as Tuft’s London collection - http://gis.nitle.org/resources/articles/article_03.htm), but in the I-CHASS platform, the researcher could simply log into the portal, select the collection of manuscripts as the data source, “Geocode Documents” as the analysis tool, and the task will be dispatched to an available supercomputer that will then go through each document, identifying any mentions of places or major landmarks, and then assign an approximate geographic location to the document based on these mentions. In this way, a tool developed for the digital histories can be instantly leveraged by a scholar in

the digital literary field, who otherwise would not likely have access to this kind of technology.

Thus, through the production of a single centralized portal environment that provides a common platform for data storage, access, and analysis, I-CHASS is taking strong steps towards beginning to bridge the gap between the needs of digital history and digital literature.