Perseids And Arethusa: Building Tools That Build Digital Humanists

Authors: Bridget Almas, Marie-Claire Beaulieu, Gernot Höflechner

Abstract

Can a methodology which allows students, scholars and teachers to use tools as we build them contribute to the development of those individuals as "digital" humanists? Over the last year as we have been developing the Perseids Platform and the Arethusa Annotation Framework we have been actively engaging the users at every step of the process and releasing features for them to work with well before the tools are fully finished. Despite the challenges this introduces for both the developers and the users, we have found that using tools mid-development opens a window into exactly how the tools let users manipulate and shape the data. The process also allows both students and researchers to understand their role in the creation, curation and annotation of texts through the scientific process of creating and using the data. This process has also exposed the critical role the humanist plays as product designer and tester of the tools we develop to support the research and publication process. Or, as Stephen Ramsay (Ramsay, 2011) says, as "builders and makers."

The Perseids platform is a collaborative editing environment built from a loose coupling of open source tools and services.[1] Two core values in the development of the platform have been (1) to reuse existing tools wherever possible and not reinvent the wheel and (2) to put the data first, focusing on the creation of data that is reusable and can be preserved through the use of stable identifiers for all of our texts, annotations and related objects, and serializing data according to standard formats. While we aim to produce prototype disseminations of the data created on the platform as web-servable digital editions, this focus on the data, and the reuse of existing tools has put production of a seamless user interface and polished presentation of output as a lower priority. We also have adopted an Agile[2] development methodology, which requires us to release new features for use early and often.

The result is that the end users see much of the "inner-wiring" of the tools. For example, in IMGSPECT[3], our image-to-text mapping tool, students identify regions of interest on an image and map them to the text of the transcription. This tool creates stable identifiers (in the form of CITE URNs[4]) for the region of interest and creates a template TEI transcription for the word or characters represented in the image which references the URN of the region of interest in the @facs attribute of the TEI markup element. Although ultimately we plan for automatic population of the underlying TEI XML file for the entire transcription from the image annotations, currently users have to copy the template markup into their file manually. While the value of such extra work is not always obvious, the process affords an opportunity for the students to experience all the steps involved in building a dataset. In particular, the process emphasizes the need to justify editorial choices in transcribing a text and creating data that conforms to the current best practices and can therefore be reused by others. In doing so, students are using new technology while following the most traditional scholarly principles. Indeed, producing good data that is communicable, interoperable, and can form the basis for sound interpretation is the goal of all scholarship. In this way, working with Perseids allows students to put the highest standards of scholarship into practice and to enter the worldwide conversation about the production of knowledge. By exposing the URNs that link the text to the image, we

[1] http://sites.tufts.edu/perseids/
[2] http://en.wikipedia.org/wiki/Agile_software_development
[3] https://github.com/PerseusDL/imgspect
[4] http://www.homermultitext.org/hmt-doc/cite/

allow them to see that what they are doing is producing data that not only justifies their choices in their own edition but also allows this data to be reused by others.

Arethusa[5] is a framework for linguistic annotation and curation that provides a highly configurable, language-independent, extensible infrastructure for close-reading, annotation, curation and exploration of open-access digitized texts. Arethusa is back-end independent, but it has been developed in collaboration with the Perseids project, and integrated with Perseids where it provides an annotation interface for morpho-syntactic analyses and will soon also act as a broker between the Perseids back-end and various other front-end annotating and editing activities, including translation alignments, entity identification and text editing. Many of the design requirements for the Arethusa morphosyntactic annotation environment were directly informed by unanticipated scholarly and pedagogical uses of its precursor tool, the Alpheios Treebank Editor.[6] Alpheios exposed the XML of both the treebank annotations it worked with and the configuration files that defined the tag sets available for use. The ability to work with these XML structures allowed the users to see the direct interplay between the data and the tools used for editing them, pushing the limits of the Alpheios interface. As design on the new Arethusa interface started, we actively engaged these users in testing and designing the new features they had asked for. This is not necessarily a novel approach - in [Thomas and Solomon, 2014] the RoSe project team noted that engaging students in this way greatly improved the usability of their system. However, in addition to the benefits afforded by the tool, we also found that this process developed the humanist users' analytical skills and experiences, forcing them to evaluate questions such as whether or not the features they were requesting merited changes in existing standard linguistic representations of morpho-syntactic annotations, and implications on interoperability. It also required them to consider the division of responsibility between user interface functionality and representation of the data manipulated by the interface. The ability to edit the underlying data directly provoked questions about data integrity, as the morpho-syntactic annotations are just one component of a larger publication which includes TEI XML transcriptions, translation alignments and other annotations.

From a pedagogical standpoint, integrating students into the development efforts has meant that both professors and pupils had to be comfortable with using a technology that is in constant evolution. For this reason, we favor a hands-on approach that focuses on the tasks to be accomplished by the students rather than general instruction about the tools or about digital humanities. In this we agree with [Mahoney and Pierazzo, 2012], who say "What we should be teaching under the umbrella of the 'digital humanities' are not skills—although they too play their part—but new methodologies and new ways of thinking." For example, students in an intermediate Greek class were given step-by-step instructions about how to transcribe a Greek inscription as an xml document rather than a lecture about xml, even though they had never been introduced to it. The instructions gave basic xml markup indications, such as <w>, <l>, and <lb>. The professor and students accomplished each task simultaneously so that the students could refer to the on-screen demonstration as they transcribed their own text. After each step, students saved their work and visualized the concrete result of their edits in the preview screen. They were thus able to measure their progress and to gain an immediate understanding of the effect of their work. This approach has been put into practice when using Perseids in classes such as Classical Mythology, intermediary Greek, medieval Latin, and many others. We have found it to be productive, as it empowers the students to work with the software from the the very beginning and to start to feel comfortable with it quickly. The approach has also offered great opportunities to gather feedback from the students and keep developing the tools in order to make them more intuitive and user-friendly.

[5] http://sosol.perseids.org/tools/arethusa/app/#/
[6] https://github.com/alpheios-project/treebank-editor

Finally, the ability to work directly with the data has led the Perseids and Arethusa users to think differently about the process of publication. Giuseppe Celano, a scholar who was an integral member of the Arethusa development process, has been exploring the possibilities for sharing treebanking analyses in the form of open micro-publications, realizing that "through Perseids, every scholar is provided a free platform allowing a micro-publication which is fully shareable, and so anyone is given the chance to be not only the user but also the contributor of a publication[7]."

Bibliography

Mahony, S. and Pierazzo, E. (2012). "Teaching Skills or Teaching Methodology?" In *Digital Humanities Pedagogy: Practices, Principles and Politics* (2012), pp. 216-225
Open Book Publishers. http://openbookpublishers.com/htmlreader/DHP/chap08.html

Ramsay, Stephen. (2011). "On Building." http://stephenramsay.us/text/2011/01/11/on-building/

Thomas, L. and Solomon, D. (2014). "Active Users: Project Development and Digital Humanities Pedagogy." In *CEA Critic,* Volume 76, Number 2, July 2014, pp. 211-220. DOI: 10.1353/cea.2014.0014

[7] From an email, 9/27/2014