Developing a New Integrated Editing Platform for Source Documents in Classics

Bridget Almas

Marie-Claire Beaulieu

Abstract

The Department of Classics at Tufts University and the Perseus Project have jointly designed and tested an integrated platform on which students can collaboratively transcribe, edit, and translate Latin and Greek texts, creating vetted open source digital editions. This project, while giving students the opportunity to work with original untranslated documents, also contributes to the efforts of the scholarly community worldwide to meet the challenge of publishing large numbers of primary source documents online while preserving high editorial standards. The platform integrates the SoSol software, originally developed to edit papyrological texts, and the CITE architecture, originally developed by the Center for Hellenic Studies to support the Homer Multitext Project. The present paper discusses our objectives in developing our online platform, the scholarly, pedagogical, and technical challenges we faced in the course of our work, and the results we obtained.

Introduction

The Classics scholarly community currently faces a double challenge. First, with a very small group of specialists worldwide, how can we process the vast numbers of source documents available to us? Thousands of ancient Greek and Latin inscriptions and thousands of medieval and later Latin manuscripts are known and published—many under Creative Commons licenses—and more are found every year in the course of archaeological digs and cataloguing efforts in libraries worldwide. Due to the sheer number of these documents, a very small proportion is thoroughly studied, leaving a wealth of information about the ancient and medieval worlds completely in the dark. Indeed, producing accurate editions, translations, and commentaries of such documents with traditional methods requires a high degree of competency in Latin and/or Greek, familiarity with complex editing conventions, and above all, lots of time. Our second challenge is one of pedagogy. Considering the need to protect and conserve our source documents, how can we train students to work with original materials? How can we offer students a hands-on learning experience that allows them to hone their language and research skills by working on texts that are not part of the traditional canon of classical authors?

The development of new digital tools and the pursuit of innovative teaching methods that integrate learning processes with scholarly research, as advocated by Blackwell and Martin (2009, esp. 9-16; see also Bellamy 2012), can help resolve both difficulties. Because digitized documents can be manipulated by students in a virtual online environment, a task that serves as a learning experience in class can also become an opportunity to contribute to the efforts of the scholarly community worldwide. In keeping with these principles, the Tufts Classics Department and the Perseus Project recently undertook to develop an integrated online platform for the collaborative editing of source documents in Classics. Using this platform, students can edit, translate, and write commentaries on ancient and medieval

documents. A built-in review process allows them to collaborate closely with their peers, instructors, and citizen scholars around the world, along the lines of open collaborative editing described by Robinson (2010). The final result of the students' work is a complete vetted scholarly edition which gets published online, producing a tangible learning outcome for the student and adding to the knowledge base of the scholarly community.

Due to the dual nature of our documents as texts and physical objects (especially in the case of inscriptions and manuscripts), we decided to build the platform by integrating software and standards from two existing projects focused on each of these aspects. We chose the SoSOL text-editing platform, which was originally developed to edit texts preserved on papyrus, and the CITE services (Collections, Indexes, and Texts, with Extensions), originally developed by the Homer Multitext Project (HMT), which define (among other things) standards and services for citations and creation of links between texts and images.

Objectives

Several separate but related needs drove our work on integrating the software and services from the above-mentioned projects. Most of our work focuses on the first two of these with a view to supporting the third and fourth goals in subsequent work.

1. To support collaborative work by students, along the model of the HMT project, thus allowing students to conduct substantive linguistic research with a tangible outcome, the publication of a digital edition of their work.
2. To work not only with inscriptions and manuscripts but also with more general textual sources, such as the Greek, Latin, and Arabic collections in the Perseus Digital Library, for which subsets of the TEI Guidelines such as the TEI-Analytics subset (being developed by the Abbott Project[1]) are more suitable.
3. To support work on a growing range of historical sources in multiple formats and languages. These include more than 1,200 medieval manuscripts for which the Walters Art Gallery (250 MSS) and the Swiss e-codices project (900 MSS) have published high resolution scans under a Creative Commons license.
4. To support a large and international community of digital editors, including students, advanced researchers, and citizen scholars. The spring 2012 user base for the Perseus Digital Library exceeded 300,000 users, with approximately 10% (30,000) working directly with Greek and Latin sources. The 90-9-1 rule[2] predicts that 9% of an online community will contribute occasionally and 1% will make the majority of new contributions. This would imply active communities of 30,000 for Perseus as a whole and 3,000 for the Greek and Latin collections.

SoSOL and CITE

[1] http://monkproject.org/MONK.wiki/Abbot%20and%20TEI-Analytics%20texts.html
[2] http://en.wikipedia.org/wiki/1%25_rule_%28Internet_culture%29

SoSOL and CITE are two separate frameworks, developed independently, for working with digital representations of ancient sources. They each approach the problem set from different directions, resulting in little overlap between what the two have to offer, and a great deal of potential for integration.

The SoSOL platform was designed to provide support for the collaborative editing of the different types of XML data being integrated from multiple sources under the Papyri.info platform (see Sosin 2010). Supported data types include transcriptions, translations, metadata, commentary and bibliographies, each adhering to the TEI/EpiDoc schema (see Bodard 2008), but with different conventions and restrictions applied. Publications made up of one or more of these data types are guided through an editing lifecycle by a workflow engine built on top of a git version control repository. Support for a simple role-based user model is provided, leveraging the OpenID specification by delegating authentication to Social Identity Providers. Editors can search a catalog of pre-established publication identifiers to select items to edit, or can create their own publication. Each user works on the publications in their own clone of the underlying git source repository until they are ready to submit a revised publication for approval, at which point their submissions are passed to an editorial board for review, and can either be returned to the editor for further work and corrections, or finalized and updated in the master branch of the repository.

The CITE (Collections, Indexes, and Texts, with Extensions) architecture provides a framework both for digitizing textual sources and for creating mappings between those sources and their digital facsimiles on the level of the citation (see Smith 2009; see also Blackwell and Smith 2009). It consists of technology-independent but machine-actionable Uniform Resource Notation (URN) schemas for canonical citation,Application Programming Interfaces (APIs) for network services that identify and retrieve objects identified by canonical URN, and implementations of those APIs on a variety of platforms. This architecture was developed by the Center for Hellenic Studies (CHS) in part to enable the work of the Homer Multitext Project (HMT). In developing the architecture, the CHS team intended to support a wide range of ancient source material in addition to manuscripts, and with the CTS (Canonical Text Services) URN syntax we are able to express in a single identifier both the position of the work in a hierarchy reflective of the Functional Requirements For Bibliographic Records (FRBR), and the position of a node or continuous range of nodes within a work. For example, the URN urn:cts:greekLit:tlg0012.tlg001.perseus-grc1:1.1#μῆνιν[1] represents the first instance of the word token ‘μῆνιν’ on Book 1, Line 1 of Perseus’ first Greek edition of Homer’s Iliad. The CITE URN syntax applies the same theory to non-document objects, and supports a citation scheme for images, enabling, in a single identifier, identification of both the image itself and specific coordinates on that image.[3]

Integration of SoSOL and CITE

In keeping with agile development methodologies[4], we have taken an iterative approach to the integration. We started with the following code bases:

[3] See http://www.homermultitext.org/hmt-doc/cite/index.html for a complete explanation of the CITE Architecture.

[4] http://en.wikipedia.org/wiki/Agile_software_development

- a forked clone of the git repository of the SoSOL platform's JRuby code base
- the Groovy/Java/Google App Engine reference implementation of CTS and CITE APIs from the HMT Project

The first deliverable to create was a prototype implementation that re-used the existing SoSOL code for EpiDoc transcriptions almost in its entirety by sub-classing it and changing only the structure of the document identifiers to correspond more closely to the CTS URN syntax. We also substituted a CTS text inventory for the Papyri.info catalog. Coding the prototype gave us a means to explore the design of the SoSOL platform code and assess its viability for reuse.

The next step was to analyze whether we could also extend this work to support the larger Perseus corpus, which will be using the TEI-Analytics XML schema instead of EpiDoc, and for which we will need to support collaborative editing not only at the level of the entire text but also at the level of a citation or passage. The latter leverages the CTS API heavily. However, as CTS is a read-only API, we needed to develop a set of parallel write/update/ delete functionality that could be used to update and create new editions of CTS-compatible texts. To experiment with this, we augmented the XQuery based implementation of the CTS APIs from the Alpheios project[5], which was written by the developer working on this project. We also coded prototypes of additional extensions to the SoSOL code to work with texts and passages that use the TEI-A XML schema rather than EpiDoc, and to present a passage selection interface.

Completing these two steps gave us confidence that the integration was in fact viable, and an NEH Digital Humanities Start-Up Grant has enabled us to move the work beyond the prototype stage to actual implementation.

Overcoming Integration Challenges

Through the work on the prototype, we were able to identify some key interoperability challenges for the two platforms.

For SoSOL, this has centered on the identification and isolation of the papyri-specific assumptions of the platform. These have primarily been in the following areas:

- identifier scheme
- cataloging system
- style sheets for display
- differing concepts of what makes up a "Publication"

For CTS, the primary integration challenge so far has been in augmenting it with a compatible Create/Update/Delete system. The challenges also include the need to identify or define a canonical citation scheme for the inscriptions, although this is not specifically a platform integration issue but instead a more general one related to the creation of digital editions.

[5] http://alpheios.net

The first deliverable of the implementation stage of the project was to integrate the prototype code with the master branch of the SoSOL repository that had continued to evolve during our protoyping efforts, and with which our forked clone was now out of sync. Through this process, we were able to both take advantage of various enhancements made to the SoSOL code in the interim and to reduce the amount of changes necessary to the main code base to support the new data and identifier types. This process also required some significant rewriting of the prototype code, but this was not surprising as the creation of production quality code was not the main objective of the prototype. We are now working on a branch of the master SoSOL repository, rather than a fork, and expect to be able to integrate the branched code back into the master branch fairly soon.[6]

Testing the Workflow

Professor Beaulieu's project to engage students in editing ancient funerary inscriptions has provided an excellent opportunity to explore this work. The job of mapping her collection of images to transcriptions in order to produce digital editions leveraging those mappings parallels in many ways the work of the HMT project and the current best practices in the field of epigraphy (see Cayless et al. 2009) and is a good fit for the CITE services and APIs. In addition, the TEI-based EpiDoc XML standard to be used for digitizing the inscriptions is already well-supported by the SoSOL platform. We were able to reuse large parts of the XML validation and display code from the papyri publication support on SoSOL while focusing on the addition of support for the CTS identifiers. This incremental approach allows us to lay the groundwork for the eventual support of the full collection of Perseus texts integration while at the same time producing something more immediately applicable and available for use by a smaller, controlled community of student assistants who can effectively serve as Beta testers for the platform.

Once the initial implementation was complete, the next step was to deploy the SoSOL and CTS services on a Perseus server with a functioning interface that Professor Beaulieu and her assistants could use to select an inscription upon which to work (see Fig. 1), and then enter the XML for the transcription (see. Fig. 2), translation (see Fig. 3a and b), and commentary (see Fig. 4). This deliverable has been fulfilled and they have been able to complete the creation of a digital transcription and translation of the Nedymos epigram through the SoSOL interface.

Reacting to Change

The iterative approach to development has enabled us to react more quickly to changing circumstances in the landscape of related tools and standards, in particular as applicable to two major aspects of the design plan: integration with image tools and the data model for the mappings between images and text.

Initially we had planned to integrate the SoSOL editor with the Java based ImageJ tool to provide an interface for selecting coordinates on an image and creating CITE URNs that represented those coordinates. The ImageJ tool, however, had to be installed locally on the user's workstation and also required additional calculations to produce the coordinates in the format required by the CITE URN syntax. In this syntax, a component of

[6] See http://git-scm.com/documentation for information on working with git repositories.

the URN identifies the region of interest as specified by four values: the x and y distance from a fixed point on the image as the starting point of the region, along with the width and height of the region. Since the initial design plan was made, however, the HMT project developed a superior Image Citation tool for working with the images, which, being web-based, was more easily integrated with the SoSOL editor, and had the additional advantage of being able to calculate the CITE URNs for the coordinates automatically from the user's selection of a region of interest on the image with their mouse. We were able to fairly quickly integrate this Image Citation tool into the SoSOL interface, and it can now be used from within this interface to select a region of interest on an image and create a CITE URN for that selection when editing or viewing the transcription. The integration currently leverages the "facs" attribute on the text element of the EpiDoc schema to identify the image to display for a given text. In future iterations we will add the ability to select from a list of multiple images and more fully integrate the Image Citation tool so that the CITE URNs can be immediately saved in annotations.

The second major change is in the data model to be used to store the mappings between the regions of interest on the image and the word tokens in the text transcriptions. Originally we planned to store these in flat table indices, but the expanding adoption of the Open Annotation Core (OAC) Data Model specification for annotations has led us to change course on that part of the design. The OAC Data Model leverages the concept of an RDF triple to create associations between related resources and annotations and enables you to define a set of connected resources: one or more annotation "targets" and an annotation "body". The annotation conveys that "the body is somehow about the target" and also allows for inclusion of related provenance metadata, semantic tagging, etc. (see Clark 2012) Representing the mappings between regions of interest on inscription images and transcriptions of the words in the inscription as OAC annotations provides us an with an implementation-independent way to store this data and facilitates its eventual reuse. Figure 5 provides an example of one of the mappings between a region of interest on an image of the Nedymos sarcophagus and the transcription of the epigram found there. In this example, the target of the annotation is the region of interest on the image as expressed by a CITE URN (at a resolvable URI location): `http://data.perseus.org/inscriptions/urn:cite:epifacs:epifacsimg.DSC03651:0.15,0.0844,0.2917,0.0844` and the body of the annotation is a CTS URN (also at a resolvable URI location) pointing to the transcribed word that is located at that point on the image, in this case the first instance of the word "Σκῆνος" on line 1 of the epigram: `http://data.perseus.org/inscriptions/urn:cts:greekEpi:igvii.2543-2545.perseus-grc1:2543.1:Σκῆνος[1]`. In the final digital edition display, these mappings provide the ability to double-click on a word in the edited text and pull up the image of that word on the document (see Fig. 6). The simple relational table data expected by the Groovy libraries used for the original display prototype can be easily created from the OAC data, but we can now also experiment with simpler code based on direct XSLT transformations of the OAC data itself, simplifying the ultimate deployment architecture (see Blackwell 2012).

Adding support to SoSOL for the OAC data model also sets the stage for using SoSOL to manage the editing workflow for a wide variety of different annotation types, including syntactic and lexical data as well as more general notes and commentary. In fact, our initial work on adding OAC support to SoSOL has recently been directed towards managing annotations identifying text reuse in Athenaeus' *"Banquet of the Sophists"* at a

hackathon sponsored by the University of Leipzig and the German Archaeological Institute.[7]

Evaluation and Next Steps

Deploying and using the SoSOL interface for this inscription has enabled us to better understand the actual workflow we will need to support for the work on the inscriptions, and has also uncovered some differences between this workflow and the one currently supported by the SoSOL platform for papyrological work. Among other things, we have identified the need to make some decisions about how we want to handle the commentary and bibliography for the inscriptions, and we have also recognized the need for some design changes to the interface introduced by the CTS approach of keeping the translations in separate documents from the source editions. These changes will be included in the next iteration, during which we will also work on completing support for storing image to text mappings as OAC annotations and continue to move forward with the support for TEI-Analytics and citation-based editing that will be required for the larger Perseus corpus. We also plan to integrate the SoSOL platform with external annotation editors, such as the Alpheios Treebank and Translation Alignment editors, to support the creation of additional lexical and grammatical stand-off annotations, and also plan to explore integration options for additional external tools, such as spell checkers, which can be used to facilitate the transcription process (see for instance the new tool for automated incomplete word suggestions being developed at the University of Leipzig: Büchler, Kruse, and Eckart 2012).

Having used these tools to produce the XML and image mapping data for the Nedymos inscription, we are now also able to begin scoping the requirements for the eventual display of the digital edition. We have used the Groovy based reference implementation of a facsimile browser from the HMT project and the Alpheios browser plugins to experiment with the options and to produce screenshots through which we are able to review and discuss the requirements in a concrete way. In the next iteration we will decide upon an implementation approach for the display code and for supporting automatic integration of the display and editing environments.

From a scholarly standpoint, the work accomplished so far has established that our method allows the production of high-quality editions of source documents which respect the best practices in the field. In particular, adherence to the EpiDoc standard ensures consistency in the textual encoding and the compatibility of our editions across platforms. Furthermore, the SoSOL software includes a function for the production of an apparatus criticus for each document which we intend to adapt for our specific purposes. This will allow us to document our editorial choices as well as list choices made by other editors in addition to any other relevant information which influences the reading of the text. As described by Monella (2008), the flexibility of the digital medium allows us to present details of textual variance much better than the traditional printed apparatus criticus. Finally, and most importantly, our editions give the audience access not only to an edited product (i.e., the editor's choices and understanding of the document) but also to the original document itself, with direct links between an edited text and its image on the stone or manuscript. Thus, when consulting our editions and translations, our audience can evaluate their accuracy at a glance.

From a pedagogical standpoint, we see a bright future for this method. It allows us to expose students to original, untranslated texts at every step in their training. For instance, a

[7] http://www.e-humanities.net/events/athenaeus-hackathon.html

student in an introductory Latin class can edit and translate simple, one-line inscriptions or basic religious texts such as a Book of Hours, while an advanced student can tackle longer and more complex texts such as family encomia on sarcophagi or legal texts. In this manner, we offer students a motivating learning experience in which they can produce original work at every stage in their training and help the scholarly community to process the vast amounts of virtually untouched source documents available.

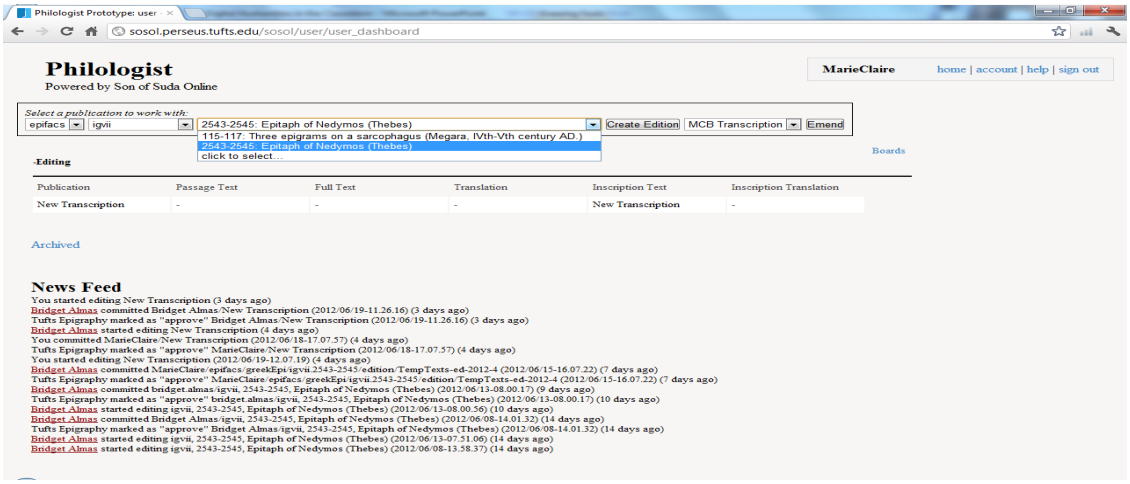

Fig. 1 Text selection

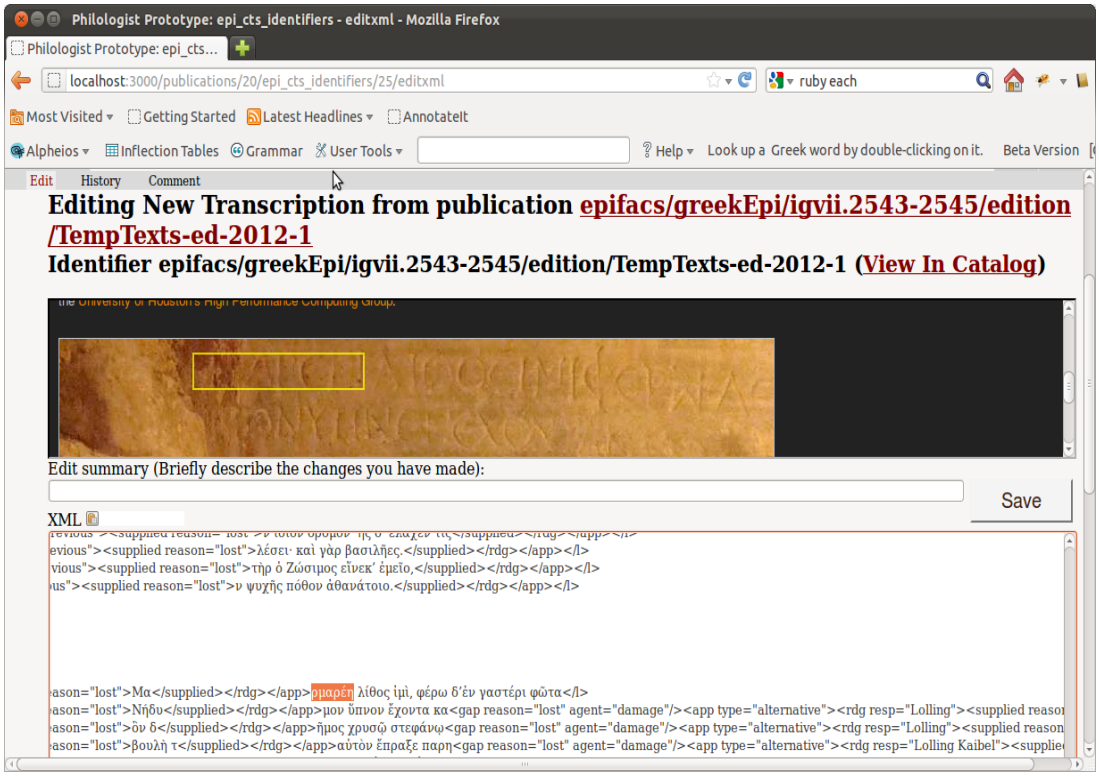

Fig. 2 Transcription and text mapping

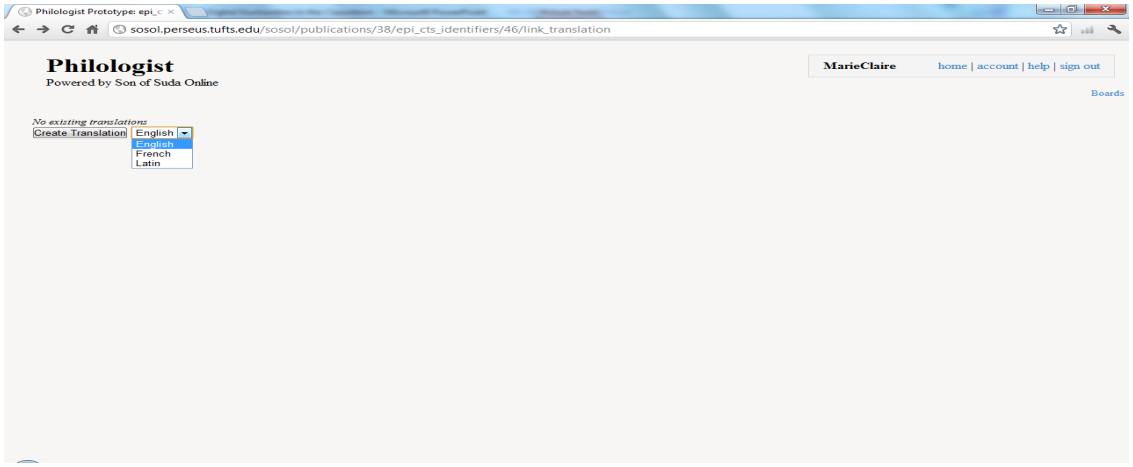

Fig. 3a Selection of a language for the translation

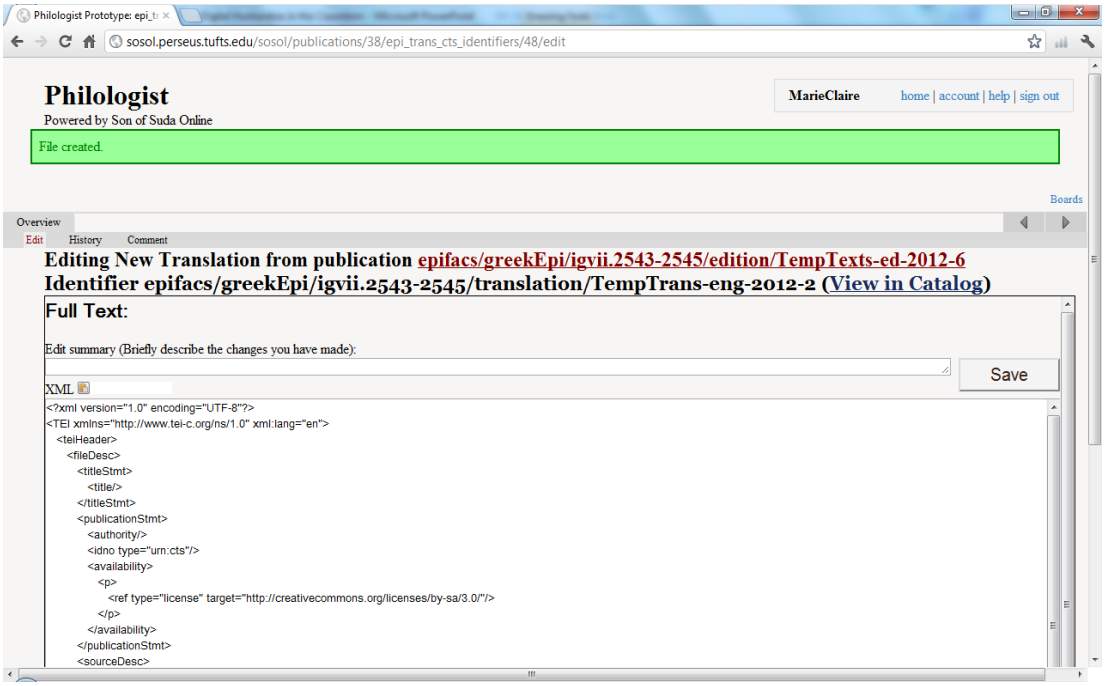

Fig. 3b Entering the translation text in xml

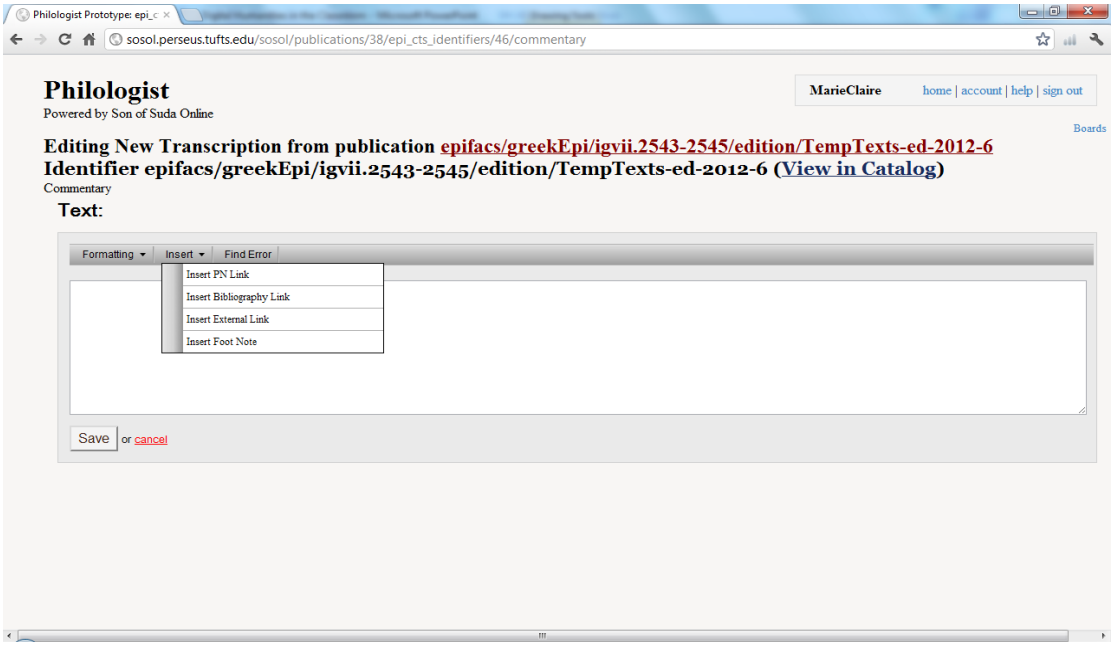

Fig. 4 Entering a commentary

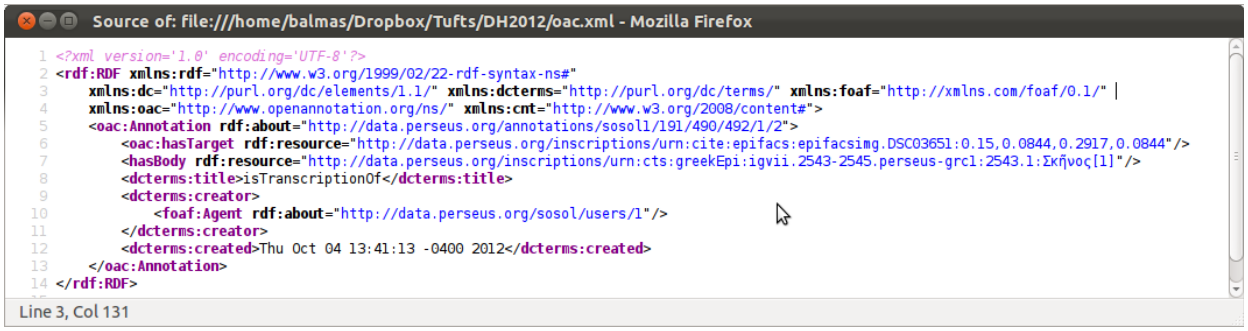

Fig. 5 Mapping of CITE urn of image coordinates to CTS urn of transcription expressed as an OAC triple.

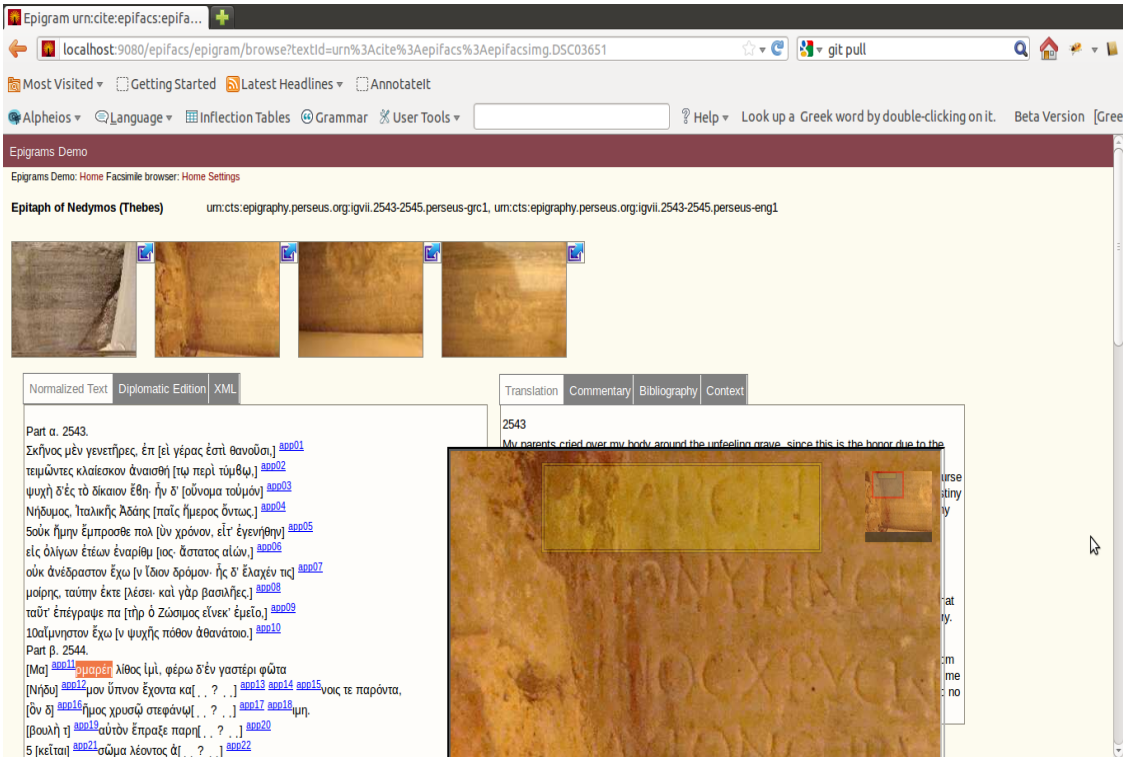

Fig. 6 Bringing up the image of a word by double-clicking

Funding

This work was supported by a Digital Humanities Start-Up Grant from the National Endowment for the Humanities [grant HD-51548-12]. Any views, findings, conclusions, or recommendations expressed in this article do not necessarily represent those of the National Endowment for the Humanities.

This work was also supported by a Tufts Innovates Grant from the Office of the Provost at Tufts University, and by a National Leadership Grant for Libraries from the Institute of Museum and Library Services.

References

Bellamy, Craig (2012). The Sound of Many Hands Clapping: Teaching the Digital Humanities through Virtual Research Environment (VREs), *Digital Humanities Quarterly*, 6 (1).
http://www.digitalhumanities.org/dhq/vol/6/2/000119/000119.html

Blackwell, Christopher and Martin, Thomas (2009). Technology, Collaboration, and Undergraduate Research. *Digital Humanities Quarterly,* 3 (1).
http://digitalhumanities.org/dhq/vol/3/1/000024/000024.html#N10159

Blackwell, Christopher and Smith, David Neel. Homer Multitext – Nine Year Update. *Digital Humanities 2009 Conference Abstracts*, (June 2009): 6-8.
http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf

Blackwell, Christopher (2012). HTML CTS Kit. *The Homer Multitext Blog.* July 18, 2012.
http://homermultitext.blogspot.com/2012/07/html-cts-kit-abstract-announcing-for.html

Bodard, Gabriel (2008). The *Inscriptions of Aphrodisias* as Electronic Publication: A User's
Perspective and a Proposed Paradigm. *Digital Medievalist*, 4.
http://www.digitalmedievalist.org/journal/4/bodard/

Büchler, Marco , Kruse, Sebastian and Eckart, Thomas (2012). *Bringing Modern Spell Checking Approaches to Ancient Texts - Automated Suggestions for Incomplete Words: Proceedings of Digital Humanities 2012*, Hamburg, Germany, 2012, pp. 137-138.

Cayless, Hugh, Charlotte Roueché, Tom Elliott, and Gabriel Bodard (2009). Epigraphy in 2017. *Digital Humanities Quarterly*, 3 (1).
http://www.digitalhumanities.org/dhq/vol/3/1/000030.html

Clark, Tom, Tim Cole, Jane Hunter, Neil Fraistat (2012). W3C Open Annotation Core Data Model. *Community Draft*, 09 May 2012.
http://www.openannotation.org/spec/core/

Monella, Paolo (2008). Towards a Digital Model to Edit the Different Paratextuality Levels within a Textual Tradition. *Digital Medievalist*.
http://www.digitalmedievalist.org/journal/4/monella/

Robinson, Peter (2010). Editing Without Walls. *Literature Compass*, 7: 57-61.
http://dx.doi.org/10.1111/j.1741-4113.2009.00676.x

Smith, Neel (2009). Citation in Classical Studies. *Digital Humanities Quarterly*, 3 (1).
http://digitalhumanities.org/dhq/vol/3/1/000028/000028.html

Sosin, Joshua (2010). Digital Papyrology. *Congress of the International Association of Papyrologists*, (August 2010), Geneva, Switzerland.
http://www.stoa.org/archives/1263