Reusing modern tools and techniques to reproduce and research ancient texts

Anna Jordanous, Centre for e-Research, King's College London, UK [anna.jordanous@kcl.ac.uk],
Alan Stanley, University of Prince Edward Island, Canada [astanley@upei.ca]
Charlotte Tupman, Dept. of Digital Humanities, King's College London, UK [charlotte.tupman@kcl.ac.uk]

The Sharing Ancient Wisdoms project (SAWS)[1] is establishing a research workflow of editing, linking and publishing semantically-enhanced TEI/XML-based digital editions of ancient manuscripts that contain wise sayings. Of particular research interest is the ability to investigate links between these manuscripts. Digital edition creation, storage, annotation and, most recently, the adding of RDF and Linked Data are hot topics in digital humanities research, making SAWS workflows and supporting technologies highly transferable.

Updating and maintaining published files, currently stored in a web filestore, has become troublesome. Additionally, the lengthy, involved procedures of transcribing and editing manuscripts have caused delays. More intuitive interfaces for editing and annotating files would be helpful. Scholars studying manuscripts in right-to-left (RTL) languages have also reported difficulties working with standard XML editors. In this work we explore if SAWS's research workflows can be made more efficient and intuitive through the use of repository tools developed within Islandora[2] for the Editing Modernism in Canada (EMiC) research project.[3]

Background: The SAWS use case

The Sharing Ancient Wisdoms (SAWS) project analyses the tradition of *gnomologia*: collections of wise sayings and philosophical instructions that appear in ancient manuscripts. Scholarly interest focuses on semantic links within and between parts of these collections of sayings. Collections (or individual sayings) would be selected from earlier manuscripts and copied into a new manuscript, often being reorganised, translated, reordered, modified or reattributed during the copying process. The manuscripts collectively form a complex network of interrelated texts, which when analysed can reveal much about the dynamics of the cultures that created and used them. SAWS enables investigation of the relationships between specific sayings, tracing the links through different textual variants and languages using RDF to link sections of text to other relevant sections of text via a subject-predicate-object relationship defined as part of an ontology.[4]

Research outputs for SAWS are digital editions of the manuscripts being studied which publish both the manuscripts and the semantic relationships within the manuscripts, for more useful and information-rich digital editions than in current digital editions. Currently SAWS scholars produce digital editions by manually transcribing the original documents and encoding them in TEI/XML using a custom schema. All files are stored in a traditional web filestore and displayed in a Javascript application. Semantic annotations are extracted using an XSLT and stored in a RDF triplestore, to be shared, reasoned with and visualised.

SAWS is not currently using a repository, but its requirements to organise and store documents are becoming more difficult to manage with traditional web filestores, particularly as more documents are submitted for publication within SAWS. An additional issue is that the transcription and editing of XML files in an editor is a time-consuming and painstaking process. As this task is often undertaken by scholars without a technical background, potential barriers arise hindering the adoption of digital research tools by less digitally-enumerate scholars. A particular issue has arisen when working with manuscripts languages which read Right-To-Left (RTL), e.g. Arabic; technical difficulties were encountered when editing TEI files containing RTL text and Left-To-Right (LTR) XML tags. SAWS are interested in whether tools recently developed by Islandora for the Editing Modernism in Canada project (see next section) can help solve these

[1] http://www.ancientwisdoms.ac.uk. Last accessed March 2013.
[2] http://islandora.ca/ . Last accessed March 2013.
[3] http://editingmodernism.ca/ . Last accessed March 2013.
[4] Anna Jordanous, K. Faith Lawrence, Mark Hedges, and Charlotte Tupman. Exploring manuscripts: sharing ancient wisdoms across the semantic web. 2nd Int. Conf. on Web Intelligence, Mining and Semantics (WIMS '12), Craiova, Romania. 2012.

issues and simplify SAWS workflows to enable for less-technical scholars to use digital research tools.

Background: Islandora & Digital Humanities research

The Islandora open source repository management software assists users in creating, accessing and managing collections of documents, without requiring users to be highly technically skilled. More intuitive, GUI-driven interactions with Fedora repositories are implemented through PHP using a Drupal front end.

Islandora is sponsored by EMiC to develop a suite of applications for managing and critically analysing Canadian modernism. Within Islandora, the Digital Humanities (DH) Solution Pack provides a WYSIWIG online interface to create, edit and annotate TEI documents within a repository-based environment, and to simplify the addition of semantic links. This Solution Pack employs the CWRCWriter[5] and SharedCanvas.[6] The Islandora Critical Editions module exposes a GUI allowing the addition and viewing of RDF entities and TEI tags. No knowledge of XML is required. Entities tie textual offsets to objects from authority lists, user-entered notes, external links, or date ranges through RDF. Image annotations are in OAC RDF.[7]

Modern methods for modern materials, applied to research on ancient documents

We are currently exploring how SAWS can repurpose the Islandora/EMiC research tools and workflows; our talk will demonstrate the processes and code used. We hope to reproduce SAWS research results and outputs in a more intuitive and effective environment. The SAWS and EMiC projects share many crossovers in terms of research goals for studying texts using TEI-based digital editions with semantic enhancements. It is not necessarily true that the two research workflows will necessarily align perfectly; there are various differences in the research goals, especially considering cultural contrasts between ancient and modern times. EMiC focuses on the study of modern typewritten (OCR-friendly) texts in modern languages, as opposed to the handwritten (possibly damaged) manuscripts studied in SAWS, in ancient languages.

Initial explorations have however proven promising, especially in terms of making it simpler for scholars to produce digital editions of the manuscripts and encode their scholarly knowledge in the digital editions. We continue to pursue these explorations more deeply until June 2013. Observations made so far include:

- Creation, updating and managing of files becomes much simpler via the Islandora repository software, compared to SAWS' current use of a web filestore.
- We have received positive initial feedback from SAWS scholars regarding improved ease of use in editing files, adding RDF links and viewing the files during the editing process. Production of Linked-Data-enhanced TEI documents can be facilitated using the Digital Humanities Solution Pack within Islandora. The repository tools make this process easier than the current SAWS approach.
- OCR within Islandora was hoped to be very helpful in making easier the painstaking task of transcription of manuscripts. However this OCR has proven very tricky with ancient documents, due to manuscript image quality, variability in scribes' handwriting, missing parts of the document, etc.

Concluding remarks

This SAWS/Islandora collaboration investigates the enhancement of TEI-encoded documents and a more user-friendly environment for editing, managing and linking texts, through the reuse of Islandora repository-based tools developed for the EMiC project. The outcomes should apply across a wide variety of textual research projects. We hope that our experiments and subsequent documentation will encourage future reuse of the EMiC/Islandora tools and the SAWS TEI-and-RDF workflows, within research that produces semantically enhanced digital editions of textual material. We expect these repository-based tools to prove especially useful amongst Digital Humanists wishing to work with non-technical scholars.

---

[5] http://www.cwrc.ca/ . Last accessed March 2013.

[6] Sanderson, R. Albritton, B. Schwemmer, R. Van de Sompel, H. "SharedCanvas: A Collaborative Model for Medieval Manuscript Layout Dissemination". *Procs of 11th ACM/IEEE Joint Conference on Digital Libraries*, Ottawa, Canada, June 2011.

[7] http://www.openannotation.org/ Last accessed March 2013.