Semantic Cartography: Using RDF/OWL to Build Adaptable Tools for Text Exploration

Ashton, Andrew
Andrew_Ashton@brown.edu
Center for Digital Scholarship, Brown University, USA

Texts encoded using the Text Encoding Initiative Guidelines (TEI) are ideally suited to close examination using a variety of digital methodologies and tools. However, because the TEI is a broad set of guidelines rather than a single schema, and because encoding practices and standards vary widely between collections, programmatic interchange among projects can be difficult. Software that operates effectively across TEI collections requires a mechanism for normalizing data - a mechanism that, ideally, preserves the essence of the source encoding. Brown University's Center for Digital Scholarship will address this issue as one part of a project, funded by a National Endowment for the Humanities Digital Humanities Start-Up Grant, to develop tools for analyzing TEI-encoded texts using the SEASR environment (National Center for Supercomputing Applications).

SEASR is a scholarly software framework that allows developers to create small software modules that perform discrete, often quite simple, analytical processing of data. These modules can be chained together and rearranged to create complex and nuanced analyses. Both the individual modules (or components, in SEASR's parlance) and the chains (or flows) running within a SEASR server environment can be made available as web services, providing for a seamless link between text collections and the many visualizations, analysis tools, and APIs available on the web. For literary scholars using TEI, this approach offers innumerable possibilities for harnessing the semantic richness encoded in the digital text, provided that there is a technological mechanism for negotiating the variety of encoding standards and practices typical of TEI–based projects.

The central thrust of Brown's effort is to develop a set of SEASR components that exploit the semantic detail available in TEI-encoded texts. These components will allow users to submit texts to a SEASR service and get back a result derived from the analytical flow that they have constructed. Results could include a data set describing morphosyntactic features of a text, a visualization of personal relationships or geographic references, a simple breakdown of textual features as they change over time within a specific genre, etc. SEASR flows can also be used to transform parts of TEI documents into more generic formats in order to use data from digital texts with web APIs, such as Google Maps. Because these components deal with semantic concepts rather than raw, predictable data, they require a mechanism to map concepts to their representations in the encoded texts. Furthermore, in order for these modules to be applicable to a variety of research collections, they must be able to adapt to different manifestations of semantically similar information. Users need the freedom to tell the software about the ways in which data with semantic meaning, such as relationships or sentiment, are encoded in their texts. To do this, we will build a semantic map – a document that describes a number of relationships between 1) software (in this case, SEASR components), 2) ontologies, and 3) encoded texts (see Figure 1).

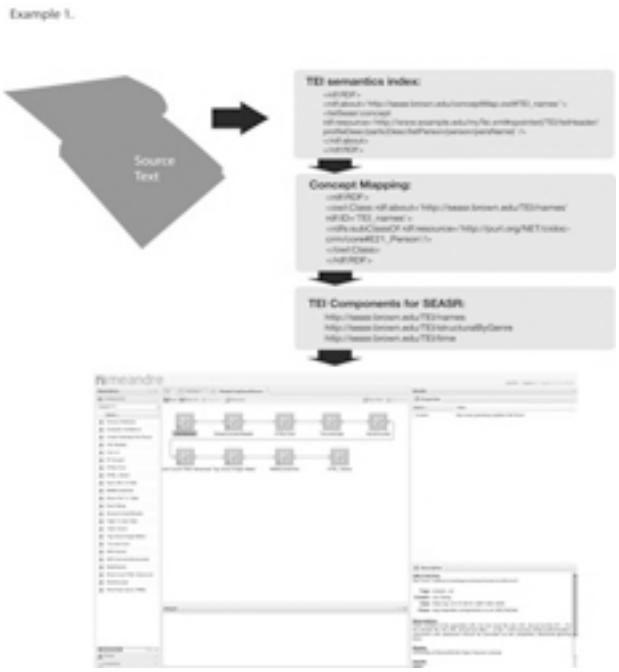

Figure 1

The semantic map is modeled on the approaches of previous projects working with semantically rich TEI collections, most notably the work of the Henry III Fine Rolls Project, based at King's College London. The creators of that project developed a method for using RDF/OWL to model the complex interpersonal and spatial relationships represented in their TEI documents (Vieira and Ciula). They use semantic-web ontologies, such as CIDOC-CRM and SKOS, to define relationships between TEI nodes. Vieira and Ciula offer examples in which TEI nodes that describe individuals are related to other nodes that describe professions, using the CIDOC-CRM ontology to codify the relationship. (Vieira and Ciula, 5) This approach serves well as a model for mapping some of the more nebulous concepts of interest to the SEASR components to the variable encoding of those concepts in TEI.

The links between the encoded texts and the SEASR components are forged using two techniques that exploit the Linked Data concepts around which SEASR - and many other emerging software tools - are designed (Berners-Lee). The first is a RDF/OWL dictionary that defines the ontology of the semantic concepts at work in the TEI component suite. RDF/OWL allows us to make assertions about the concepts associated with any resource identified by a URI (Uniform Resource Identifier). Within SEASR, any component or flow is addressable via a URI, making it the potential subject of a RDF/OWL expression. For example, a component that extracts personal name references from a text can be defined in an RDF/OWL expression as having an association with the concept of a personal name, as defined by any number of ontologies. The second part of the semantic map is an editable configuration document that uses the XPointer syntax to identify the fragments of a particular TEI collection that correspond to the semantic definitions expressed in the RDF/OWL dictionary. In this example, collections that encode names in several different ways can specify, via XPointer, the expected locations of relevant data. When the SEASR server begins an analysis, data is retrieved from the TEI collection using the parameters defined in the semantic map, and is then passed along to other components for further analysis and output.

In addition to the semantic map and the set of related analytical components, the planned TEI suite for SEASR includes components to ease the retrieval and validation of locally defined data as they are pulled into analytical flows. One such component examines which of the TEI-specific components in a flow have definitions in the RDF/OWL dictionary. The result is passed to another component, which uses Schematron to verify that the locations and relationships expressed in the semantic map are indeed present in the data being received for analysis. A successful response signals SEASR to proceed with the analysis, while an unsuccessful one returns information to the user about which parts of the text failed to validate.

Our approach is different from that of other projects, such as the MONK Project, which have also wrestled with the inevitable variability in encoded collections (MONK Project). For MONK, this issue was especially prominent as the goal of the project was to build tools that could combine data from diverse collections and analyze them as if they were a uniform corpus. This meant handling not only TEI of various flavors, but other types of XML and SGML documents as well. To solve this problem, MONK investigators developed TEI Analytics, a generalized subset of TEI P5 features designed "to exploit common denominators in these texts while at the same time adding new markup for data structures useful in common analytical tasks" (Pytlik-Zillig, 2009). TEI-A

enables developers to combine different text collections to allow large-scale analysis by systems such as MONK. As a solution to handling centralized, large-scale data analysis, TEI-A is an invaluable achievement in light of the maturation of mass-digitization efforts such as Google Books and HathiTrust. However, our goal in creating SEASR components is fundamentally different than that of MONK, and thus warrants a different approach. Our SEASR tools are designed to be shared among institutions but to be used differently by each, as a part of a web interface for a particular collection. Furthermore, the granularity of the concepts of interest to the TEI components for SEASR makes it infeasible that we could easily map such encoding to a common format, such as TEI-A. Hence, the semantic map – an index that acts as a small-scale interpreter between local collections and the abstract semantic notions marked-up within them.

In developing TEI tools for SEASR, we address several issues of immediate interest to scholars and tools-developers in the Digital Humanities. It is certainly useful to create text analysis tools for this new software environment. But of broader interest to scholars and users of digital text collections are the semantic mechanisms that permit interplay between community-based tools and their own collections. Several issues require close scrutiny as this model develops: with careful forethought, we need to ensure that our tools are viable outside of the SEASR framework; we need to consider whether the XPointer syntax has a future in the evolving Semantic Web ecology, and likewise consider how our curated scholarly collections can interact more seamlessly with that environment. Ultimately, the tools that we develop will be available in a public repository for institutions experimenting with SEASR.

Funding: This work was supported by the National Endowment for the Humanities Digital Humanities Start Up Grants program.

References

Berners-Lee, T.. *Linked Data – Design Issues*. http://www.w3.org/DesignIssues/LinkedData.html (accessed 14 November, 2009).

MONK Project. http://www.monkproject.org (accessed 14 November, 2009).

National Center for Supercomputing Applications (NCSA). SEASR: Software Environment for the Advancement of Scholarly Research. http://www.seasr.org (accessed 14 November, 2009).

Pytlik-Zillig, B. (2009). 'TEI Analytics: converting documents into a TEI format for cross-collection text analysis'. *Literary and Linguistic Computing*. **24(2)**: 187-192. doi:10.1093/llc/fqp005.

Vieira, J.M. and Ciula, A. (2007). 'Implementing an RDF/OWL Ontology on Henry the III Fine Rolls'. *OWLED 2007*. Innsbruck, Austria, June 2007. http://www.webont.org/owled/2007/PapersPDF/submission_6.pdf.