Discursive Metadata and Controlled Vocabularies

Mylonas, Elli
elli_mylonas@brown.edu
Brown University, USA

Wendts, Heidi
heidi_wendt@brown.edu
Brown University, USA

Bodel, John
john_bodel@brown.edu
Brown University, USA

While formulating an Epidoc compliant template for the encoding of ancient inscriptions, it became apparent that it was necessary to accommodate discursive information about various characteristics of an inscription as metadata in the header of a document, and to specify the same characteristics using a controlled vocabulary, to facilitate searching, sorting and indexing. The msDesc features of the TEI guidelines do not actually allow this type of encoding to occur in several crucial places. However, it is possible to achieve both goals by repurposing, and perhaps straining the usage of some TEI features. We will describe the problem and our solution in more detail, in order to document one project's solution to a common problem, but also to suggest that the TEI Guidelines might be modified to allow this as a more normal use.

Epidoc, a TEI P5 schema that has been developed for epigraphical and papyrological materials is widely used for encoding classical and other western inscriptions. Historically, there have been two parallel and converging ways to encode this type of documentary evidence. The first treats the transcription of the text together with descriptive information about the support, context, decoration and history as content, the way it might be if it were published in a book, and enclosing it all within the <text> element. In this type of encoding the TEI header information is brief, serving to document the source publication, and not the inscription. The primary example of this type of encoding is InsAph, which originated as the digital version of a print volume, and represented the publication of record for its inscriptions. The second approach treats the text of an inscription as content, and places contextual information such as the description of the surface the inscription was written on, its date, format and origin as structured metadata in the TEI header. US Epigraphy, which originated as an aggregation of inscriptions, most of which had already been published, is an example of this approach. These approaches have different advantages: the first results in a more readable and more nuanced description of the inscription. The second, in which the placement of information is more predictable and controlled, allows better processing, searching and indexing.

The US Epigraphy project records Graeco-Roman inscriptions that are known to be in United States collections so that they may be located and studied or used in teaching. As such, the metadata that allows the inscription to be searched and sorted by its characteristics is of paramount importance. The project is developing its corpus using an iterative process, by which inscriptions are first recorded as an ID number with bibliographic citations, then metadata and images are added. Transcription and more detailed descriptions, a necessarily slower process requiring more epigraphical expertise, are added as a third step. This progression ensures that the corpus is as complete as possible, and that information is added in a sequence that provides as much information as possible about as large a number of inscriptions as possible.

US Epigraphy, following the TEI P5 version of the Epidoc schema and encoding practices, relies heavily on the TEI header and uses the <msDesc> component of the header to record metadata about an inscription.

In <msDesc> there are elements to indicate the genre to which the inscription text belongs (<msItem class="xx">), the type of support on which it is inscribed (<objectDesc form="xx">) and the material of which it is made (<supportDesc material="xx">). These three elements are used to indicate parallel types of information, but unfortunately, they don't exhibit parallel behaviors.

<msItem> has an attribute to indicate text genre and it can accommodate more discursive

detail in a child <p>. The attribute, @class, is a specialized attribute of type “data.code” that allows the msItem to point to a controlled vocabulary of text genres. This is handled through a complex mechanism as follows: the text genres are listed using a <taxonomy> element in <profileDesc>. <textClass>, also part of <profileDesc> then points to a genre in the taxonomy, and msItem/@class in turn, points to <textClass>. This is complicated, but it allows a controlled, and less precise value to co-exist with a more nuanced but less processable description of the text genre. Also, crucially, it maintains the controlled list of genres in the document, and not in the schema.

Conversely, <objectDesc> and <supportDesc> have specialized attributes @form and @material whose values belong to the class “data.enumerated,” forcing their values to be maintained in the schema. This is undesirable, as it means that an encoder, or a content specialist would have to modify the schema in order to change a controlled vocabulary. Changing the values in an enumerated attribute also means that it will no longer be possible to validate different epigraphical projects with the same schema, even though their document structures are fundamentally the same.

The ideal solution is to be able to maintain taxonomies within the document, and refer to values within them using an attribute such as @ana, whose value belongs to class “data.pointer.” <taxonomy> provides a powerful and appropriate classification structure, but in the guidelines it is defined as containing only information on text genres, and forming part of the <msItem><textClass> construct. @ana can point to interpretive elements such as <interp> and <fs>, but not <category>, which is the constituent part of <taxonomy>.

Currently, it is possible to create several taxonomies, and to access them using the xi.include mechanism, so that all files and all encoders are using the same controlled vocabularies at all times, and updates are immediate. It is also possible to point to elements in the taxonomies from <objectDesc> and <supportDesc> using an @ana attribute, since @ana is globally available, and points to a valid URI. However, although this validates, it isn’t semantically correct according to the TEI guidelines. A more satisfying solution is to redefine specialized attributes like @support and @material to behave like @ana, and be able to point to controlled vocabularies such a those contained within <taxonomy>.

It is important, when encoding highly structured but also potentially idiosyncratic materials, like inscriptions or papyri, to be able to use both controlled and full-text descriptions. This should be enabled by the markup, but should also be encouraged as good encoding practice. It is also expedient and easier to avoid errors for encoders and programmers to have parallel structures describing similar types of information.

As corpora like US Epigraphy, InsAph, DDDP and similar collections become more concerned with how they will be mined and processed, and are no longer content with creating digital facsimiles to facilitate access, this type of information management is becoming more important. This poster has focused on a few elements and their accompanying attributes. They are not the only places where this problem arises, however. The solution that is presented here is by no means an ideal one. Indeed, it is only permissible insofar as it results in valid TEI documents. There are several other possible approaches. The best solution will be one that results in a set of best practices that can be re-used in other, similar situations.

References

Epidoc. http://epidoc.sourceforge.net/ (accessed 11/2/2009).

InsAph. http://www.insaph.kcl.ac.uk/index.html (accessed 11/2/2009).

TEI Guidelines. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index-toc.html (accessed 11/2/2009).

(11/2/2009). *US Epigraphy*. http://usepigraphy.brown.edu.