Modeling Scholarship

Julia Flanders

Workshop on Knowledge Organization and Data Modeling in the Humanities, March 15-17, 2012

Our discussion so far has inventoried a number of different kinds of digital objects, considered as targets of modeling practice:

- Simple objects, modeled through "surrogates"
- Informational relations, modeled through links, RDF
- Complex objects that are in effect a surrogate plus metadata

I would like to focus here on another kind of modeling: representations of **intellectual systems.** Thinking back to Allen's talk yesterday, I think what I mean by "intellectual" may be extending what he means by "intentional": in other words, systems in which the system itself may be an object of scrutiny while in use, or in which (as Kari put it) the noise we make while trying to communicate signal is itself of interest to us.

The TEI is an interesting example of such a system and it is particularly interesting because it offers a formal way of modeling types of information that are pervasively present in scholarship, and may be given a great deal of careful attention that results in detailed formal description, but have not typically formalized as data in scholarly practice. In other words, they have been formalized for use by humans, but not by computational processes.

What I'd like to do in this presentation is to take a closer look at the TEI as an "intellectual" modeling system, and consider how it represents these complex vectors of information, and how this information might be used more effectively in digital scholarship.

Let's step back and ask first: What are we modeling when we encode a TEI document? [slide 1]

1. Through our use of markup (on its own, without regard for the schema) we may be modeling a document or data source: selecting pieces of it for representation, describing those pieces using a formal language that is internally consistent and that allows us to associate a semantics with the markers we use
2. And, in the case of the TEI, the information we are modeling here in the XML data may include not only the document or data source itself, but also a self-conscious representation of our transcriptional, editorial, and

interpretative activities with respect to that document, undertaken as part of the creation of the digital object. It may also include relational information connecting this document to others, or to other data sources.

3. By using a markup system that has reference to a schema, we are also modeling the "type" or "genre" of documents (and we have to acknowledge that both of these terms deserve scrutiny on their own that there isn't time to give here): in other words, we are locating this document within a system that identifies some documents as being "the documents we are modeling here" (and by implication others that fall outside that category); another way to put this is to say that we are modeling this document **as an instance of a genre**, whether or not in some larger sense it "really is" a member of that genre; we are assimilating it to the genre, claiming ownership of it in that respect. Broadly speaking, we might term this the "document ecology", the set of characteristics that constitute the commonalities among this set of documents or statements.
4. In this way we are also modeling our own intentions with respect to that document (both on its own and as a member of a collection, whether implied or actual): by making a claim that this document belongs in this class of documents, we also say that we will treat it as such (we will process it as such, we will read it as such, we will interpret it as such) and these statements also affect the information that will be made visible: the kinds of expressiveness the document can retain within the communicative system we are setting up for it.
5. And as we generate successive versions of such models, we are also making claims with respect to that document projected through time.
6. And, at the same time, these statements always implicitly, and sometimes explicitly, have meaning in relation to other projects' statements about and treatment of their own documents: the use of the TEI itself expresses an intention that the encoding of the document be intelligible within a larger system of meaning defined by the TEI, and the specifics of that use (what elements are used and not used, and in what ways) also situates the encoding within a spectrum of use (is this, comparatively speaking, a very detailed encoding or a very impoverished encoding?). We might call this the "social ecology" of our documents": in TEI this is accomplished through the ODD.

The ways in which markup models documents and information about documents is familiar and in the interests of time I'm going to set it aside. I'd like to look more closely at the schema and the ODD.

As Wendell Piez and Alan Liu have shown in detail, historically schemas as constraint systems arise in process, as a result of the need to regulate the manufacture of pieces of a work process independently and formally (rather than by seeing whether the work process itself results in a functional outcome). So instead of waiting to assemble the lawnmower to find out whether it will work, we test each piece against a gauge; we can thus discover flaws and inefficiencies in the work process. **Schemas help us regulate process**. Hence schemas also model a diachronic information space: any given schema models a stage in a process (even if that process has only one stage), and the entire process can be modeled as a series of schemas. [slide 2]

It follows from this that the schema, considered from this perspective, also models the expertise or intention being exercised at that stage in the process. For example: in a work flow for a journal article that begins with authorial encoding, followed by a layer of editorial encoding, the latter process might involve a schema with elements for consistent keywording of topics, process metadata, etc. The schema that regulates the final publication would need to enforce the presence of publication metadata.

Having made this suggestion, it's interesting to note that in the TEI (for historical reasons) schemas appear to possess a kind of timelessness and agentlessness, because the TEI is framed as if it represented a set of convictions about documents (rather than a set of functional or processing requirements). For many individual scholarly users of the TEI (as distinct from large-scale projects focused on production), schemas aren't in fact conceptualized as part of a **work system**. This may perhaps be because the academic setting of that use tends to obfuscate the dimension of work in favor of a more timeless view in which the model of the document represents intellectual convictions above all. Time, in the TEI, is perceived as changes in those convictions, developments in our research trajectory, at both the individual level (an improving understanding of my texts) and at the level of the TEI itself, which offers a steady narrative of improvement (refinement, "new features"), undergirded by technological progressivism: the march from P3 to P6.

Let's turn now to the ODD system and the ODD customization file. The ODD system itself is distinctive, and in some ways entirely unique, as a way of defining a markup language. [slide 3] Any given ODD customization file, taken singly, models a single set of choices about document constraint, aimed at expressing a set of decisions about the modeling of individual documents, and at creating convergence in the modeling of a single set of documents. [slide 4] Any given customization file also models the delta, the vector of difference, between our local situation and “TEI Central,” the set of defaults instantiated in the

unmodified TEI. [slide 5] Multiple ODD files, taken together, model divergences between multiple data sets: these might be data sets from different projects, or [slide 6] they might be multiple stages in the development of a single data set: stages in a work flow, or stages in the development of the project's thinking about how to model the data. As a result, the ODD also reinvigorates our ability to use the schema not as a set of timeless convictions but as a set of functional constraints that operate within a work flow or the developmental narrative **of the project** (rather than of the TEI).

In the modeling of divergences here, we can also see another important but underexplored relationship being modeled, namely that of debate and dissent. The ODD models debate within the TEI community about what features are fundamental to our understanding of text (i.e. the diversity of aims and methods that produces the complexity of the TEI in the first place). It also models changes in the terms of that debate over time (registered in changes to the TEI schema overall), and also dissent on the part of any specific individual or project from any specific modeling decision the TEI community has agreed on (e.g. whether a specific element may go in a specific place). Tthis debate and dissent is modeled with considerable explicitness in the ODD customization, since the customization file records what elements and attributes are included and excluded, what changes to classes are made, and what controlled vocabularies and new or renamed elements have been created.

We thus have here a scene of considerable texture and complexity, considered as either a cross-sectional snapshot, or as a temporal sequence. Looking at it as a cross-section, we can see debate at a given moment; uncertainty, hypothesis, speculation; choices, alternatives; intentions (measured against actions). Looking at it as a sequence, we can see work process, developmental process, history of debate, history of actions. In a sense, what we are seeing here is the emergence of a set of tools for modeling something like historiography: a representation of how theory and practice change over time.

What can we learn from the TEI's example here?

The TEI's potential to serve as such a tool set has arisen, first, by virtue of its situation at the center of a very complex modeling problem (humanities textual data). But it has also arisen because by its nature the TEI is designed to handle not only the modeling of that data, but also the markers of transcriptional, editorial, and interpretative self-awareness: the non-transparency of the modeling process is itself part of what is being modeled (and again, I would claim that this is a distinctive feature of humanities data modeling). And third, it has arisen through the TEI’s response to scholarly pressure (i.e. pressure from its scholarly users) to provide ever more nuanced ways to capture these contours of

scholarly responsibility. In other words, the TEI reflects the ongoing debates within the digital humanities about how the digital medium itself can serve as a vector for scholarly ideas and scholarly work.

The proposed new genetic module in the TEI is a good example of this: many of the elements it adds are expressly designed to support the conduct of debates about editorial strategy and interpretation, and Elena Pierazzo may be able to tell us more about this.

There are a few questions I would like to raise in closing, which I feel still need to be answered:

- Is there an advantage to modeling such an intricately connected field of information within a single representational system? or are there parts of this information that would be better factored out and handled separately? (Is the TEI like a gigantic lintball here, or a naturally cohering system of information?)
- And ultimately I'd like to consider whether this kind of complex and layered modeling might also suggest methods we could use in other digital humanities contexts.