Theoretical Issues in Text Encoding: A Critical Review

1. Introduction

Text encoding has long had an important role in the humanities computing community. Initially this importance was simply based on the need to develop systems for representing culturally significant texts in a format that would allow them to be analyzed by computer. However, as these systems for encoding text needed to be as subtle and sophisticated as both the structure of the texts themselves and the hypotheses and theories that were being computationally tested, it is not surprising that decisions about text encoding sometimes seemed to reflect not just incidental variations in technology and logistics but fundamental differences in method and approach to the textual material.

A number of specific issues and topics have arisen in the last twenty years or so and can now be identified as familiar longstanding debates within the humanities computing community. Most of these are fairly robust in that they have no widely accepted resolution and are regularly reiterated and refined. We believe that these are foundational topics that reveal much about both text encoding and humanities computing, not merely distractions from the real work of humanities computing. Discussions range across journals, email lists, conference panels, project documentation, and technical reports, so a systematic survey of these topics will be of value to the humanities computing community.

We have begun a project to survey and critically review the work that has been done on these topics. The project is in its earliest stages and our objective in this poster presentation is to invite participation from the humanities computing community. At this point we are particularly interested in suggestions of new topics that we may have missed, important distinctions within the topics listed, and relevant work we might overlook when writing the full review.

In what follows the reader will notice that we have avoided mentioning even the most obvious researchers, positions, and publications with respect to any particular topic. Balance, fairness, and completeness are essential to a project of this sort, and we are not yet prepared, even if we had the space, to present a balanced selection of researchers and references for any of our topics. We will soon, however, be posting drafts of sections, complete with references and researchers, in an online working version of this survey, at http://eprg.isrl.uiuc.edu/markuptheory/ .

2. Interpretive nature of markup

Is some markup `interpretative'? If so, what sort of markup is interpretive, in what sense is it interpretative, and what does that mean for encoding practice? This is perhaps the classic example of a recognized theoretical topic in humanities text encoding. Some encoding theorists claim, for instance, that the markup that predominates in the TEI is inappropriately interpretative, compromising the relevance and value of that encoding system for humanists. Others argue that all markup is interpretative, but this is not a problem: markup allows scholars to express an interpretation of a text, advancing theories, and exposing their interpretations to criticism. Various other positions have also been taken.

Although there are a few journal articles that take up this issue in passing and a panel discussion on the topic was held at ALLC/ACH 1996, many of the influential analyses and exchanges are in the `gray literature' of humanities computing, including committee reports within the Text Encoding Initiative and several sustained discussions on Humanist.

3. Hierarchical nature of text

SGML/XML is a grammar that yields a tree without cycles or overlaps. While some writers have explicitly argued that text itself has a hierarchical structure, a number of difficult cases suggest that the content objects of texts do not form a hierarchy. The early influential discussion of this topic was ‘SGML-Based Markup for Literary Texts’ , by David Barnard et al., in *Computers and the Humanities*, 22 (1988). Since then quite a number of encoding theorists have made important contributions and work continues. Among the questions discussed are:

- Is text hierarchical or not? Always? Usually? In what sense?
- If text is not always or typically hierarchical, is this a problem for SGML/XML vocabularies such as the TEI?
- How should non-hierarchical structures be encoded?

Modified versions of the hierarchy thesis have been proposed, empirical surveys of the prevalence of hierarchy conducted, and alternative encoding systems developed.

4. Kinds of markup

Markup theorists have used a number of terms for kinds of markup, such as *explicit*, *implicit*, *procedural*, *presentational*, *punctuational*, *descriptive*, *prescriptive*, and *authorial*. Questions are raised about possible misunderstandings of the fundamental nature of markup, and some writers have argued that we limit the value of text encoding when we neglect to exploit certain sorts of markup.

5. Non-linguistic features of text

Citing for example Blake's attention to page design and the rat's tail in Alice in Wonderland, some argue that presentational features of texts (what Jerome McGann calls *bibliographic codes*) are themselves constitutive of text and not just properties of its rendition. McGann's discussion of this issue is considered by some to be a serious challenge to the TEI approach, although others have argued that the challenge is merely apparent. This discussion has recently grown to include the most general problems of understanding the interplay of text and images in the production of meaning.

6. Correspondence of descriptive markup to concepts guiding authors

Some markup vocabularies are presented as expressing concepts that an author easily perceives. However, it has been argued that often the relevant notions are far from consciousness, compromising the value of descriptive markup, at least for authoring but perhaps also for publishing and analysis. Moreover, some conventions of scribal writing systems, such as capitalization, italicization, underlining, and quotation marks, are routinely used in semantically ambiguous fashion, and authors cannot always explain why they used a certain type of markup. Forcing disambiguation would be not only onerous but perhaps semantically falsifying as well. Although this topic is not encountered so frequently as it once was, we hope to identify the original sources of this discussion and connect them to related work, such as recent discussion in the computer-supported cooperative work (CSCW) community of problems with excessive `formality'.

7. Vagueness, ambiguity, underspecification, and uncertainty

How can a markup vocabulary or syntax allow for such phenomena as vagueness, ambiguity, uncertainty? For example, could one create an element without an exact location, thereby avoiding false precision? The point at which a plot turns is a text element of this sort which one might want to encode. What is the right way to write a markup vocabulary or syntax that allows for ambiguity or underspecification? Conventions of scribal markup are semantically ambiguous, and underspecification and ambiguity are common in human language. What is the right way to write a markup vocabulary or syntax that allows for encoder uncertainty?

8. The relationship between markup and text

Does markup offer a translation of a text, an abstract representation of it, or both? Is markup part of the text or added to it? Do the answers to these questions depend on whether the markup is applied by the author of the text or by another? Or on whether the text is created originally using a computer markup language rather than transcribed from a document?

9. Data structures and data models

SGML/XML markup serializes a data structure, but some claim that it does not provide a *data model*, that this is a major flaw in applications such as the TEI.

10. The general nature of text

What is the general nature of text? Is it an abstract object or a useful fiction for managing material objects? Is it a social construction, an `objectively existing' entity, or both?

11. Markup as old wine in a new bottle

Does the use of markup make a genuine contribution to contemporary humanities scholarship? Or is markup in fact tied to an approach to literature that went out of style a century ago? We hope that our discussion of the other topics in this critical review will provide more substance to this `metaissue'.

Date: (revised May 31, 2004) Author: Kevin Hawkins and Allen Renear (revised KSH).