Theoretical Issues in Text Encoding—A Critical Review

Kevin Hawkins
University of Illinois at Urbana-Champaign
Graduate School of Library and Information Science
Electronic Publishing Research Group
Slavic and East European Library
http://www.students.uiuc.edu/~kshawkin/

Abstract

This project will survey the important theoretical issues in text encoding, as identified from the perspective of the humanities computing community. Below is an outline of topics to be covered with selected references for each. We welcome suggestions for additional topics or references, including those from perspectives other than humanities computing.

1 Introduction

Text encoding has long had an important role in the humanities computing community. Initially this importance derived from the need to develop systems for representing culturally significant texts that would allow them to be analyzed by computer. Such systems for encoding text needed to be as subtle and sophisticated as both the structure of the texts themselves and the hypotheses and theories about them that were being computationally tested. Not surprisingly, decisions about text encoding often seemed to reflect fundamental differences in method and approach toward the textual material. Later, with the advent of text processing and electronic publishing, text encoding presented new opportunities for innovations in scholarly communication. The development of interactive networked hypertext brought further considerations to bear, and text encoding practices seemed to some to confirm or refute various ambitious theories of authorship, culture, and communication. The fundamental interest of humanists in theories of cultural objects and the nature of representation inevitably ensured that text encoding—which seemed in its very nature intrinsically involved in these fundamental topics—would elicit controversy.

A number of specific issues and topics from the last twenty years or so can be identified as more or less classic text encoding-related debates within the humanities community. Most of these are fairly robust in that they have no widely accepted resolution and are regularly reiterated and improved. We believe that these are foundational topics that reveal much about both text encoding and humanities computing. However, it is difficult to review these topics since the discussions are often hard to track by citation alone and range across journals, email lists, conference panels, project documentation, and technical reports. In fact, catching up on this background is something of a rite of passage for newcomers to humanities computing, involving the inevitable, though not unpleasant, participation in oral tradition at convivial extracurricular events. Although we would not wish to reduce the opportunities for socialization and entertainment, we think that a more systematic treatment is of value.

Our project intends to identify each topic, briefly reviewing (from an historical perspective) the logic of the discussion to date and indicating the most important references and their role in the discussion. The survey will be posted electronically and updated by the editors. Suggestions and additions will be encouraged. Currently we have a very preliminary list of topics with associated references. Your suggestions of topics, references, or categorization are encouraged. [KSH, AHR]

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 theoretical problem in text encoding for computing humanists. Some encoding theorists claim, for instance, that the markup that predominates in the TEI [Text Encoding Initiative] 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.

Selected references:

4. ____________________________
5. ____________________________

3 Hierarchical nature of text

SGML [Standard Generalized Markup Language] / XML [Extensible Markup Language] is a grammar that yields a tree without cycles or overlaps. But while it has been argued that text itself has a hierarchical structure, a number of difficult cases suggests that the content objects of texts do not form a hierarchy. Is text hierarchical or not? If not, is this a problem for SGML/XML vocabularies such as the TEI?

Selected references:


4 Markup as translation or abstract representation

Does markup offer a translation of a text, an abstract representation of it, or both?

Selected references:

1. 
2. 

5 Kinds of markup

Markup theorists have used a number of terms to categorize kinds of markup, such as explicit, implicit, procedural, presentational, punctuational, descriptive, prescriptive, and authorial. We will describe and compare the various schemes.

Selected references:

8. 
9. 

6 Correspondence of descriptive markup to concepts guiding authors

Some markup vocabularies seem to be presented as expressing concepts that an author easily perceives. However it has been argued that often the relevant notions are far from consciousness and that fact compromises the value of descriptive markup, at least for authoring but perhaps also for publishing and analysis. Moreover, some conventions of writing systems—capitalization, italicization, underlining, and especially quotation marks—are routinely used in semantically ambiguous fashion. Forcing disambiguation would be not only onerous but perhaps semantically falsifying as well.

Selected references:

4. 
5. 

7 Vagueness, ambiguity, uncertainty

Is it possible to write a markup vocabulary that allows for vagueness, ambiguity, generality, underspecification, and the like?

Selected references:

1. 
2. 

8 Alternative techniques in markup

Should a markup language provide alternative techniques for marking up “the same thing”?

Selected references:

1. 
2. 

9 Data structures and data models

SGML/XML markup serializes a data structure, but some have claimed that it does not provide a "data model" and that is a major flaw in applications such as the TEI.

Selected references:


10 Non-linguistic features

Citing for example Blake’s attention to page design and the rat’s tail in Alice in Wonderland, some argue that presentational features ("bibliographic codes") of texts are themselves constitutive of the "logical text", and not just properties of its rendition.

Selected references:


11 Social construction

Are texts real entities or are they socially constructed?

Selected references:


12 Abstract vs. concrete

Are texts concrete physical entities (or equivalence classes of such things) or are they abstract objects?

Selected references:

3. 4.

13 ______________________________________
14 ______________________________________

Selected references:

1. 2.

14 ______________________________________

Selected references:

1. 2.