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:

1. TEI AI3 W5 The TEI Guidelines (Version 1.1 10/90): A critique. Report, TEI Literary Studies Work Group (AI3), 18 October 1991. Published on the Worldwide Web at http://www.tei-c.org/Vault/AIlist.html.
2. Sperberg-McQueen, C.M. "Text in the electronic age: Textual study and text encoding, with examples from medieval texts". *Literary & Linguistic Computing* 6, 1 (1991), 34-46
3. Lancashire, I. Sperberg-McQueen, C.M., et al. Discussion on the mailing list HUMANIST, November 1995 – January 1996. Available on the Worldwide Web at http://www.princeton.edu/~mccarty/humanist/.
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:

1. DeRose, S., Durand, D., Mylonas, E., and Renear, A.H. "What is text, really?". *Journal of Computing in Higher Education* 1, 2 (1990), 3-26.
2. Huitfeldt, C. "Multi-dimensional texts in a one-dimensional medium". *Computers and the Humanities* 28, 4/5 (1994), 235-241.
3. Renear, A., Durand, D., and Mylonas, E. "Refining our notion of what text really is: The problem of overlapping hierarchies". Published on the Worldwide Web at http://www.stg.brown.edu/resources/stg/monographs/ohco.html.
4. Barnard, D., Burnard, L., Gaspart, J.-P.; Price, L. A., Sperberg-McQueen, C. M., Varile, G. B. "Hierarchical encoding of text: technical problems and SGML solutions." *Computers and the Humanities*, 29, 3 (1995), 211-231.

5. Sperberg-McQueen, C. M. and Huitfeldt, C. "Concurrent document hierarchies in MECS and SGML". *Literary and Linguistic Computing* 14, 1 (1999), 29-42.
6. McGann, J. *Radiant Textuality: Literature after the World Wide Web*. Palgrave Macmillan, New York, NY, 2001.
7. Cover, R. Markup languages and (non-) hierarchies. Technology report, Cover Pages, 2003. Published on the Worldwide Web at http://xml.coverpages.org/hierarchies.html.
8. _____
9. _____

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:

1. Goldfarb, C. F. "A generalized approach to document markup." In *Proceedings of the ACM SIGPLAN SIGOA Symposium on Text Manipulation* (Portland, OR, June 1981), Association for Computing Machinery, 1981, pp. 68-73. (Adapted as "Annex A. Introduction to generalized markup" in ISO 8879.)
2. Lamport, L. "Document production: Visual or logical?". *Notices of the American Mathematical Society* (June 1987), 621-624. Available on the Worldwide Web at http://research.microsoft.com/users/lamport/pubs/pubs.html#document-production
3. Coombs, J., Renear, A., and DeRose, S. "Markup systems and the future of scholarly text processing". *Communications of the ACM* 30, 11 (November 1987), 933-947. Available on the Worldwide Web at http://www.oasis-open.org/cover/coombs.html. (Reprinted with new commentary in *The Digital Word: Text-Based Computing in the Humanities*, G. Landow and P. Delany, eds., Cambridge, MA, MIT Press, 1993, pp. 85-118.)
4. Cournane, M. "The application of SGML/TEI to the processing of complex, multi-lingual text". PhD dissertation, University College Cork, Ireland, 1997.
5. Renear, A. "The descriptive/procedural distinction is flawed," *Markup Languages: Theory and Practice* 2, 4 (Fall 2000), 411-420.
6. Piez, W. "Beyond the 'descriptive vs. procedural' distinction". In *Proceedings of Extreme Markup Languages 2001* (Montréal, Canada, August 2001), B. T. Usdin and S. R. Newcomb, eds.
7. Caton, P. "Markup's current imbalance". *Markup Languages: Theory and Practice"* 3, 1 (Winter 2001), 1-13.
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:

1. Coombs, J., Renear, A., and DeRose, S. "Markup systems and the future of scholarly text processing". *Communications of the ACM* 30, 11 (November 1987), 933-947. Available on the Worldwide Web at http://www.oasis-open.org/cover/coombs.html. (Reprinted with new commentary in *The Digital Word: Text-Based Computing in the Humanities*, G. Landow and P. Delany, eds., Cambridge, MA, MIT Press, 1993, pp. 85-118.)
2. DeRose, S. "Structured information: navigation, access, and control." Berkeley Finding Aid Conference, April 1995. Available on the Worldwide Web at http://sunsite.berkeley.edu/FindingAids/EAD/derose.html.
3. Shipman, F. and Marshall, C. "Formality considered harmful: experiences, emerging themes, and directions on the use of formal representations in interactive systems". *Computer Supported Cooperative Work* 8, 4 (Fall 1999), 333-352. Available on the Worldwide Web at http://www.csdl.tamu.edu/~shipman/cscw.pdf.
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:

1. Raymond, D., Tompa, F., and Wood, D. "From data representation to data model: Meta-semantic issues in the evolution of SGML". *Computer Standards & Interfaces* 18, 1 (January 1996), 25-36.
2. Buzzetti, D. "Digital representation and the text model". *New Literary History* 33, 1 (2002), 61-88.
3. Discussion [on whether XML/SGML defines abstract structures (trees) or sets of strings (elements)] on mailing list xml-dev, March 2003. Available on the Worldwide Web at http://lists.xml.org/archives/xml-dev/.
4. _____
5. _____

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:

1. McGann, J. "The rationale of hypertext". In *Electronic Text: Investigations in Method and Theory*, K. Sutherland, ed. Clarendon, Oxford, UK, 1997, pp. 19-46.
2. Renear, A. McGann, J., and Hockey, S. "What is text? A debate on the philosophical and epistemological nature of text in the light of humanities computing research". Available on the Worldwide Web at http://www.iath.virginia.edu/ach-allc.99/proceedings/hockey-renear2.html.
3. McGann, J. *Radiant Textuality: Literature after the World Wide Web*. Palgrave Macmillan, New York, NY, 2001.
4. _____
5. _____

11 Social construction

Are texts real entities or are they socially constructed?

Selected references:

1. Huitfeldt, C. "Toward a machine-readable version of Wittgenstein's Nachlass: Some editorial problems". *Philosophische Editionen. Erwartugen an sie-Wirkungen durch sie, Beihefte zu editio Band 6*, H. Senger, ed., Max Niemeyer, Tübingen, 1994, pp. 37-43.
2. Pichler, A. "Advantages of a machine-readable version of Wittgenstein's Nachlaß". *Culture and Value: Beiträge des 18. Internationalen Wittgenstein Symposiums.* (Kirchberg am Wechsel, Austria, August 1995), K. Johannessen and T. Nordenstam, eds., pp. 690-695.
3. Renear, A. "Theory and meta-theory in the development of text encoding". *Interactive Monist Seminar*, M. Biggs and C. Huitfeldt, eds., November 1995 – January 1996. Published on the Worldwide Web at http://hhobel.phl.univie.ac.at/mii/pesp.html.
4. _____
5. _____

12 Abstract vs. concrete

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

Selected references:

1. Renear, A. "Theory and meta-theory in the development of text encoding". *Interactive Monist Seminar*, M. Biggs and C. Huitfeldt, eds., November 1995 – January 1996. Published on the Worldwide Web at http://hhobel.phl.univie.ac.at/mii/pesp.html.
2. Huitfeldt, C. Presentations at workshop "Theory and Metatheory in the Development of Text Encoding". ACO*HUM: The Future of the Humanities in the Digital Age (Bergen, Norway, September 1998).
3. _____
4. _____

13 _____

_____

_____

Selected references:

1. _____
2. _____

14 _____

_____

_____

Selected references:

1. _____
2. _____