Text Representation and Textual Models

Dino Buzzetti
Department of Philosophy
University of Bologna
via Zamboni, 38
I-40126 Bologna BO
Italy
buzzetti@philo.unibo.it

The practice of text encoding, based on the Guidelines of the Text Encoding Initiative,[TEI] has prompted the need of a critical reflection on the adequacy of current encoding theory.[RMD] In its turn, the debate on the theory of markup [RTWb, RTWa, BH] has elicited new investigations on markup schemes as document grammars.[SH] Therefore, one central issue of markup theory seems to reside in clarifying the relationship between document grammars and textual structures.

A basic ambiguity hovers over the discussion on text encoding--an "inadvertent shift" from document to textual structures, from textual representation to textual model.[BR] It seems to lurk in the very definition of "what is markup" endorsed by the TEI: if markup is defined as "all the information contained in a computer file other than the 'text' itself," how can "'any' aspect of a text of importance to a researcher" ever "be signalled by markup"? [BS,2] Either markup represents that information which "'is not' part of the text"[CRdR,934] and is 'other' than text, or markup represents aspects of that information which "'is' part of the text, and is 'the same as' text."[BR] But not both, unless 'text' is taken in two different senses, as information "coded as characters or sequences of characters,"[D,1] i.e. as a digital expression on the one hand, and as the "content of the expression"[CRdR,934] on the other. The representation of any information content is not the information content which is represented by that representation and one should be wary of identifying textual structures with document structures or document grammars.

Reference seems appropriate here to Hjelmslev's fourfold distinction between "form" and "substance" respectively of the "expression" and the "content" of discourse.[H,47-70] Hjelmslev's model bears the stamp of Saussure's notion of the text as "a two-level signifier-signified construction" in the stress on the "inseparability of expression and content," but also enables us to distinguish between the form of the representation and the form of the content represented thereby.[S,39] If we find "acceptable" a "very broad definition" of textual structure as "the set of the latent relations among the parts" of a literary work,[S,34] we should ask ourselves how such relations can be accounted for in terms of the document grammar enforced by a certain encoding scheme. Is, for instance, the OHCO model, [dRDMR] or any model appropriately "refining our notion of text"[RMD] an adequate one? Can it provide a suitable data structure to process textual information for text critical or interpretational purposes?

It has been argued that it is "difficult" for markup "to express structure that is not a subset of character positions in the text,"[RTWa,9-10] and that it is so precisely because markup "inherits"[RTWa,2] the "order of text,"[RTWa,10] defined as the linear order of the string of characters in which it is embedded. In an SGML-like "strongly embedded" kind of markup, the position within the character string "is information bearing", whereas a "weakly embedded" kind of markup is informative regardless its position and it could be placed "out-of-line."[RTWa,3-4] All this means that SGML-based encoding "reduces the structural properties of a text to the structural properties of a document," [BR] because "the properties" of the structures which this kind of markup can describe "are largely derivative of the properties of the document," the linear stream of characters "in which it is embedded." [RTWa,3-4] Again, the form of the representation cannot be mistaken for the form of the content which is thereby represented.

The structural properties of an encoded document expressed by means of a strongly embedded markup scheme can therefore be described as essentially 'notational'. The notational properties of decimal, as opposed to binary or any other form of numerical notation, do not affect the arithmetical properties of numbers. In other words, "markup is not a data model, it is a type of data representation"[RTWa,16] and different forms of text representation should match data models capable of expressing the set of latent relations constituting textual structure.

Both text critical and interpretational procedures require non-linear models of the text.[Ba, Bb] Variant readings of a complex textual tradition cannot, in general, be represented in a linear way, neither can complex specimens of authorial manuscripts. Competing interpretational perspectives can detect concurrent textual structures. Also "historical text", or textual information enabling historical enquiry, can be conceived as multi-layered and demands non-linear models.[Ta,55-57] An "extended string" data type able to provide non-linear data models for processing textual information has been introduced precisely for this purpose.[Tb]

The extended string internal representation allows import/export of different forms of marked-up data for the same data model, of different "formats" for the same "formalism,"[J,75-76] and makes it possible to use concurrent encoded representations depending on different interpretations of the same text; or to combine encoded transcriptions of the several witnesses of a complex textual tradition into a sort of "logical sum." [Ta,56]

The difference between encoding a running text representation and relating its relevant segments to an external knowledge representation organized as a database [Ta,55-57] can be related to the distinction between the form of the expression and the form of the content. But in many cases the form of the exppression reveals a structural feature of its content. That is why this distinction is easily overlooked. What is then the conceptual status of markup? Is it a sort of metalinguistic description, or is it a direct expansion of our writing system employed to express intrinsic features of textual content? In the latter case markup is to be thought of as an extension of text representation, in the former as a separate and external representation of its structure.

This question has far-reaching implications, bearing as it does upon the capacity of language to be self-reflexive, but more significantly here it concerns directly the status of markup as a formal language and a document grammar.

Bibliography

[Ba] Buzzetti, D., 'Il testo "fluido": Sull'uso dell'informatica nella critica e nell'analisi del testo', in Luciano Floridi, ed., *Filosofia & informatica, Convegno Nazionale della Società Filosofica Italiana: Roma, 23-24 novembre 1995*, Torino, Paravia, 1996, pp. 85-93.

[Bb] Buzzetti, D., 'Digital Editions: Variant readings and interpretations', in *ALLC-ACH'96 Conference Abstracts*, University of Bergen, 1996, pp. 254-56.

[BH] D. Biggs, M., and C. Huitfeldt (eds.), 'Philosophy and Electronic Publishing: Theory and metatheory in the development of text encoding', in *The Monist*, 80:3 (1997), pp. 348-367.

[BR] Buzzetti, D., and M. Rehbein, 'Textual Fluidity and Digital Editions', in *Text Variety in the Witnesses of Medieval Texts*, edited by M. Dobreva, Proceedings of the International Conference, Sofia 21-23 September 1997, forthcoming.

[BS] Burnard, L., and C. M. Sperberg-McQueen, *Living with the Guidelines: An introduction to TEI tagging, Text Encoding Initiative*, Document Number: TEI EDW18, March 13, 1991.

[CRdR] Coombs, J.H., A.H. Renear, and S. J. DeRose, 'Markup Systems and the Future of Scholarly Text Processing', in *Communications of the ACM*, 30(1987), pp. 933-47.

[D] Day, A.C., *Text Processing*, Cambridge, Cambridge University Press, 1984.

[dRDMR] DeRose, S.J., D.D. Durand, E. Mylonas, and A.H. Renear, 'What is Text, Really?', in *Journal of Computing in Higher Education*, 1:2(1990), pp. 3-26.

[H] Hjelmslev, L., *Prolegomena to a Theory of Language*, 2nd Engl. ed., Madison, University of Wisconsin Press, 1961.

[J] Joloboff, V., 'Document Representations: Concepts and Models', in J. André, R. Furuta, V. Quint (eds.), *Structured Documents*, Cambridge, Cambridge University Press, 1989, pp. 75-105.

[RMD] Renear, A.H., E. Mylonas, and D. Durand, 'Refining our Notion of What Text Really Is: The problem of overlapping hierarchies', in *Research in Humanities Computing* 4, Oxford, Clarendon Press, 1992, pp. 263-280.

[RTWa] Raymond, D.R., F.W. Tompa, and D. Wood, 'Markup Reconsidered', paper presented at the *First International Workshop on Principles of Document Processing*, Washington DC, October 22-23, 1992.

[RTWb] Raymond, D.R., F.W. Tompa, and D. Wood, 'From Data Representation to Data Model: Meta-Semantic Issues in the Evolution of SGML', in *Computer Standards and Interfaces*, 10 (1995).

[S] Segre, C., *Introduction to the Analysis of Literary Text*, Engl. transl. by J. Meddemmen, Bloomington, Indiana University Press, 1988.

[SH] Sperberg-McQueen,C.M., and C. Huitfeldt, forthcoming.

[Ta] Thaller, M., 'The Processing of Manuscripts', in Id., ed., *Images and Manuscripts in Historical Computing*, St. Katharinen, Max-Planck-Institut für Geschichte in Kommission bei Scrpta Mercaturae Verlag, 1992 (Halbgraü Reihe zur historischen Fachinformatik, A14), pp. 41-72.

[Tb] Thaller, M., 'Text as a Data Type', in *ALLC-ACH'96 Conference Abstracts*, University of Bergen, 1996, pp. 252-54.

[TEI] Sperberg-McQueen,C.M., and L. Burnard (eds.), *Guidelines for Text Encoding and Interchange*, Chicago and Oxford, Text Encoding Initiative, 1994.