Digital Representation and the Text Model*

Dino Buzzetti

I

Does the digital representation of the text demand the definition of a model? In and of itself, every representation and, consequently, every form of text representation entails the implicit or explicit assumption of a model, at least if we accept the postulate that the "map is not the territory."[1] So, the conventional image of a text, handwritten or typed, is itself a text model. The same thing can be said, then, of its digital representation or, to be more precise, of every form of its digital representation, regardless of its specific kind. The problem of the model presents itself, therefore, as a problem of adequacy with respect to the conventional model and representation.

With respect to the conventional representation, an adequate digital representation should in no way impoverish the informative content of the text. If the digital text representation is not original, that is, if we consider a reproduction as opposed to a text produced directly in digital form, the first fundamental criterion for adequacy is constituted by the exhaustivity of the representation. In order to obtain the exhaustivity of the representation, markup is usually resorted to. In fact, textual information in machine-readable form is represented primarily by way of the binary coding of the characters: in this way, the "text" is conceived, from a computational point of view, as a type of data and the treatment of the text, that is, the "storage and processing of textual material," comes to consist in the treatment of "information coded as characters or sequences of characters."[2] It is evident, however, that the computational notion of the text as a type of data does not coincide with the notion of the text as a product of literary activity. The pure and simple character sequence is not adequate enough to represent all of the information contained in the "literary material as originally written by an author" (*TP* 1). Hence the need to furnish additional information by way of embedding markers defined by a given markup language.

* An earlier version of this paper has appeared in *Il ruolo del modello nella scienza e nel sapere* (Rome, 1999), pp. 127–61.

A second criterion for adequacy, equally important, concerns the liability of the digital representation to automatic processing and its functionality with respect to the critical operations of reconstructing or interpreting the text. For particular analytical purposes, the digital text representation may provide distinct advantages and be preferable to conventional representation itself. For example, the possibility of combining digital images with transcripts of the text renders the mimetic function of diplomatic transcripts superfluous and modifies its purposes. Combined with an image, a diplomatic transcript no longer serves to "reproduce the original," but rather to extract information from it and to represent it in an automatically processable form. In this light, the diacritical signs or the forms of markup are no longer conceived as an aid in visibly reconstructing an absent document, but rather as a means of "modelling" the physical and textual information contained in the original for the purpose of further processing.[3] The image itself, to the extent that it is a *digital* representation of visual information, does not provide merely a "facsimile" or "physical reproduction" of the original, but rather a set of "structured data," that is, a "logical representation" of the document's contents.[4]

The structure of the digital representation becomes very important with regard to the conditions of adequacy, because it imposes precise conditions upon the procedures used in the automatic processing of the informational content of the document. The representation's form must serve the analytical operations necessary to the study of the text. Even the form of the conventional text representation poses some problems (particular typographic stratagems have been proposed, for example, for the critical edition of texts handed down by a fluid or very complex tradition)[5] and the same problems occur, with equal if not greater prominence, with digital text representations. At all events, no form that sets fundamental limits on any analytical and scholarly operation can be considered a suitable form of representation. An adequate digital text representation must therefore be compatible with the application of the formal procedures of information processing which give algorithmic form to current methods and practices of textual criticism and interpretation.

The practice of markup, which became widespread with scholars who apply computational procedures to the study of a text, has revealed some difficulties which derive from fundamental theoretical options. These difficulties concern the conditions of adequacy of both kinds. On the one hand, a complete awareness of the theoretical status of markup, in its various forms, is lacking, and this has negative consequences upon the exhaustivity of the text's representation and the pertinence of markup. On the other hand, the forms of representation obtained

through markup have not always been compatible with the computational procedures required by the application of automatic methods to the critical study of the text.

II

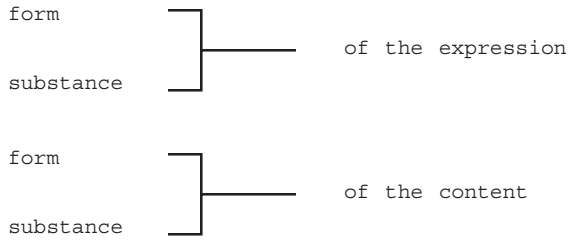
Essentially, the problems concerning the adequacy of the digital text representation may be formulated as such: (a) what are the formal characteristics of the data type through which the text representation is obtained? (b) what formal characteristics of the structural model of the text are needed in order to implement the operations necessary for its critical analysis? (c) which formalism, compatible with the syntax of the data and with the semantics of the model, may be implemented? The form of representation defined by the data type must be functionally compatible with the structural properties of the represented object defined by the type of model.

III

This specification of the problems presupposes a clear distinction between the structure of the representation and the structure of the object represented, a distinction that does not seem to have been addressed with sufficient clarity in the *Guidelines* of the *Text Encoding Initiative* (TEI),[6] the most organic proposal posited thus far with regards to the "encoding of texts" and the "data interchange in humanities research."[7] In the very definition of "what is markup" proposed by the editors of the *Guidelines*, we sense a "basic ambiguity"[8] which prevents us from discriminating between the formal properties of the "representation" of information, and the formal properties of the "information" represented (TF 30), that is, between the data considered as "information coded in a special way" (*TP* 1) and data considered as the model or the "abstract structure" of the information (TF 30). In fact, if markup is defined, as it has been by the editors of the *Guidelines*, as "all the information contained in a computer *file* other than the text itself," how is it possible that "*any* aspect of the text of importance to a researcher" can ever "be signaled by markup"?[9] One or the other: either, as others too have asserted, markup is held to be information that "*is not* part of the text"[10] and is *different* from it, or markup represents certain aspects of the information which "*is* part of the text, and is *the same as* text" (TF 31). But certainly not the two together, unless inadvertently, as appears the case here, *text* is assigned

two different meanings and construed in two different ways: on the one hand, as a sequence of encoded characters, that is, as an expression or digital representation of the information; or, on the other hand, as the "content of the expression" or the digital representation of the information (MS 934). Indeed, it should be clear that "the representation of information, or of content, is not the content or the information represented" or expressed "by such representation" (*TR* 219).

Still, the two meanings, though distinct, may be considered complementary and in fact both of them must be reinstated within an adequate notion of the text. The theory of the "literary"[11] text sometimes presents the distinction between the representation of information and the information represented as a distinction between the "expression" and the "content" of the text (*IA* 39). With regards to this, it may be appropriate to recall the description given by Hjelmslev,[12] of a "four-tiered model" which "bears the stamp of Saussure (expression = signifier; content = signified," (*IA* 39):



This model allows us to clearly distinguish between the form of the representation and the form of the content represented.

Another Saussurian element of the model proposed by Hjelmslev is manifested in "the stress on the inseparability of expression and content"(*IA* 39): on the one hand, "text is not to be identified with any of its several forms of representation;" on the other, the "textual representation is by necessity concrete and needs a physical support" (TF 29). Thus, there can be no text without representation; instead, the text *is* representation, but representation which depends upon an abstract content and is severally determined in the singularity of a concrete expression. All of which allows for the characterization of the text as *invariant*, to generalize an observation made by Cesare Segre, who refers to the "original constitution" of the text as a "series of letters and punctuation marks" that are the bearers of "graphic meanings," which in turn are "carriers of semantic meanings" (*IA* 24). So, if we refer to these original "signifiers" of the text,[13] that is, "if graphic signs (letters, punctuation, etc.) are looked at as signifiers for sounds, pauses, etc., and if we reflect on the fact that such signs can be transcribed over and over and in different ways (for

example, with different scripts and characters) without the value changing, we will be led to conclude that it is the text which is the *invariant*, the succession of values, with respect to variables like characters, script, etc." (*IA* 24, italics added). If we accept as "values," with respect to various expressions of the text, the semantic signifieds they bear, we may conclude that the whole of these signifieds constitutes "the *invariable* content of all of the material representations" of the text (*TF* 29).

From this we may infer the notion that the "content" or "signified" of the text is the unifying principle, in some respects, of its different expressions.[14] This notion may be applied, in the same sense, to single words, phrases and passages, that is, both to the text as a whole and to its individual parts. This allows us to compare the notion of text variation to the notion of synonymy and to establish a connection between the variant readings of the text, or the variable forms of the expression, on the one hand, and interpretative variation, or the variable forms of the content, on the other. The relevance of these considerations to the analysis of the digital text representation will be elaborated upon shortly, in section twelve. For the moment it will suffice to shed some light, from an abstract point of view, upon the correlation between the form of the expression and the form of the text's content, a correlation which does not, however, generally require any isomorphism as a necessary condition. The concurrent and complementary nature of the expression's form and the content's form undoubtedly constitutes a necessary prerequisite of the text, but the same cannot be said of their isomorphism. On the contrary, we must insist that this is the case only under very particular conditions.

IV

The distinction between "expression" and "content" may be suitably applied to the digital text representation as well. In its digital form, the text may be considered a "datum" and with regards to data, it is possible to distinguish between "data representation" and "data model."[15] The datum may be defined "in informatics" as the "representation (of information) in a certain code."[16] Here, "representation" refers to a combination of signs, that is, a combination of "objects," constituted by the physical states "upon which the executors of algorithms," that is, the computers, "operate"; and "information" refers to the "signifieds" which are "attributed to data," that is, to the combination of signs, or objects, or physical states which constitute their material expression (217). The form, or the organization of the objects which constitute the concrete representation of information, in short, the *form of the information's*

expression, is that which is currently called a "data structure," whereas the *form of the information's content*, the structure and the organization of the informative content, is that which is currently called a "data model."[17] Thus, the notion of "datum," as well as the notion of "text," are ambiguous notions and may refer to both the expression and the content of the information—an observation which is not surprising if we consider that the digital representation of a text is by definition a collection of data.

One must necessarily take into consideration this dual dimension of the data, the expression and the content, when reflecting upon the nature and the logical status of markup. Another way of expressing the distinction between expression and content, with reference to the data, is by distinguishing between "format" and "formalism." Even with regards to this it has been made evident that, even though "both terms (format and formalism) have a different meaning, they are often confused." A *format* may be defined as a "syntax" according to which a certain "representation" of information has been "codified," while a *formalism* may be described as what furnishes an "interpretation" of that "representation" of the information, that is, what assigns a particular "meaning" to the "sequence of digital values" of which it is constituted.[18] The "format" of the representation is what determines the "data structures," that is, the structural forms of the expressions of which the representation is constituted; "formalism," on the other hand, is what determines the "data model," that is, the "abstract form" of the content represented by such expressions (MR 13). "Format" and "data structure" are syntactical notions and refer to the expression of the digital representation; "formalism" and "data model" are semantical notions and refer to its content. A data model "needs to represent knowledge"[19] about the objects represented by the data, and consists in the formal specification of the abstract properties of the objects represented, together with the operations that can be defined upon such elements and the constraints which are applied to the various classes of objects. A data model may, therefore, determine an "algebra" for its objects[20] by utilizing a mathematical formalism completely "independent" of the particular form of "data representation" (FDR 4–5).

V

With these distinctions in mind, it is now possible to attempt an adequate description of the nature of markup. It has been affirmed that "markup is simply the denotation of specific positions" in the character sequence which constitute the text's expression by way of embedding

certain "tokens," which themselves consist of characters (MR 4). Markup may generally be defined then as "the use of embedded codes, known as tags, to describe a document's structure" (MR 1). Hence, what we are dealing with is "one of several possible techniques for representing [the] structure" of a "document," that is, the structure of that which properly constitutes the text's expression (MR abstract). One may say that "markup thus belongs not to the world of formalisms but to the world of representations," or that "markup is not a data model" but a "type of data representation" (MR 4, 16). In fact, while a data model directly describes "the semantics, and without taking into account the data representation" (FDR 5), it is evident that "the issues of markup are issues of *representation* of structure and not of its abstract form," considered independently of the expression (MR 13, italics added).

One may, however, note two fundamentally different ways of assigning a structure to a document and, consequently, two different types of markup. For the structure may be "placed" or "nested" in the text, considered as an expression or a string of codified characters, so as to be "strongly or weakly embedded" into it. In both cases, the tags which denote the structure are contained in the data representation. Though, in the first case "the position" in which they are embedded in the character sequence is "information bearing;" in the second case, however, while the tags are "informative," their "location within the text" or within the character sequence "is not information bearing." Thus, we may speak of "strongly embedded" markup and "weakly embedded" markup, respectively (MR 3–4).

It is thus evident that the "properties" of the structures which may be described by a form of strongly embedded markup "are largely derivative of the properties of the documents," or of the string of characters, "in which it is embedded" (MR 4). In fact, "because it shares the data representation," it is "difficult" for strongly embedded markup "to express structure that is not a subset of character positions in the text" (MR 9). Being indissolubly linked to the position in which it is embedded makes it possible for strongly embedded markup to inherit "many of the text's properties" (MR 2), meaning by "text" a character sequence and above all its linear "order" (MR 9, 10). But since "high level structures are not always composed of low level features," the structure of the text "is not always reducible to a functional description of the system's subcomponents," as, for example, the relative position of characters in their linear ordering (MR 9). Strongly embedded markup does not seem, then, entirely adequate for expressing the text's structure, and its "problems" may be considered the "result" of the "inherited properties" of the character sequence (MR 2).

The same thing can be said in another way by making reference to

the distinction between the *expression* and the *content* of the data and considering, respectively, their specific *form*, that is, the linear order of the succession of codified characters on the one hand, and the structure of that which the various strings of characters signify on the other. We may thus affirm that the dependency of strongly embedded markup upon the *form of the data expression* seriously limits its ability to represent the structure of the text. The *form of the data content* cannot generally be described by a function of the position of specific strings of characters within the linear sequence which contains them. In fact, the form of the data content is not necessarily isomorphic to the form of their expression.

In essence, a strongly embedded markup language is a language for representing data structures and not a language for representing data models. A strongly embedded markup language refers to the representation of information and not to the data content; it serves to describe the form of the data expression and not to describe factual "constraints" upon the structural elements of the information represented, or to specify "the effect of the operations" which may be executed by combining these formal objects (FDR 6). A strongly embedded markup system may therefore be described as a form of representation of the data *format*, and not as a form of representation of the *formalism* which defines the operators applicable to that data. In the end, it serves fundamentally to express structural properties depending on the "notational" features of the representation of the information content.

VI

Strongly embedded markup is thus essentially a representation of the *form of the data expression* and serves to make the structure of the text representation clear. It serves to represent the text representation and notably its structure. It will be necessary in section twelve of this essay to reflect more deeply upon the status of markup as *representation of the text representation*; but, if it is true that "the structural properties of an encoded document expressed by means of a strongly-embedded markup scheme" are "essentially 'notational' properties" (*TR* 220), then it must be concluded that the notational properties of the text representation, being the properties of the form of the data expression, must in no way be confused with the properties of their content. Analogously, we may observe that the notational differences between decimal numeration, binary numeration or any other system of numeration do not in any way affect the arithmetical properties of natural numbers. The operation

described below is always the same, independent of the system of notation adopted, decimal or binary, Arabic or Roman (fig. 1).

In the same way, the linear structure of the expression of a second degree equation used to describe a conic section is completely independent of the geometric properties of the figure represented (fig. 2).

These examples clearly demonstrate that it is absolutely inappropriate to have the structural properties of a representation's content dependent upon structural forms which the syntax of the representation language assigns to its expression. What a strongly embedded markup language does allow to be made clearer is thus only the *form of the text expression*, that is, the logical structure of a document. Furthermore, the structure assigned to the document by a strongly embedded markup system depends upon the syntactic characteristics of the markup language used to represent it. In fact, "there is no one logical structure for a document" (*MR* 16), and the form of the text's expression severally represented depends upon the expressive capacities of the strongly embedded markup system which allows it to be exposed. Consequently, it is not legitimate to identify the structure of the text's *content* with the structure of the text's *expression* which a particular markup language is able to represent.

VII

The expression and content of the text are, however, confused, as it has been shown, in the *Guidelines* of the TEI and in the encoding practice that ensues. The responsibility for the confusion can be ascribed to the ambiguous definition of the text "as an ordered hierarchy of content objects, or 'OHCO'" and to the hasty assumption that this definition was the "basic model of the text."[21] A "content object" is a portion of a "document" (WT 5) that contains or is contained within other content objects, or portions of the document, and that forms with them a "hierarchy" of containment relations, the smallest elements of which are "ordered" in succession, in the sequence of characters which form the document. The notion of "content object" essentially allows for

decimal	4	+	2	=	6	Arabic
binary	100	+	10	=	110	Arabic
decimal	IV	plus	II	is equal to	VI	Roman

Fig. 1.

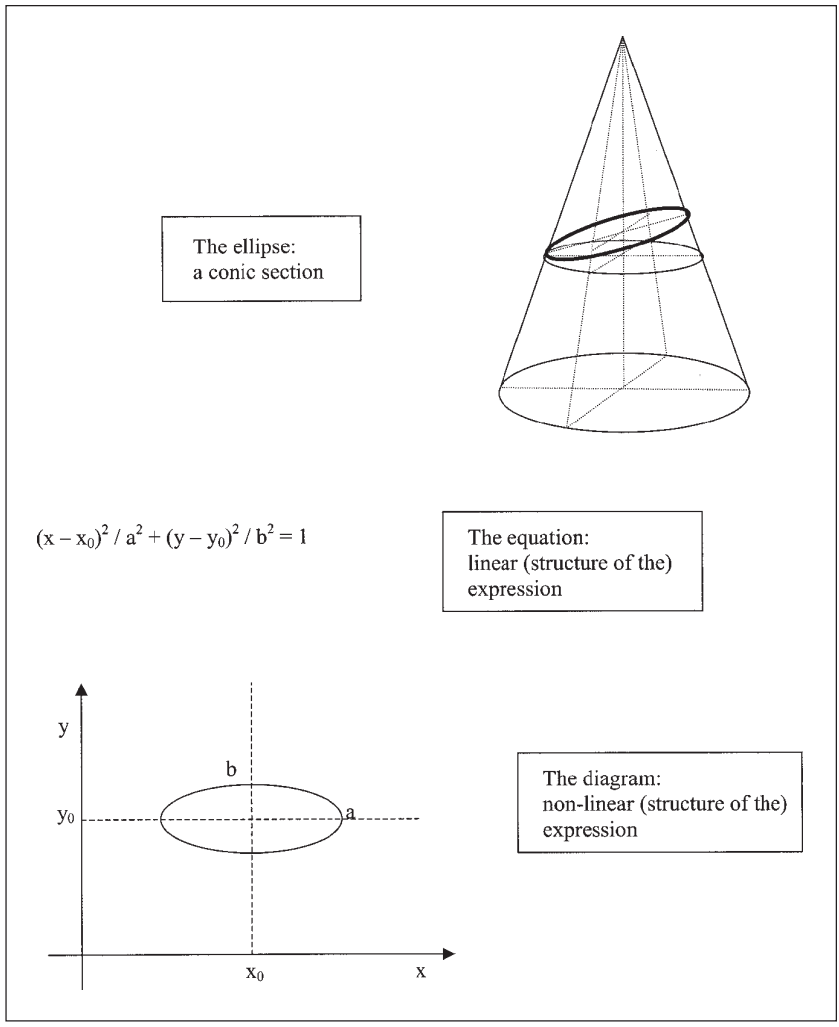


Fig. 2.

the conception of a document as a data structure and precisely as a tree structure, or graph, in which terminal nodes are arranged in linear order.

This definition was implicitly assumed in the ordinary practice of encoding with the "promotion of SGML," the ISO standard for descriptive markup systems,[22] as "a standard for encoding textual data" (WT 18). In fact, the TEI "relied upon the use of SGML as the fundamental language for the description of the text" (WT 13), and "SGML defines a document in terms of its OHCO structure." What SGML represents is "a

hierarchical document structure" described "with mnemonic names for the content objects of the data" which form it (WT 12). In essence, "one fundamental premise of SGML is that texts are composed of discrete content objects, and that supplying meaningful names for these delimited textual objects, their attributes and their hierarchical relationships, independent of possible appearances, is one of the most powerful means of transforming text into information units that may be addressed sensibly by knowledgeable software."[23] It is clear that this fundamental presupposition assimilates the expression to the content of the text and confuses the structure determined by the content objects—a structure formed by portions of the document consisting of strings of characters—with the "content structure" (WT 23) of the document, which is a structure made up of abstract elements of objects of knowledge. In fact, this is based upon the "thesis" that, "in some relevant sense of 'book', 'text', or 'document' (perhaps *qua intellectual objects*), such things are 'ordered hierarchies of content objects,'"[24] where the assimilation of the text to the book, or to the document that carries out its material representation, is clearly evident.

It is not SGML that is responsible for this confusion, but the way it has been used to represent the content of the text rather than its expression. SGML is described as a "metalanguage,"[25] one that "does not define a markup language" but rather the formal syntax for the use of strongly embedded tags. SGML is therefore "a language for defining *markup languages*" (LG 2) for strongly embedded markup systems, which allow for the structure or form of the data expression in terms of positions relative to the linear ordering of the encoded characters. In short, SGML is the formal specification of the syntax of the data expression. Therefore, SGML is a "representation language" conceived in order to describe "the logical structure of a document" (*SRS* 15), that is, the structure of the text's expression. The TEI, instead, "implicitly takes this structural model to be a fundamental model of the text representation, thus making the form of the text representation dependent upon a data structure" which describes *de facto* the structure of its expression.[26] An "OHCO structure" is not a model of the text, but a possible model of its expression.

The "OHCO thesis," however, is no longer accepted without reservation. In fact, the act of encoding texts itself has given rise to "some practical problems that seem to call this thesis to question" (*RN* 263). In essence, "the way in which texts were analyzed into objects by the text processing theorists and standard developers," a method founded upon the notion that the "text" is a string of characters, has revealed itself to be "fundamentally different from the way in which they were analyzed into objects by the literary and linguistic encoding community." So, "the

tendency of SGML to assume that documents could be represented as a single hierarchical structure created real practical problems for text encoding projects." But it is not only a matter of concluding that the OHCO thesis "is false" because you can assign "many hierarchical structures," all of which have "a plausable claim to be the 'logical' structure" (*RN* 269), to a single document, but rather, it is a matter of admitting that "the relevant logical structures of a text are not without exception hierarchical" (*RN* 279, n.13). Indeed, not only must we agree that "the same document conforms to several overlapping structures,"[27] each of which is "strict hierarchical," that is, composed of "objects always 'nested' and never overlapped" (*RN* 269); but we must also recognize that the textual structures are not usually of this type and that every "non-hierarchical structure constitutes a problem for an SGML-based encoding system for literary texts" (SB 266–67). SGML is a data structure representation language and the problem clearly arises from the confusion between the structure of the text's expression, which may be represented by a strongly embedded markup system, and the structure of the text's content, which does not generally conform to a hierarchical or linear type of model. In addition, SGML is not a data model representation language; indeed, it "defines no operators" (FDR 6, 15) and provides "no semantics" (MR 17). The use of strongly embedded markup systems generally poses, therefore, serious limitations upon the adequacy of the text representations: the use of "a universal linear representation of hierarchical structure 'is' an imposition which drastically curtails the representation of non-hierarchical structure."[28]

It is true that "any kind of data structure," linear or non-linear, "can be represented" linearly by a "stream" of characters that "conform" to the formal syntax of SGML (FDR 6), which is "a technique for the sequential, hierarchical representation of every data, with embedded tags representing beginnings and ends" (EM 129). However, this does not mean that any kind of structure "may be strongly embedded in the linear sequence of the characters" (*TF* 34) that form a text, or exclusively represented in function of the position of the tags within it. The syntactical structure which SGML assigns to the representation is, therefore, able to express not any given structure, but the form, or structure, of its representation. In fact, a non-linear structure may be linearly *represented*, but cannot be linearly *defined*, that is, calculated exclusively as a function of a purely linear order, such as that of the position of the characters in the linear sequence which constitutes the text's expression.

In summary, we may assert that strongly embedded markup systems prove to be inadequate in reference to both the exhaustivity and the functionality of the text representation and model. No form of strongly

embedded markup is able to represent non-linear structures as a function of positions assigned within the character sequence that constitutes the text's expression, and no form of strongly embedded markup is thus able to satisfy the criteria for the exhaustivity of the representation. In the final analysis, strongly embedded markup causes the possibly non-linear form of the data content to collapse upon the exclusively linear form of the data expression. The linearization of the *expression* cannot, however, bring about the linearization of the content and its structure. So strongly embedded markup is not able to represent non-linear structures of the data *content.* But it does not even allow for the acquisition of functionally adequate text representations. In fact, the linear representation of a non-linear structure of data, the format of the data expression obtained by strongly embedded markup systems, is not usually associated with the definition of operators which may be applied to the elements of the representation. The absence of operative definitions associated with the text representation makes it impossible, without further specifications, to apply algorithmic procedures, defined by a coherent data model, to the elements and data structures that it distinguishes. It is certainly possible to combine a formalism with "a system" of document management "which recognizes texts [strings of characters] containing a structure (usually indicated by the presence of markup);" however, this cannot provide anything more than a "formalism (usually a context-free grammar)" (MR 4) fit to the treatment of linear data structures, as for example the "OHCO structures." As a result, systems of strongly embedded markup are generally unable to satisfy the criterion of functionality for a text representation.

VIII

As opposed to strongly embedded markup, "weakly embedded markup" does not set *a priori* constraints on the functional adequacy of the text representation. By weakly embedded markup we mean "a structure not embedded [in the stream of characters] which conforms to the syntactic requirements of a certain markup standard." A weakly embedded "tag," is therefore "information bearing," but "its position" within the character sequence "is not informative." It may be placed "at any point within the text [the stream of characters], or even outside of it;" it is precisely for this reason that it is also known as "out-of-line" markup and is "more properly considered a specific type of external structure." Thus, it is possible "to distinguish between the internal and external structure" of the text, but it should be made clear that what we are referring to here is the linear sequence of characters which makes up the text's *expression*

(MR 3–4). The reasons for which an "internal" structure of the text's expression, isomorphic to its linear order, is generally unable to correspond to the content's structure, have already been made clear. Consequently, only a structure "external" to the text's expression is able to supply an adequate model of its content.

An external structure of the text's expression usually constitutes a data model. Reference to a data model overturns, in essence, the relationship between the data structures defined by the text representation and the particular formalism which regulates its manipulation. It is no longer the structure assigned by the markup to the text's expression that determines the operations applicable to the textual data, rather the operations applicable to the structural elements of the text's content determine the data structures that express its representation. In this way a document, or more broadly, the expression of the text, is generated from a formal model which processes its informative content. In short, the document or the text's expression depend upon the formal model by which they are produced. Herein lies also the concept of a kind of digital edition which is essentially based upon a generative conception of the text representation.

The idea of adopting a type of markup external or "parallel" to the "stream" of characters (*EM* 131), as opposed to the practice of strongly embedded markup, has been forcefully supported by Theodor Nelson, the acclaimed inventor of the term "hypertext," used "to describe the idea of the non-linear writing/reading in an informatic system."[29] According to Nelson, the structures of sequential and hierarchical data imposed by strongly embedded markup are absolutely "impeded" and "add obstacles" to the "exact representation of human thought, especially that thought put into words and writing" (*EM* 132–33). An adequate representation of the text's content may therefore be obtained only through forms of parallel markup; in this case, the "tags can be like those of SGML," but are certainly "not embedded" in the text, for they are rather "treated as separate" and placed "in parallel streams with reference positions in the text data stream"(*EM* 131).

Similar reasons are to be found in Manfred Thaller's proposal for the representation and treatment of a "historical text."[30] A historical text is described as "the formally treatable representation of the current assumptions of a researcher about what his documents actually contain" (*HI* 63), and presupposes reference to an external structure which organizes their content. This structure must be "built into" the "natural" representation of the text (*HI* 83), regarded as a "collection of linearly ordered strings" of characters (*HI* 65). An adequate representation of the text must then be able to "define the relationship between a 'text' as a running representation of a handed down document, and a 'text' as

converted into a database according to some abstract model" of data (*HI* 64). Overturning the perspective and taking the view of the data model able to represent the content's structure, a codified linear representation of the text appears to be a "realization" (*HI* 65) or "external representation" of the "internal representation" obtained within that model.[31] Actually, the proposed model provides for a "non-linear data type" (*HI* 61) consisting of an "extended string" (*TDT* 252) formed by ordered combinations of "tokens" which "can carry an arbitrary complex set of attributes." In turn, a "token," which is usually formed by a single string of characters, may be made up of a "non-ordered set of strings" (that is, all of the forms associated to a certain lemma).[32] This data type allows for the concurrent non-linear representation of different overlapping hierarchical structures, each of which may be exported or imported as a string of codified characters from a system that uses this model. Several hierarchical or non-hierarchical structures may therefore be simultaneously represented by the system and distinctly visualized by means of various linear representations of encoded strings.

Both of these solutions bring to mind the idea of "virtual"[33] or "active" documents which, not long ago, were considered "future applications" (MR 9), but which, with the development of the World Wide Web, find ever increasing implementations. A virtual document may be considered as a document "dynamically" produced through different "generation techniques" out of an "underlying" combination of "data" or "knowledge bases,"[34] that is, as "a document for which no persistent state exists and for which some or all of each instance is generated at run time."[35] It is interesting that particular attention has been paid to two specifically distinct forms of virtual documents: "hypertextual documents" and "conceptual documents,"[36] both of which correspond, for the most part, to the two solutions previously illustrated, and which refer, significantly, to two different types of digital editions, hypertextual editions[37] and editions in database form,[38] for which there already exist theoretical proposals and concrete instances.

IX

The considerations regarding the nature and the various forms of markup which have been made thus far have shown that a data model adequate to the text representation may be connected with its linear expression only through forms of external and weakly embedded markup. Markup, no matter what its kind, is essentially notational to the extent that it associates a structure with the text's expression and is not necessarily connected to any formalism. In addition, the forms of

strongly embedded markup generate data structures to which formalisms designed to manage structures made of non-linear objects are not applicable. The structure that may be assigned to the linear representation of the text by a strongly embedded markup system depends upon the form of the text's expression, which is neither isomorphic nor coincides with the form of its content. The attempt to use the processing model applicable to the expression of the text's digital representation as a processing model applicable to its content generates ambiguities and raises insoluble difficulties.[39] The processing of the expression and the processing of the content demand, respectively, two different formalisms. The expression's structure is in fact defined by a linear order, whereas the content's structure is defined, in general, by a multidimensional matrix.

In order to achieve congruence between the expression's form and the structural properties of the content represented thereby, it is necessary to find a formalism able to comply with the syntax defined by the data types and structures that form the linear representation of the text, and also able to operate according to the semantics assigned to it by the data model that structures its content. This is achieved with the use of weakly embedded markup systems capable of combining the external structure of the content model, usually non-linear, to the stream of the text's expression, which is essentially linear. So, for example, in order to overcome the presence of "overlapping hierarchies" in the text, some solutions have been proposed, which try to provide not only (a) "a convenient file format for recording overlap," that is, a suitable form of *markup*; but also (b) "a notation for expressing constraints on documents with overlap," that is, a formalism acting upon the objects represented, as well as (c) "plausible data structures for representing documents with overlap," or, in other words, a syntax for assigning an adequate structure to the text's expression.[40] So, in conclusion, a suitable digital representation of the text seems to require a weakly embedded markup system and a non-linear data model. The markup system must be capable of projecting the non-linear structure of the model upon the linear expression of the text, and the model must be able to represent the non-linear structure of its content. The data model may thus be conceived as a deep structure, able to generate distinct superficial structures or linear representations of the text, all connected to one another and all related to the same deep structure.

The textual model implicitly presupposed by any digital representation, in effect, coincides with the data model used to generate it. Every representation of the text presupposes a model of it, and the text's expression may be considered as the algorithmic structure which allows its processing. The text, every text, considering the material structure

with which it is constructed, may be conceived as the processing system of its conceptual content. Alphabetical technology is but one example of this construction. Digital technology offers greater possibilities to formalize such procedures.

X

If, then, the requirements which ensure the exhaustivity of the digital representation of the text reside primarily in its formal and structural characteristics, then the fundamental requirements which will guarantee its functional adequacy must be ensured by the definition of a data model capable of formalizing the critical and analytical procedures applicable to the study of the text. The data model upon which the digital representation of the text is founded must be capable of transposing, by way of algorithms, the procedures for textual criticism and interpretative textual analysis. The model must be able to satisfy the needs of the philologist and the editor, as well as those of the historian and the literary critic.

Under close inspection, both procedures appear to be based upon a one-to-many relation. The editor, in fact, must "unite the various sequential representations of the text drawn from the individual pieces of evidence into a single, coherent and non-sequential representation. This way, the entire structure of the intra-textual relationships among the various versions of the text, usually represented by way of the apparatus, may be reconstructed, processed and analyzed with a device capable of organizing into assorted structures, both sequential and non-sequential," the entire textual tradition, that is, the factual data at the editor's disposal for reconstructing the text. But "even with regard to the interpretation, it is a matter of combining into a single, coherent representation, numerous different structural representations." It is a matter of "bringing together a variety of structural forms, articulated and interconnected in many different ways, into a single representation providing for the application of coherent procedures in order to reveal single structural forms and allow for movement from one analytical structure to another" (*ITF* 87–88). So, "the task of a digital edition is best understood as an attempt to match the computational model of text representation and processing with the conceptual model of text reconstruction and textual criticism, on the one side, and, on the other, with the procedures and methods of text analysis and interpretation" (*TF* 36). Since the editor "must make a coherent unity of the several documents and forms of representation through which the text was transmitted to him," and the interpreter "must, on the other hand, go

from a single, coherent form of text representation to several ideational and interpretative structures which are compatible to one another," one may conclude that, in both cases, a single "comprehensive representation must be placed in relation to several and mutually compatible structures, in one case making the manifold one, and in the other, descending from one to manifold." In both cases the data model "may be the same" (*ITF* 88). A data model based upon, for example, "extended strings" (*TDT* 252) is able, on the one hand, to "reduce to a consistent unity a multiplicity of different and possibly overlapping hierarchical representations," thus responding to the needs of textual criticism, or is able, on the other hand, to "derive from a single sequential representation, a multiplicity of different structural representations,"[41] which instead meets the needs of textual analysis (fig. 3).

But if it is true that both of the operations, that of the editor and that of the interpreter, are essentially based upon the same one-to-many relation, and that they can therefore use the same data model, the fact remains that the one acts upon the structure of the expression (the textual variants) and the other acts upon the structure of the content (its many interpretations). Both presuppose a single, coherent, non-linear structural representation; but in the first case, this consists in the "logical sum" (*HI* 64) of the different concrete sequential representations that make up the complete textual tradition; in the second case, instead, it somehow constitutes the "hermeneutic invariant," that is, the abstract and conceptual "foundation"(*ITF* 88) of "the 'affinity'[42] between the ideation, or the forms of the author's ideation, and the interpretation, or the forms of the reader's interpretation," an affinity "which makes the text's comprehension possible" (*ITF* 88). However, the instrument used to algorithmically apply the two practices, the one regarding the expression and the other regarding the content, to the digital representation of the text is the same in this case as well and, in fact, is nothing more than markup. Markup makes the two practices perspicuous and connects them to the linear representation of the text. But even with regard to this, what is once again relevant, in order to more precisely define the dual function of markup, is the distinction between the expression and the content of the text. The markup, in fact, expresses the relation between the model and the expression of the data.

XI

A precise idea regarding the relation between the text's expression and content comes to us from an ancient neo-Platonic maxim: "Every dialogue is a cosmos and every cosmos is a dialogue."[43] We may conclude

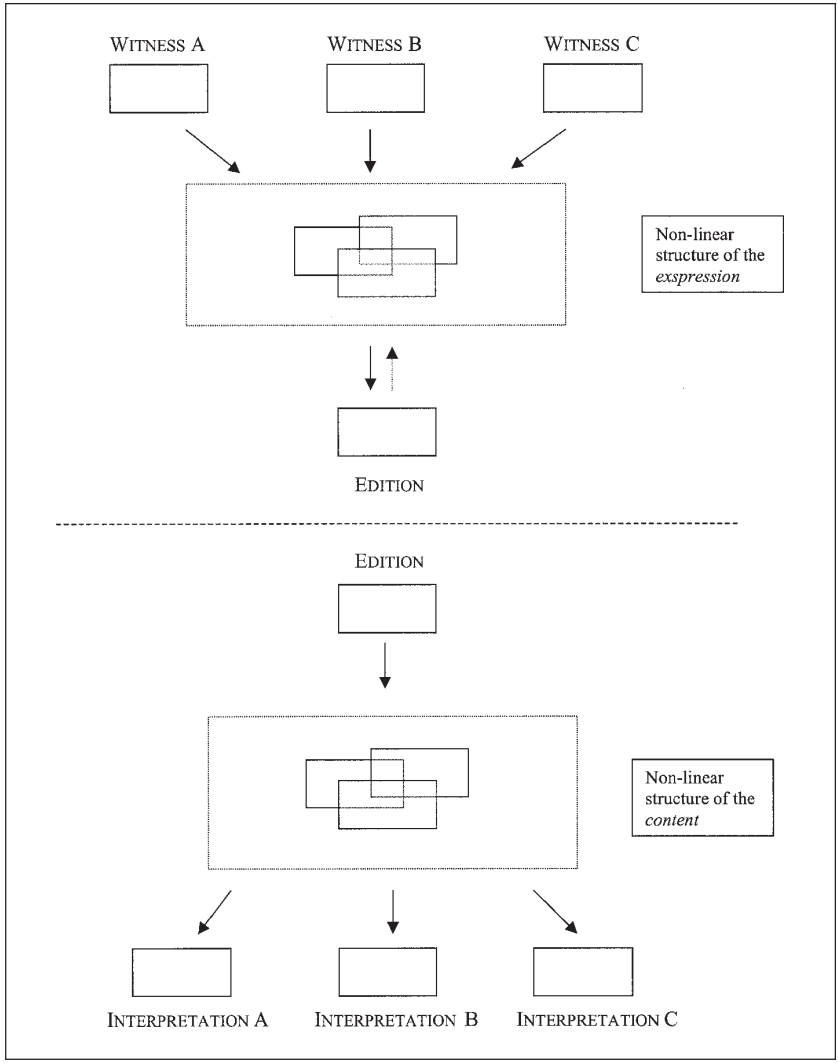


Fig. 3.

that every structural order, every model assigned to the world, is potentially reflected in the text which represents it. At the same time, in every text and in every interpretation of it lies a possible representation of the world, a way of assigning a definite order and a definite structure to it. So, to a given content corresponds an infinite possibility of expressions, and to a given expression corresponds an infinite number of interpretative approaches. The indefiniteness of the relationship

between the expression and the content is what assures the dynamism and mobility of the text: to any single expression many different contents may correspond, and to any single content many different expressions may correspond. If then a "very broad definition" of the text's "structure" may "prove acceptable," and if one intends it as "the set of latent relations" among its parts (*IA* 34), then again the number of possible structural determinations applicable to the text is potentially infinite. And if the text (or rather its realization through a concrete form of expression) may be conceived as the linearization of a complex content, then the markup may be considered the instrument most fit to expose their potential structural relations. Hence the importance of clearly comprehending the logical and linguistic status of markup in relation to the linear expression of the text.

Markup is "simultaneously embedded and separable" from the text; "it is part of the text and yet it is distinct" from it (MR 3). It is *representation* of structure and, at the same time, it is itself a *structure*. Indeed, on the one hand, it may be described as a "technique for representing structure" (MR abstract) and, on the other, as "structure" that is part of the text. It is representation of structure to the extent that it is "the denotation of specific positions" which identify single and determined strings within the linear sequence of the characters. And it is structure to the extent that the denotation of the position is carried out by way of "some assigned tokens" embedded and included in the sign sequence which makes up the text expression (MR 4). As *representation* of some structure, markup is a metalanguage; as a *structure*, and as it exposes some implicit structural features of the text, it constitutes an extension of it and increases its expressive resources. Markup, therefore, *denotes* structure and *is* structure. Just consider a very common form of markup, such as punctuation (MS 935; MR 9) or, say, the use and function of any diacritical sign. This dual expressive valency indistinctly characterizes every type of markup, so that both strongly embedded markup and weakly embedded markup can *denote* and *be*, respectively, the structure of the expression or the structure of the content of the text.

An adequate description of the logical status of markup and the ambiguous nature of its expressive function, demands some brief technical illustrations. By way of the direct expression of the structural aspects of the text, markup increases, so to speak, the expressive resources of the object-language, introducing a type of expression which entails a form of second-order predication, that is, a kind of predication which refers not to what the language represents, but to the way in which it ordinarily predicates, or to the forms of its first-order predication. A second-order object-language expression is therefore a self-referential expression which refers to abstract or structural properties of

the language as a form of representation, and thus to its method of representation, whether by way of the expression's form or by way of the content's form. A formal demonstration of the equivalence between second-order object-language expressions and metalinguistic expressions referring to the meaning of first-order object-language expressions corresponding to them, cannot be provided here. It should suffice to observe that the metalinguistic use of a certain tag describes the structural function carried out by the same tag when it is used as a direct expression of the structure of the text's content or the text's expression.

XII

The indissolubility of expression and content comes out once again in markup and has broad implications for its use. Markup is at once *representation* and *representation of a representation*. Markup also represents the *way of representing* of a given text. It expresses explicitly a self-reflexive dimension of the text. As *representation*, markup represents the text's content or expression, directly expressing its structure; as *representation of the representation*, it represents, metalinguistically, the structure of the text's expression or content. In any case, it is the representation of a structural aspect of the text, pertaining to its content or to its expression, in which a self-reflexive representational function may be exposed. In fact, the text itself is a representation, and its markup represents the manner in which this form of representation represents; and it does it in different ways, either metalinguistically, or by extending the expressive forms of the object-language. Thus the digital representation of the text may undoubtedly benefit from the use of markup, which exhibits explicitly these distinct linguistic functions. In the digital representation of the text, markup becomes the specific site for the exposure of the relation between expression and content, a relation which may be described externally, or revealed within the text through an increase of its expressive powers.

So, it is through markup that one of the most fruitful and most paradoxical dimensions of the text, and of "poetical textuality" in particular, may be expressed. Literature of a poetic nature entails sudden "shape shifting" generated by the constant tension between different conceptual structures, such as the emergence or sudden appearance of "a new and unexpected grammar" or a new "ideal order." In a poetic text there are constant and recurring shifts between "intratextual orders of textual relations,"[44] where the "sonic and visible features," together with "semantic, syntactic, and rhetorical features" constitute numerous "fields" of "textual action" which interact among

one another and make the text a site of "instability" (*RT* 175–76, 181). The poetic structure may, in essence, be seen as a device, or a "machine" made up of concrete signs which permanently assures its conceptual mobility. As a result, the text, or any of its elements, can never be conceived as "self-identical" (*RT* 175–76, 181). Unable to remain permanently identical to itself, the text may be thought of as an "algorithm," the "logic" of which is only "frameable in some kind of paradoxical articulation," such as the following:

$$A = A \text{ if and only if } A \neq A.^{45}$$

Just as we might say of an illustration, along with Wittgenstein, that "we *see* it as we *interpret* it,"[46] in the same way we might say of a text that we *understand* it as we *interpret* it. The famous example of perceptive "reversibility"[47] cited by Wittgenstein, the image which "can be seen as a rabbit's head or a duck's" (194), can be metaphorically referred to the text[48] to the extent that it allows forms of interpretative reversibility induced by the "interplay" of different fields of textual action (*RT* 183). The dynamic instability of the text's "structure," which "contributes to defining its essence and its meaning," as it assumes from time to time "an unexpected meaning" or it produces "an unforeseen effect,"[49] leads the text to be associated in the "field" (*RT* 183) of "perception" itself, "which foreshadows the formation of ideas and the development of thought"[50] with those "forms/non-forms"[51] that belong to "the class of "ambiguous" structures, natural and artistic (*SI* 37), whose distinctive character may "be defined as the coexistence, at a critical point, of two aspects or schemas of reality that are mutually exclusive" (*SI* 34).

Now, these critical zones may be highlighted in the text by the "complex networks" (*RT* 188) of textual relations manifested by markup. The latency of the textual relations (*IA* 34), the *virtuality* of the structure, is expressed in the indefiniteness of the relationship between the expression and the content and in its dual law of compensation. In the interpretative process, the indefiniteness of the content is compensated for by the definiteness of the expression; in the editorial and transmissive process, instead, the indefiniteness of the expression is compensated for by the definiteness of the content. So the conceptual "shifts" (*RT* 173) of interpretation are compensated for by the stability of the edition referred to, and the "fluidity"[52] of the textual tradition is compensated for by the presumed fixity of the content, as if it were conceived of once and for all in the author's original ideative act. The unstable relationship between the "mutation" (*RT* 173) and the "invariability" (*IA* 34) of the expression and content determines the *mobility* of the text, which is but the reflection of the indefiniteness of its structure.

The markup, then, becomes the site of the virtuality of the structure and the specific form of the unfolding of what is contained implicitly in the digital representation of the text. The markup highlights and registers the "instability" of textual phenomena; it visibly reflects the history of the transmission and the interpretation of the text; it reflects its expressive development and conceptual vitality. The entire critical apparatus of the digital edition is thus dependent upon the correct utilization of markup.

Markup, to the extent that it is an expression of the compensation between identity and variation, of both expression and content, with respect to the indefiniteness of the text's structure, lends itself to the representation of the non-identity of the text with itself. As we noted, the tags may be used as either metalinguistic descriptions of the text, or as self-referential extensions of its object-language. It is the latter use which makes it an instrument fit to represent the osmosis between textual and hermeneutic fluidity, between the variability of the textual readings and the variability of the text's interpretations. The tags' dual nature, metalinguistic and non-metalinguistic, may be placed in relation to the convertibility of expression and content at a level of linguistic representation which is typical of self-referential expressions. Assumed to be second-order self-referential expressions of the object-language, the tags used to express different conceptual and interpretative structures of the text may be assimilated to distinct textual variants. Reciprocally, different variant readings of the text may be assimilated to the expression of distinct textual interpretations. In essence, the markup may transform the interpretative variations into textual variations and textual variations into interpretative variations. Just as the relation one-to-many between the expression's identity and the content's variation can be turned into the relation one-to-many between the expression's variation and the content's identity, in the same way the markup can be considered a varying expression of a content which is always identical to itself, or a manifestation of the content's variation of a single and always identical expression of the text. In this way, markup becomes an instrument for use in transforming the implicit variation of the interpretation of an identical expression into the explicit fluidity of the expression of an identical content. Variously encoded textual portions, generated by different interpretations of one single expression, may be considered synonymous expressions of one single content. It is this characteristic, linked to the self-referential use of tags, which allows for the representation, through the markup, of the non-identity of the text with itself. In more technical terms, we could say that the markup may be used to represent an "endomorphism"[53] acting upon a combination of distinct textual units or definite elements in the text. In the final analysis, the

self-reflexivity of the markup is the condition for expressing, in explicit form, the non-identity of the text with itself. In formula,

$$(A = A \text{ if and only if } A \neq A) \langle\!\!-\!\!\rangle A \xrightarrow{\textit{markup}} A.$$

Expressing the compensation between variability and invariability, the markup allows for the exploration of the "negative space of textuality" (*RT* 190) generated by the indefiniteness of its structure. The one-to-many relation itself can express, in both senses, the compensatory correlation of the expression's structure with the content's structure and may be used as a foundation of the functional adequacy of possible operative models of the text.

The need to "deal analytically" (*RT* 190) with the text's instability thus finds one of the instruments for its realization by resorting to adequate forms of markup. The new "devices of page order" introduced by the "writing revolution" of the 12th century used to carry out the same function with regard to writing, the traditional form of text representation. Through its new "visual architecture," the transcribed text yielded "to the mental image of its structure," and with "the will to use visual articulation as a means of interpretation" came "the text as an object" in the form of a book, that visual "materialization" of a mental "abstraction" which, around 1460, by way of mechanical techniques, "was reified in printed form."[54] The critical apparatus of the printed editions carries out a similar function with respect to exhibiting and visualizing the series of relations which connect the different pieces of evidence of a textual tradition. In the same way, in a digital representation, markup may be used for exhibiting and visualizing the implicit textual relations and for representing the diverse phenomena of textual mobility. As we have seen, the correct use of markup and the adequacy of digital representation presuppose, however, recourse to a suitable text model. Thus, the formalization of the critical and interpretative procedures, and the specification of the operative model, remain essential for the production of digital editions capable of guaranteeing the exhaustivity and functionality of this new and more complex form of text representation.

University of Bologna
Translated by Jon & Marella Morris

NOTES

1 This is "the original statement for which Korzybski is most famous"—the statement that "the map is not the territory." This statement "came out of a very wide range of philosophic thinking, going back to Greece and wriggling through the history of European thought over the last two thousand years" (Gregory Bateson, *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology* [1973], [London, 1992], p. 455).

2 A. Colin Day, *Text Processing,* (Cambridge, 1984), p. 1; hereafter cited in text as *TP*.

3 Dino Buzzetti, "Image Processing and the Study of Manuscript Textual Traditions," in *Historical Methods*, 28 (1995), 148.

4 See Buzzetti, "Image Processing," p. 148, and Matthew G. Kirschenbaum and others, "Refining Our Notions of What (Digital) Images Really Are," in *ACH-ALLC'99 Conference Proceedings* (Charlottesville, 1999), p. 56.

5 See, for example, Georg Steer, "Textkritik und Textgeschichte. Editorische Präsentation von Textprozessen: Das Nibelungenlied. Der Schwabenspiegel. Die Predigten Tauler," in *Methoden und Probleme der Edition Mittelalterlicher Texte,* ed. von Rolf Bergmann und Kurt Gärtner (Tübingen, 1993), p. 118, and Dino Buzzetti and Malte Rehbein, "Textual Fluidity and Digital Editions," in *Text Variety in the Witnesses of Medieval Texts,* ed. Milena Dobreva (Sofia, 1998), p. 21; hereafter cited in text as *TF*.

6 *Guidelines for Electronic Text Encoding and Interchange*, ed. C. M. Sperberg-McQueen and Lou Burnard (Chicago, 1994).

7 Lou Burnard, "An Introduction to the Text Encoding Initiative", in *Modelling Historical Data*, ed. Daniel I. Greenstein (St. Katharinen, 1991), p. 83.

8 Buzzetti, "Text Representation and Textual Models," in *ACH-ALLC'99 Conference Proceedings* (Charlottesville, 1999), p. 219; hereafter cited in text as *TR*.

9 Burnard and C. M. Sperberg-McQueen, *Living with the Guidelines: An Introduction to TEI Tagging* (Text Encoding Initiative, Document Number: TEI EDW18, March 13, 1991), p. 2 (italics added); hereafter cited in text as LG.

10 James H. Coombs, Allen H. Renear, and Steven J. DeRose, "Markup Systems and the Future of Scholarly Text Processing," in *Communications of the ACM,* 30 (November 1987), 934 (italics added); hereafter cited in text as MS.

11 See Cesare Segre, *Introduction to the Analysis of the Literary Text*, with Tomaso Kemeny, tr. John Meddemmen (Bloomington, 1988); hereafter cited in text as *IA*.

12 See Louis Hjelmslev, *Prolegomena to a Theory of Language*, tr. Francis J. Whitfield (Madison, 1961).

13 The idea of the text as constituted by a "series of elements" which may be considered independently or as interdependent from one another has been developed by Tito Orlandi: the text is described as a "system" composed of a series of "(sub)systems" and the model of the text as an "inseparable combination of many levels" of complete and coherent formal representation. See also T. Orlandi, "Testi, modelli, e sistemi," in *Il ruolo del modello nella scienza e nel sapere* (Rome, 1999), pp. 73–90.

14 For an analogous view, see Arne Naess, *Interpretation and Preciseness: A Contribution to the Theory of Communication* (Oslo, 1953). I owe this reference to a conversation regarding the above with Claus Huitfeldt at the "Wittgenstein Archives" in Bergen.

15 See Darrell R. Raymond, Frank W. Tompa and Derick Wood, "Markup Reconsidered" (paper presented at the First International Workshop on Principles of Document Processing, Washington, D.C., October 22–23, 1992), http://www.csd.uwo.ca/staff/drraymon/publications.html; hereafter cited in text as MR.

16 Mauro La Torre, *Principi di informatica* (Firenze, 1994), p. 333; hereafter cited in text.

17 It is in this sense that the two notions—data structure and data model—are used by Darrell R. Raymond and others in "Markup Reconsidered."

18 Vania Joloboff, "Document Representations: Concepts and Models," in *Structured Documents*, ed. Jacques André, Richard K. Furuta, and Vincent Quint (New York, 1989), pp. 75–76.

19 L. Goldschlager and Andrew Lister, *Computer Science: A Modern Introduction* (London, 1988), p. 248.

20 Darrell R. Raymond, Frank Tompa and Derick Wood, "From Data Representation to Data Model: Meta-Semantic Issues in the Evolution of SGML," in *Computer Standards and Interfaces*, (18 January 1996), 25–36. http://www.csd.uwo.ca/staff/drraymon/publications.html; hereafter cited in text as FDR. What is meant here by "algebra" is a type of mathematical structure that may be generically described as "an aggregate of objects together with certain operations to combine them" (I. N. Herstein, *Topics in Algebra* [New York, 1975], p. 2).

21 Steven J. DeRose, David G. Durand, Elli Mylonas, and Allen H. Renear, "What Is Text, Really?," in *Journal of Computing in Higher Education*, 1.2 (1990), 6; hereafter cited in text as WT.

22 See ISO 8879:1986, *Information processing—Text and office systems—Standard Generalized Markup Language (SGML)*. Descriptive markup may be defined as the use of "mnemonic names" for objects endowed with content (DeRose and others, "What Is Text?," p. 5) or more generally for different elements of text representation.

23 Robin Cover, Nicholas Duncan, and David Barnard, "The Progress of SGML (Standard Generalized Markup Language): Extracts from a Comprehensive Bibliography," in *Literary and Linguistic Computing*, 6 (1981), 198.

24 Allen Renear, Elli Mylonas, David G. Durand, "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies," in *Text Encoding Initiative: Background and Context*, ed. Nancy Ide and Jean Veronis (Dordrecht, 1995), p. 265; hereafter cited in text as *RN*.

25 Joan M. Smith, *SGML and Related Standards: Document Description and Processing Languages* (New York, 1992), p. 15; hereafter cited in text as *SRS*.

26 Buzzetti, "Il testo 'fluido': Sull'uso dell'informatica nella critica e nell'analisi del testo," in *Filosofia & informatica*, ed. L. Floridi (Torino, 1996), p. 91; hereafter cited in text as *ITF*.

27 David Barnard, Ron Hayter, Maria Karababa, George Logan and John McFadden, "SGML-Based Markup for Literary Texts: Two Problems and Some Solutions," in *Computers and the Humanities*, 12 (1988), 266; hereafter cited in text as SB.

28 Theodor Holm Nelson, "Embedded Markup Considered Harmful," in *XML: Principles, Tools, and Techniques* (Cambridge, Mass., 1977), pp. 129–30; hereafter cited in text as EM.

29 Pierre Lévy, *Les technologies de l'intelligence: L'avenir de la pensée à l'ére informatique* (Paris, 1990), p. 34.

30 Manfred Thaller, "Historical Information Science: Is There Such a Thing? New Comments on an Old Idea," in *Discipline umanistiche e informatica*, ed. Tito Orlandi (Rome, 1993), pp. 51–86; hereafter cited in text as *HI*.

31 Thaller, "Text as a Data Type," in *ALLC-ACH'96 Conference Abstracts* (Bergen, 1996), p. 252; hereafter cited in text as *TDT*.

32 Thaller, "A Proposal for a Humanities Text Processing Protocol," in *ACH-ALLC'99 Conference Proceedings* (Charlottesville, 1999), p. 139.

33 See *HTF/VDI: Workshop on Virtual Documents, Hypertext Functionality and the Web*, organized by Maria Miroslavjevic, Fabio Vitali, and Carolyn Watters, at *WWW8: The Eighth*

International WorldWideWeb Conference, Tuesday, May 11, 1999, Toronto, Canada, http://www.cs.unibo.it/~fabio/VD99/index.html.

34 Maria Miroslavjevic, Fabio Vitali, and Carolyn Watters, "Introduction," in *HTF/VD*, http://www.cs.unibo.it/~fabio/VD99/intro.html.

35 Carolyn Watters and Michael Shepherd, "Research Issues for Virtual Documents," in *HTF/VD*, http://www.cs.unibo.it/~fabio/VD99/shepherd/shepherd.html.

36 Sylvie Ranwez and Michael Crampes, "Conceptual Documents and Hypertext Documents Are Two Different Forms of Virtual Document," in *HTF/VD*, http://www.cs.unibo.it/~fabio/VD99/ranwez/ranwez.html. A "hypertext document" is identified with a virtual document generated by the "navigation" of the user; its "final form depends entirely upon the user," while a "conceptual document" is described as a virtual document dynamically constructed by an "engine" which chooses, organizes, and assembles its information bricks.

37 See Jerome McGann, "The Rationale of Hypertext," in *Electronic Text: Investigations in Method and Theory*, ed. K. Sutherland (Oxford, 1997), pp. 19–46 and his "Imagining What You Don't Know: The Theoretical Goals of the Rossetti Archive," http://www.iath.virginia.edu/~jjm2f/chum.html.

38 See Dino Buzzetti and Andrea Tabarroni, "Informatica e critica del testo: il caso di una tradizione 'fluida'," in *Schede umanistiche*, N.S., 1.2 (1991), 185–93, which develops the idea of "edition as database," starting from the concept of "database as edition" introduced by Manfred Thaller (see Max-Planck-Institut für Geschichte, *Halbgraue Reihe zur Historischen Fachinformatik,* ed. M. Thaller, Series C: *Databasen als Editionen*, St. Katharinen, Scripta Mercaturae Verlag.)

39 Referring to the scheme illustrated by Tito Orlandi (cf. "Testi, modelli, e sistemi"), we may say that this would create some confusion between the two different "levels" of the text model, corresponding to two different "subsystems" which demand different formalisms and diverse implementation.

40 C. M. Sperberg-McQueen and Claus Huitfeldt, "GODDAG: A Data Structure for Overlapping Hierarchies," in *ACH-ALLC'99 Conference Proceedings* (Charlottesville, 1999), p. 198.

41 Buzzetti, "Digital Editions: Variant Readings and Interpretations," in *ALLC-ACH'96 Conference Abstracts* (Bergen, 1996), p. 255.

42 On the notion of "affinity" see Paul Yorck von Wartenburg, in *Briefwechsel zwischen Wilhelm Dilthey und dem Grafen Paul Yorck von Wartenburg. 1877–1897* (Haller, 1923), p. 191, cited by Hans Georg Gadamer, *Il problema della coscienza storica* (1963), from the Italian translation by Giangaetano Bartolomei (Napoli, 1969), p. 57.

43 See *Anonymous Prolegomena to Platonic Philosophy*, ed. Leendert Gerrit Westerink (Amsterdam, 1962), p. 30. Regarding this theme see also James A. Coulter, *The Literary Microcosm: Theories of Interpretation of the Later Neoplatonists* (Leiden, 1976), pp. 101–103, and Luc Brisson, "Le discours comme univers et l'univers comme discours: Platon et ses interprètes néo-platoniciens," in *Le texte et ses représentations* (Paris, 1987), pp. 121–28 (cited in *Notes complémentaires* to the Budé edition of the text, edited by Westerink, Jean Trouillard, and A. Ph. Ségonds, *Prolégomènes à la philosophie de Platon* [Paris, 1990]).

44 Jerome McGann, *Radiant Textuality: Literature Since the World Wide Web* (New York, 2001), pp. 175–76, 181; hereafter cited in text as *RT*.

45 McGann, *Radiant Textuality*, pp. 175–76, 181.

46 Ludwig Wittgenstein, *Philosophical Investigations*, Book 2, tr. Gertrude E. Anscombe (New York, 1958), p. 193; hereafter cited in text.

47 Compare the term "Reversibility," in Umberto Galimberti, *Dizionario di psicologia* (Torino, 1992), pp. 812–13.

48 A poetic component may also be seen as "a duck-rabbit work" (McGann, *Radiant Textuality*, p. 180).
49 Giuseppe Caglioti, *Eidos e psiche: Struttura della materia e dinamica dell'immagine* (Nuoro, 1995), p. 45.
50 Caglioti, *Simmetrie infrante nella scienza e nell'arte* (Milan, 1994), p. 37; hereafter cited in text as *SI*.
51 See R. Thom, *Stabilité structurale et morphogenèse* (Reading, Mass., 1972).
52 See Buzzetti, "Il testo 'fluido'" and Buzzetti and Rehbein, "Textual Fluidity."
53 For the notion of "endomorphism," see F. William Lawvere and Stephen H. Schanuel, *Conceptual Mathematics: A First Introduction to Categories* (Cambridge, 1997).
54 Ivan Illich, *In the Vineyard of the Text: A Commentary to Didascalicon* (Chicago, 1993), chapters 6 and 7.