Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies

Authors

- Allen Renear, Brown University
- Elli Mylonas, Harvard University
- David Durand, Boston University

Contact

Allen Renear
Scholarly Technology Group, Box 1885
Brown University
Providence, Rhode Island, O2912
Phone: 401 863-7312 Email: Allen_Renear@Brown.Edu

Status

Final version, January 6, 1993. A slightly edited version of this paper was published in 1996 in *Research in Humanities Computing,* Oxford University Press, Nancy Ide and Susan Hockey, eds. An earlier version was presented at the annual joint meeting of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, Christ Church, Oxford University, April 1992.

Contents

Abstract

We examine the claim that 'text is an ordered hierarchy of content objects'; this thesis was affirmed by the authors, and others, in the late 1980s and has been associated with certain approaches to text processing and the encoding of literary texts. First we discuss the nature of this claim and its connection with the history of text processing and text encoding standardization projects such as SGML and the Text Encoding Initiative. We then describe how the experience of the text encoding community, as represented and codified in the TEI *Guidelines*, has raised difficulties for this thesis. Next we consider two progressively weaker versions of this thesis formulated in response to these difficulties. Ultimately we find that no version appears to be free from counterexample.

Although none of these formulations proves to be theoretically sound, they are nonetheless methodologically illuminating as each generalizes actual encoding practices, making explicit certain assumptions that, even though they have been fundamental to the working methodologies of most text encoding projects, have never been explicitly articulated, let alone explained or defended. The counterexamples to the different versions of the OHCO thesis also arise in actual encoding projects -- so although our focus is theoretical it is grounded in the methodology and problems of contemporary encoding practices. The problems discussed here have implications not only for text encoding and our understanding of the nature of textual communication, but raise very fundamental issues in the logic and methodology of the humanities.

Introduction

It has been argued in many places and in many ways that documents are 'hierarchies of content objects' (e.g. Coombs, et al. 1987; DeRose, et al. 1990).[1] Text, according on this view, is essentially composed of nesting objects such as chapters, sections, paragraphs, extracts, lists, and so on. Those of us involved in humanities computing are likely to connect this claim with projects developing standards for encoding machine-readable texts. In this context the 'OHCO thesis' has proven to be a serviceable ideology in promoting the 'descriptive markup' approach to representing literary texts in machine-readable form -- and, more specifically, it has provided a congenial framework for the Text Encoding Initiative. But before it was adduced for this purpose the view that text was a hierarchy of content objects was implicit in early efforts to theorize about the development of text processing and typesetting software (Goldfarb 1981, Reid 1980). In fact, versions of this view which are entirely independent of any interest in computing applications can be discerned in the rhetoric of the 'parts of a book' which has been prevalent in style manuals and bibliography handbooks for some time (Renear, 1993).

Surprisingly, this thesis has undergone remarkably little refinement, extension, or clarification over the years it has served as the theoretical background for text processing research and text encoding standardization. This is due in part to the fairly implicit and heuristic role it plays in these activities. The Guidelines of the Text Encoding Initiative exhibit a characteristically ambiguous stance: although they seem to privilege this view and benefit from its influence, they do not specifically invoke, explain, or defend it.

Recently practitioners of text encoding have found themselves wrestling with some practical problems that seem to call this thesis into question. One of these may be called the 'problem of overlapping hierarchies' (Barnard, et al. 1988). This problem is generally taken to be purely a technical difficulty, the solution of which is to be found in any particular instance by considering the practical trade-offs of different encoding techniques. In fact, we think that the continuing perplexity that surrounds the treatment of overlapping hierarchies is not due to the technical issues of encoding at all, but rather to a more fundamental deficiency in our understanding of just what we are doing when we prepare an encoded text. The simple tenet that 'text is a hierarchy of content objects' fails us here'. Moreover, since the assumption of hierarchy rests primarily on its practical advantages -- and sufficiency -- for a variety of applications, rather than a principled analysis, we do not even have adequate terminology for describing the problems that arise.

This paper tries to take a few steps towards refining our notions both of what text is and what we are doing when we encode a text. As a device for raising these theoretical issues we focus primarily on the problem of nonhierarchical relations as this problem is actually encountered in ongoing text encoding projects.[2] We will see how an analysis of these problems bears on the slogan that 'text is a hierarchy of content objects'. In the process we will make explicit some of the assumptions behind current encoding practices. These assumptions, which are not always consistent, turn out to raise very profound issues, not only for text encoding, but for our understanding of the nature of textual communication and the logic and methodology of the humanities. As this is only an attempt to begin the necessary groundwork for a theory of text encoding, we have focused on expressing basic intuitions rather than rigorous and exact analyses -- those will come later.

Any discussion of 'what text is' broaches topics already well-developed in literary theory, the theory of textual criticism and scholarly editing, and the ontology of literary works. We will not be surveying these areas or attempting to directly relate our findings to discussions taking place there. Our intent is only to take a small step toward the development of a theory of 'what text is' by generalizing from the lessons of actual encoding efforts. This is a somewhat indirect way of proceeding to attack the general theoretical question of the nature of text, but we suspect that there is some benefit in a fresh perspective from an entirely new direction. Perhaps someday the theories generated in this empirical fashion will encounter -- either in agreement or contradiction -- the more general a priori reasoning of literary theorists and philosophers.

Before beginning we would like to fry a red herring. Most recent criticisms of descriptive markup and the content object approach to text have been motivated by supposed methodological problems in recognizing the features to be encoded -- these are the familiar and perennial controversies surrounding the 'subjective' and 'interpretative' nature of text encoding. Although these issues are important they are not our issues in this essay. We are rather concerned here with the internal coherence of the OHCO thesis itself and not with discovery procedures for content objects. Of course it may turn out in the end that the epistemology and metaphysics of text objects are profoundly entangled with each other -- but that is a conclusion one should reach as the result of argument and analysis, and not assume at the outset of an investigation of the principles implicit in current encoding practices.Our provisional attitude towards this issue is simple. The process of preparing a machine-readable text is in all essentials exactly like the process of preparing a traditional edition. No edition can be entirely 'theory-free', although they vary in the extent to which they are tendentious. Similarly for text encoding: no encoded text is strictly speaking 'theory-free', but without text encoding there is no machine-readable text at all. It should be a commonplace that machine-readable texts are 'subjective' and 'interpretive', but not *especially* subjective or interpretative. So we endorse Michael Sperberg-McQueen's first axiom about the markup used to implement text encoding: *Markup reflects a theory of text* (Sperberg-McQueen 1991). In fact, what follows can be considered to a large extent an extended meditation on this axiom.[3]

OHCO-1

Thesis: Text is an Ordered Hierarchy of Content Objects

> **OHCO-1:** Text is an ordered hierarchy of content objects.

OHCO-1 gives the ontological question 'what is text?' an equally ontological answer: text is an ordered hierarchy of content objects (DeRose, et al. 1990). The claim here is that in some relevant sense of 'book', 'text', or 'document' (perhaps *qua intellectual objects*) such things are 'ordered hierarchies of content objects'. A book for instance is a sequence of chapters, each of which is a sequence of major sections, each of which in turn is a sequence of subsections. Within the lowest level subsections are objects like paragraphs, sentences, prose quotations, verse quotations, equations, proofs, theorems, and so on. Many of these objects can be decomposed further. This structure is hierarchical because these objects 'nest' inside one another like Chinese boxes. It is ordered because there is a linear relationship to objects -- for any two objects within a book one object comes before the other.**[4]** Finally, we call these objects content objects because they organize text into natural units that are, in some sense, based on meaning and communicative intentions.**[5]**

The OHCO view of text can be contrasted with other models of text generalized from the software, practices, or methodologies that embody them: bitmaps (raster images), characters and formatting commands (procedural markup), glyphs and white space, character transcripts (the so-called 'ASCII' or 'text only' form of a document), and layout hierarchies.**[6]**

In both text processing and textbase development the superiority of the OHCO approach over these other models can be shown easily -- it is by far the simplest and most functional way to create, modify, and format texts; it is required to support effectively text retrieval, browsing, analysis, and other sorts of special processing; and texts represented according to this model are much more easily shared among different software applications and computing systems. The motivation for the view that texts are hierarchies of content objects lay initially in reflecting on these practical benefits of treating texts as if they were ordered hierarchies of content objects.

Arguments

The early positive arguments for text being a hierarchy of content objects were advanced largely to promote a particular approach to text processing and text encoding and to discourage the competing alternatives. The partisans of content-oriented text processing and descriptive markup claimed that treating texts as if they were ordered hierarchies of content objects had many practical benefits, while alternative representational practices resulted in various inefficiencies and inadequacies. It was a short step from noting the practical advantages of treating texts as if they were OCHOs to explaining those advantages by the hypothesis that texts are OHCOs. But once this hypothesis had been motivated by its practical advantages corroborating arguments of various kinds can be found.

Arguments that text is a structure of content objects are grouped here into three broad categories: pragmatic, empirical, and theoretical. This overview is designed only to briefly review the sorts of aguments that have been made; it is neither an endorsement nor a complete presentation of them. Of most interest here are the structure of the arguments and the assumptions underlying them.

Pragmatic

Pragmatic arguments are based on the practical advantages of the OHCO model. They originate with the designers of text processing software but have a broad appeal to anyone working with texts on the computer. These arguments begin, as described above, with the observation that there are many practical advantages to modeling a text as an OHCO rather than using one of the alternative models. Text modeled as OHCOs are easier to create, modify, print according to varying specifications, transfer from one application to another, and so on. Many of the analytic procedures or specialized processing one might want to do with a text are not even possible unless the OHCO structure is represented. These phenomena of the comparative efficiency and functionality of texts represented as OHCOs are best explained, according to this argument, by the hypothesis that texts are ordered hierarchies of content objects.

The argument has this form:

- If you treat texts as ordered hierarchies of content objects many practical advantages follow, but not otherwise.
- Therefore texts are ordered hierarchies of content objects.

Obviously only those theoretically inclined will explicitly draw the ontological conclusion -- text is an ordered hierarchy of content objects -- rather than the practical one: treat a text as (or model a text as) an ordered hierarchy of content objects. In fact, one can even argue for the usefulness of descriptive markup without claiming that one is treating texts as OHCOs, and many early promoters of descriptive markup did just that -- the issue of what they treated text as just did not arise. But it is difficult to explain the effectiveness of descriptive markup without saying, for instance, that one is identifying the relevant parts of a text or its stable salient features. If one then reflects on these explanations it is hard to to deny that one is representing the text as a certain kind of thing which consists of parts (objects) arranged in a particular way, to form a certain kind of structure: an ordered hierarchy of content objects.

Empirical

Closely related to the pragmatic arguments is a class of arguments that might be called empirical. These begin by observing that content objects and their relations figure very prominently in our talk about texts, and specifically in our descriptions, explanations, theories, hypotheses, and generalizations about texts. For instance, our theories and conjectures about literature make extensive use of terms for chapters, titles, sections, paragraphs, sentences, footnotes, stanzas, lines, acts, scenes, speeches, etc. These have prominent explanatory roles in our talk about texts and in our theorizing about texts and related subjects such as authorship, literary history, criticism, poetics, and so on. If we follow the recommendations of many philosophers of science and resolve ontological questions by looking to the nominal phrases in our theoretical assertions, then we will conclude that such things -- chapters, verses, lines, etc. -- are indeed the stuff of which literature is made.[7]

The empirical argument has this form:

- Content objects and their relations are the principal theoretical entities referred to by our theories, explanations, and descriptions regarding texts.
- Therefore texts are relations of content objects.

Theoretical

The arguments in this category are generally the least convincing, perhaps because they quite explicitly reveal the abstract and philosophical nature of the project proposed by the question: 'What is text?'. Nevertheless they are surprisingly common, in some form or other, when text encoders attempt to resolve hard problems in a principled way, or to justify their methodological resolutions.

The most important theoretical argument is a classic argument from variation of the sort used to distinguish essential from accidental properties in scholastic metaphysics, or, in a more contemporary philosophical idiom, to establish 'identity conditions' for objects. It notes that if a layout feature 'of a text' (such as leading or typeface) changes, the text 'itself' still remains essentially the same, but if the number or structure of the text's content objects changes -- say the number of chapters varies or one paragraph is replaced by another -- then we no longer, strictly speaking, have 'the same text'. You and I can both be reading the 'same text', say *Moby-Dick*, even though mine is in Times and yours in Palatino, or even though mine is in 10 point type and yours in 12 point type -- so that mine has more typographical lines, pages, and line end hyphens. On the other hand if my copy has fewer or different paragraphs than yours, or has its sentences in an entirely different order, then that seems to decisively argue that we are not reading 'the same text'.[8]

The argument goes like this:

- x and y are the same text if and only if they are the same ordered hierarchy of content objects.
- Therefore texts are ordered hierarchies of content objects.

Other theoretical arguments are often made as well. For instance, it is sometimes claimed that the competing non-OHCO models listed above omit essential information about the text; that an OHCO representation can generate the other proposed representations but not vice versa; and that understanding and creating text essentially involves grasping the OHCO structure of a text, but does not essentially involve grasping any other structure -- and that each of these facts implies that texts are OHCOs. We shall not discuss these arguments directly here, though our counterexamples will also apply.

Counterexamples: Multiple Logical Hierarchies

The forgoing arguments provide the initial support for the hypothesis that texts are ordered hierarchies of content objects. If these arguments are good ones then this thesis:

1. explains the success of certain representational strategies
2. is implied by our theorizing about literature
3. matches our intuitions about what is essential and what accidental about textual identity.

These arguments seem at least promising enough to shift the burden of proof on to the shoulders of those who believe that the OHCO thesis is false. Critics of this view must come up with more persuasive alternative accounts, or, at least, counterexamples.[9]

Are there any counterexamples? In fact, there are, and in retrospect it may seem hard to understand how these counterexamples could have been ignored. But how they happened to be ignored or minimized, at least by some of us, is a story that itself is an important piece of the recent history of text encoding. It is closely connected with the fact that the principal way in which texts were analyzed into objects by the text processing theorists and standards developers of the early 1980s is fundamentally different from the way in which they are analyzed into objects by the literary and linguistic encoding community of the late 1980s. In short: the a text as seen by the SGML community is not the same as the text seen by the TEI community -- that is, the accounts that they would offer of a text's structure are significantly different.

In the Old (SGML) View Genres Determine Text Objects

During the initial development of descriptive markup systems and the content object approach to text, each document was seen as having a single natural representation as a 'logical' hierarchy of objects, as determined by the genre of the document. What text objects might occur in a document on this view is a function of the genre or category of text that that document belonged to: legal contracts had one set of objects, scientific monographs another -- poems, novels, play scripts, letters, sermons, prayers, invoices, petitions, receipts, summonses, and so on all had their own set of objects and grammars that specified the syntactical relations those objects could have. Although representations of a particular document might differ when there was some uncertainty about the structure of the document being represented, and the specificity or granularity of a representation could vary, there was a sense that a single document structure was being encoded, that the document structure was being encoded, and any substantial differences indicated a disagreement about the structure of the document.[10] And in any representation the objects always seemed to form strict hierarchical structures: i.e. objects always 'nested' and never 'overlapped' This is, indeed, the view that is represented in the OHCO thesis.[11]

As well as the 'logical' structure of a document there were also alternative 'physical' representations. These were typically created by formatting or in some other way processing the logical document. Although objects within the logical structure never overlapped with each other, and objects from each physical structures did not overlap with each other, it was possible for objects from within the logical structure to overlap with objects from a physical structure. For example, while logical objects such as sentences, paragraphs and sections do not overlap with each other and physical objects such as typographical lines, columns, and pages do not overlap with each other (in a single layout design), the objects from a logical structure frequently overlap with objects from the physical structure: a sentence, for instance, may begin in the middle of one typographical line and end in the middle of a later typographical line. The SGML standard and its associated exegetical literature reflect this view. For instance, the definition of 'document type' indicates the role of genre in driving the hierarchical structuring of a document instance: '4.102 document type: A class of documents having similar characteristics; for example, journal article, technical manual, or memo.'

Document types in SGML are given a specific document type definition that, among other things, constrains all instances of that type to be hierarchical structures of text objects ('elements'). Consistent with this view research projects on text processing designed text processing systems that maintained exactly two hierarchies, the hierarchy of logical objects and a hierarchy of intended layout objects.(Chamberlin et al., 1987).

Typical document types and text objects:

Book:
    front matter, back matter, body, chapter, section, paragraph, extract, footnote...
Article:
    title, author, affiliation, abstract, section, subsection, paragraph, extract ...
Letter:
    sender address, recipient address, salutation, body, paragraph, close, scrivener initials, enclosure note
Poem:
    title, stanza, line
Script:
    cast list, Performance history, title, stage directions, act, scene, line

The New (TEI) View: Perspectives Determine Text Objects

When researchers from the literary and linguistic communities began using SGML in their work, the tendency of SGML to assume that documents could be represented as a single logical hierarchical structure quickly created real practical problems for text encoding projects. These problems were compellingly described by Barnard and others in a 1987 article (Barnard et al. 1987).

Briefly the difficulty is that while the SGML world seemed to assume that text encoders would always represent a text as a single logical structure, there in fact turned out to be many hierarchical structures that also had reasonable claims to being 'logical'. The hierarchy which was taken to be the logical hierarchy of a document was what one might call the 'editorial' hierarchy and corresponded more or less to the 'parts of a book' (or analogues for other document types) that one found discussed in style manuals -- objects such as chapters, sections, paragraphs, etc. This was not surprising. SGML had its origins in organizations concerned with using computers to create and typeset technical documentation and other commercial publications. In this milieu the editorial structure of a text can easily be taken as its only logical structure.

What Barnard's article pointed out was that there are many features of interest to scholars which taken together do not form a single hierarchy, but which nevertheless all seem plausibly 'logical'. Consider a verse drama for instance. It contains dialogue lines (speeches), metrical lines, and sentences. But object such as these do fit in a single hierarchy of non-overlapping objects: sentences and metrical lines obviously overlap (enjambment) and when a character finishes another character's sentence or metrical line then dialogue lines overlap with sentences and metrical lines. Yet all of these objects have equal claim to be 'logical', at least in given our so far very casual notion of what is meant by 'logical' -- they certainly cannot be assigned to 'physical' hierarchies.

Consequences of the Shift: There is no Unique Logical Hierarchy

On the old view text objects were grouped into families as determined by genre or category of text element (SGML 'document type'). On the new view families are determined by the analytical or methodological perspective on the text.

Some examples of such perspectives and typical elements they might contain are:

- Dramatic: act, scene, stage directions, speech, ...
- Prosodic: poem, verse, stanza, quatrain, couplet, line, half line, foot...
- Narrative: preparatory, villainy, insufficiency, reaction, victory... (Propp)
- Rhetorical: proem, narrative, arguments, subsidiary remarks, peroration... (Korax of Syracuse)
- Discourse: opening, check, topic changing, ending...
- Axiomatic: Primitives, axioms, definitions, theorems, proofs, counterexamples, definienda, definientes, clauses...
- Syntactic: Sentence, noun phrase, verb phrase, determiner, adjective, noun, verb...

Any of these structures has a plausible claim to be the 'logical structure of the text -- for instance they all fit the notion of 'content object' both as suggested by the gloss 'having to do with meaning and communicative intention' and as contextually defined by the arguments given above in support of OHCO-1. But because there is no single logical hierarchy which contains all of these structures we can no longer claim that 'text is an ordered hierarchy of content objects'. Once the class of logical elements in a given text is expanded to include all of the different perspectives we will inevitably be beset by overlapping objects: there is no unique hierarchy of content objects which is the text. OHCO-1 is false.

OHCO-2

Thesis: Perspectives Determine OHCOs

Although the original OHCO thesis can be seen to be false, a weaker revision suggests itself almost immediately and seems, moreover, to also reflect actual encoding practice. Text encoders dealing with overlapping objects found that although objects from different analytical perspectives would overlap with each other, pairs of objects from within a single analytical perspective seemed never to overlap. For instance, in the present example prosodic objects (stanzas, lines, half lines, couplets, etc.) do not overlap with each other, nor do the linguistic objects (sentences, phrases, words), nor do dramatic/editorial objects (title, cast list, acts, scenes, stage directions, dialogue). Each perspective on the text -- prosodic, linguistic, dramatic -- seems to determine an exact hierarchy. So although there apparently is no single OHCO which is 'the text itself', apart from a reference to a methodological community or analytical perspective, the objects that are determined by these various analytical perspectives seem to organize themselves, without exception, into hierarchies. This was good news for text encoders because it meant that each perspective could be represented as a document type and was thus amenable to description within the powerful SGML formalism -- only now document types would correspond to analytical perspectives and not to genre. Moreover, SGML contains a feature, CONCUR, that allows multiple hierarchies of a document to be represented and coordinated, so it seemed that ultimately the problem of overlapping objects would not be a practical problem for encoding projects.

These considerations suggest a plausible revision of the OHCO thesis:

> **OHCO-2:** An analytical perspective on a text determines an ordered hierarchy of content

> objects.

A rough explication of the technical phrase 'analytical perspective':

> An *analytical perspective* is natural family of methodology, theory, and analytical practice

OHCO-2 reflects the commonplace truth that there is no univocal sense of 'text', 'book', or 'document' and that consequently these words do not, without further qualification, designate genuine 'natural kinds' that play useful roles in explanations and descriptions of the world. Instead, they have many different senses that play various very diverse theoretical roles and invoke different complexes of associated concepts.

OHCO-2 does seem to reflect actual text encoding practices. When it is discovered that an analysis of a text appears to have two objects overlapping, encoders will typically consider this prima facie evidence that these two objects are not things of the same sort, that they belong to different analytical perspectives and therefore should not be placed in the same document type. For instance, if sentences overlap with metrical lines that is because one is a linguistic object and one a prosodic object. If speeches overlap with pages that is because one is a dialogue object and the other a typographical object. On this view if two objects are both, say, truly prosodic then they simply cannot overlap. If text encoders find that their analysis has overlapping objects, then they typically attempt to classify one of these objects as an object in a different analytical perspective. That classification is generally made initially by appealing to one's general intuitive sense of what sort of object it is -- metrical lines, feet, couplets, all seem to belong together, as do speeches, sentences, phrases, and words. But if the object could in some circumstances overlap with those already considered exemplary members of its new classification, then that is taken as evidence that the proposed re-classification is not the right one. Objects thus tend to sort themselves into natural families which, because there is no overlap within a family, may be handled nicely with the apparatus of SGML: different document types are assigned to each family and coordinated, when a single text is involved, with CONCUR.

The principle being followed here is a logical consequence of OHCO-2.

> **OHCO-2.1:** If two objects x and y overlap then they belong to different perspectives

Because it generalizes actual encoding practice and seems to have some independent plausibility, apart from its being a corollary of the stronger and more theoretically ambitious OHCO-2, we might ask directly about the logical status of this principle: Are texts, vis a vis some perspective, themselves hierarchical or is it an a priori truth of human experience that our analytical perspectives carve the world up into hierarchies? Or perhaps hierarchical division is merely often useful and consequently common, but not required by either the structure of texts or the nature of human reasoning. Cynics will suggest that text encoders had ample secondary motivation for adopting OHCO-2.1: if perspectives did not sort themselves out into hierarchies, then encoding projects would be deprived of the considerable benefits of SGML formalisms.

Counterexamples: Enjambment for Instance

Does every perspective determine a hierarchy of content objects? Obviously there is a difficulty here in our rather rough understanding of what a 'perspective' is. If 'literary studies' is itself a perspective then indeed not all perspectives determine hierarchies: literary studies discusses sentences, themes, pages, metrical lines -- and these, as we have seen, can overlap. We might be tempted to refine our notion of analytical perspective in such a way as to exclude 'literary studies' as being a true perspective, perhaps because it does not have sufficient theoretical coherence or specificity, although such a maneuver would be suspiciously ad hoc.

There is however a quicker way of casting doubt upon OHCO-2. Discussions of many sorts about texts are filled with characterizations, descriptions, and hypotheses that explicitly relate text objects from different perspectives -- chapters and themes, speaker and meter, narrative and paragraphing. Moreover there are even technical terms, such as enjambment and caesura, that specifically refer to relationships between objects from overlapping families. Because a technical vocabulary can be plausibly considered a sign of an analytical perspective the existence of this terminology suggests that there are analytical perspectives that contain overlapping objects.

OHCO-3

Thesis: Perspectives Can Be Decomposed into OHCOs

The final repair may be too nice for some, but having defended versions of the OHCO thesis this far, and having found the OHCO approach both theoretically compelling and useful as a practical principle of encoding, we cannot resist a final, and, we think, natural, revision. Moreover, it is one that, like the others, seems to underly coding decisions and preferences as we have seen them made in text projects.

First a new piece of terminology:

> **Sub-Perspective:** x is a *sub-perspective* of y if and only if x is a perspective and y is a perspective and the rules, theories, methods, and practices of x are all included in the rules, theories, methods, and practices of y, but not vice versa.

The idea is roughly that of a sub-discipline or some other sort of unified and coherent part of an analytical perspective. For instance, literary history, literary criticism, and textual criticism might be considered 'parts of, 'areas of', or 'sub-fields of' literary studies. And each of these in turn also has sub-fields. Textual criticism, for instance, might be plausibly said to have as parts: transcription, recension, and emendation. This notion of a part or sub-field of a discipline or analytical practice is what is meant here by 'sub-perspective'.

A version of the OHCO thesis that allows perspectives with overlapping objects, but still asserts a significant role for hierarchies in our understanding of what a text is:

> **OHCO-3:** For every distinct pair of objects x and y that overlap in the structure determined by some perspective P(1), there exists diverse perspectives P(2) and P(3) such that P(2) and P(3) are sub-perspectives of P(1) and x is a object in P(2) and not in P(3) and y is an object in P(3) and not in P(2).
>
> (or: objects may overlap in a perspective, but if they do then they belong to different sub-perspectives of that perspective)

The simple model of text -- an ordered hierarchy of content objects -- had a nice platonic shape to it. The current one is like something out of Yeatsean neo-platonism: a text now seems to be a kind of system of concurrent perspectives which decompose into concurrent sub-perspectives which in turn can be decomposed ... and so on, the process perhaps continuing until atomic perspectives (foundational analytic practices?) are reached. On this view some perspectives may contain overlapping objects, but that is always a sign that the perspectives are not atomic and may be decomposed further.

Counterexamples: Strikeouts for Instance

Unfortunately even OHCO-3, the weakest version of the OHCO thesis (being the weakest constraint it allows the most varied and baroque text structures) does not seem immune from counterexample. The following have been proposed as examples of objects which can overlap with themselves -- and so cannot plausibly be teased into different analytical sub-perspectives.

- Text critical objects such as strike outs and variant readings [12]
- Narrative objects such as stories
- Reference structures such as hypertext link anchors and targets
- Poetic objects such as tropes and allusions
- Discourse objects such as topics
- Concrete writing objects such as sentences forming acrostics
- Linguistic objects such as arbitrary collocations

And as well as overlapping objects there are discontiguous objects, which are also non-hierarchical. Examples are:

- Lists broken across paragraphs
- Songs or choral odes broken across other text

In some of these cases there may be ways to defend OHCO-3 and maintain that in fact there is a stable hierarchical structure that is being misunderstood. For instance, any apparent case of overlapping instances of the same object can of course be challenged by proposing a more fine-grained classification scheme. However we suspect that in most cases it will be immediately evident that the granularity necessary to distinguish such objects, so that they can be assigned to different perspectives, will not plausibly correspond to alternative analytical perspectives, but rather only to a distinctions that are present within the text as seen from a particular analytical perspective. For variety we illustrate this with a non-textual example: tonal objects such as keys can overlap in modulatory passages, sharing notes or chords -- but as there is not, within music theory, a separate analytical perspective for every key, one consequently cannot plausibly avoid this overlap by assigning such objects to different perspectives (and therefore different hierarchies).

The difficulty and subtly of these defenses suggest that may be time to reexamine our initial provisional concern to defend hierarchy. We have retreated from saying that texts are hierarchical, to saying that perspectives are hierarchical, to saying that perspectives can be decomposed into hierarchical sub-perspectives. There are indeed enormous advantages to be had if we can approach our subject matter in hierarchical units, but these advantages, and even the fact that hierarchies seem prevalent and important, should not deter us from allowing at least as many things into our theories of text encoding as there are in the world.[13]

Practical Problems with the OHCO Thesis

Even under OHCO-3 situations arise which violate simple hierarchical notions. A number of strategies are used to represent such cases with standard SGML techniques, but none of these is completely satisfactory (Barnard, et al. 1988).

The most radical is simply to pick a single hierarchy as the 'real' document hierarchy, and flatten all other hierarchies. This is accomplished by a variety of methods generally involving the use of zero-width tags (NULL-content tags in SGML terminology). Page numbers, canonical reference schemes and similar sequences of objects that cover an entire document are often represented as 'milestone' tag. The TEI defines milestones as point labels that give a page number, or other reference information. If the flattened hierarchy is more complex some extra information must be implied by the zero-width tags. For example, a text-user, or processing software, might have to know that a 'chapter' milestone terminates any outstanding 'section' milestones. However because strictly speaking these elements have no scope or content the SGML mechanisms for indicating such syntactical relations cannot be employed.

Less violent to the objects represented is the use of CONCUR. This feature of SGML has, unfortunately, rarely been implemented, but does allow for several parallel hierarchical decompositions of a text. It also tends to create somewhat cumbersome and verbose markup. When actually tagging texts, however, CONCUR seems not to represent the text structure faithfully, despite its apparent descriptive adequacy. CONCUR-style markup requires that each hierarchy contain the entire document. Some hierarchies do not seem to stand on their own -- for instance, a text with metrical markup and no other tags would not be generally useful; it might even be considered incoherent as a representation. But CONCUR implies that such breakdowns are useful, by requiring each set of overlapping items to be represented as a complete hierarchy. These inadequacies are not surprising as CONCUR was not designed to enable the representation of multiple logical views. It was designed specifically to allow the results of formatting to be coordinated and represented in the same file as the source document.

The notion of multiple 'logical' views is generally absent from SGML. Its awkwardness in coordinating multiple hierarchies is such that Goldfarb himself has stated: 'I therefore recommend that CONCUR not be used to create multiple logical views of a document, such as verse-oriented and speech-oriented views of poetry' (Goldfarb, 1990, p. 304).

Finally, some non-hierarchical structures can be represented using the tag structure known as a 'span' in the TEI. Spans are zero-width tags that delimit the starts and ends of non-hierarchical structures. The start and end tags are linked to each other by explicit cross references. Given this technique, spans are even capable of handling the multiple-strikeout case mentioned above, as well as being the only structure that can represent hypertext links in their full generality. The qualitative analysis community has been using spans as the only form of markup for computer-aided ethnographic and anthropological analyses -- providing further support for the contention that hierarchies, while common and useful, are not inherent to all perspectives (Miles and Huberman 1984).

Conclusion

The foregoing analysis seems fundamentally sound in that analytical perspectives do seem to exist and do seem to provide fundamental insights into the nature of texts and the methodology of text encoding. And although we have retreated from the simple OHCO thesis, we note that the spirit of the OHCO hypotheses is borne out to the extent that texts qua intellectual objects still seem to be composed of structures of meaning-related features and that, moreover, these structures are often hierarchical.**[14]**

So we have the following positive conclusions:

1. Perspectives -- theories, methodologies, and analytical practices -- are at least as important as genre in the identification of text objects **[15]**
2. Perspectives frequently determine hierarchies of objects
3. Non-hierarchical perspectives can often be decomposed into hierarchical sub-perspectives

   But we note:

4. Perspectives do not always determine hierarchies
5. Non-hierarchical perspectives cannot always be decomposed into hierarchical sub-perspectives

The theory of text and text encoding methodology is still in a rudimentary state; we hope the concepts discussed here contributed to the groundwork for further discussion.**[16]**

Bibliography

Barnard, D., Hayter R., Karababa M., Logan G., and McFadden, J. (1988), 'SGML-Based Markup for Literary Texts: Two Problems and Some Solutions', *Computers and the Humanities* 22: 265-276.

Barnard, D. T., Fraser, C. A., and Logan, G. M. (1988), 'Generalized Markup for Literary Texts', *Literary and Linguistic Computing*, 3.1: 26-31.

Chamberlin, D. D., Hasselmeier, H. F., Luniewski, A. W., Paris, D. P., Wade, B. W., and Zolliker, M. L. (1987), 'Quill: An Extensible System for Editing Documents of Mixed Type'. In *Proceedings of the 21st Hawaii International Conference on System Sciences*. Washington, DC: IEEE Computer Society Press.

Coombs, J. H., Renear, A. H. and DeRose S. J. (1987), 'Markup Systems and the Future of Scholarly Text Processing', *Communications of the Association for Computing Machinery*, 30: 933-947.

DeRose, S. J., Durand, D. G., Mylonas, E., and Renear A. H. (1990), 'What is Text, Really?', *Journal of Computing in Higher Education*, 1.2: 3-26.

Fraser, C. A. (1986), *An Encoding Standard for Literary Documents*, M.Sc. Thesis, (Queen's University, Kingston Ontario).

Goldfarb, C. (1981), 'A Generalized Approach to Document Markup', in *Proceedings of the ACM SIGPLAN--SIGOA Symposium on Text Manipulation*, (New York: ACM).

Goldfarb, C. (1990), *The SGML Handbook*, (Oxford).

Huitfeldt, C., and Rossvaer, V. (1989), *The Norwegian Wittgenstein Project Report 1988*, (The Norwegian Center for the Humanities, Bergen).

Huitfeldt, Claus (1992), *MECS -- A Multi-Element Code System*, Working Papers from the Wittgenstein Archives at the University of Bergen, No 3, (seen in draft in October 1992).

Huitfeldt, Claus, (1992), 'Multi-Dimensional Texts in a One-Dimensional Medium', paper presented to Wittgensteinseminara i Skjolden, May 1992, (seen in draft in September 1992).

Koo, R. (1989), 'A Model for Electronic Documents', *Special Interest Group on Office Information Systems (SIGOIS) Bulletin*, 10.1.

McKerrow, R. B. (1927), *An Introduction to Bibliography for Literary Students*, (Oxford).

Miles, M. B. and Huberman A. M. (1984), *Qualitative Data Analysis, a Sourcebook of New Methods*, (Sage).

The Chicago Manual of Style (1982), (Chicago: University of Chicago Press) 13th ed.

International Organization for Standardization (ISO), *Information Processing -- Text and Office Systems -- Standard Generalized Markup Language (SGML)*, ISO 8879-1986, International Organization for Standardization (ISO) 1986.

Quine, W. v. O, (1953), 'On What There Is' in *From a Logical Point of View* (Cambridge: Harvard University Press).

Reid, B. (1980), 'A High-Level Approach to Computer Document Formatting'. in *Proceedings of the 7th Annual ACM Symposium on Programming Languages*, (New York: ACM).

Renear, A. (1993), 'Representing Texts on the Computer: Lessons from and for Philosophy', *Bulletin of* *the John Rylands Library* (forthcoming in 1993).

Rohr, P. (1991), 'The TextBase Paradigm: Architectural Considerations for a Second-Generation Scholar's Workstation', Senior Thesis, University of Chicago, (seen in draft).

Smith, J. (1987), 'The Standard Generalized Markup Language (SGML) for Humanities Publishing', *Literary and Linguistic Computing*, 2.3: 1971-75.

Sperberg-McQueen, C. M. (1991), 'Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts', *Literary and Linguistic Computing*, 6.1: 1991.

TEI (1990), *Guidelines for the Encoding and Interchange of Machine-Readable Texts*, C. M. Sperberg-McQueen and L. Burnard, eds. (Chicago and Oxford: TEI).

Notes

Note 1 Our thinking on the topics in this essay owes much to conversations and collaborations, spanning many years, with Geoffrey Bilder, Lou Burnard, James H. Coombs, Steven J. DeRose, Claus Huitfeldt, W. Richard Ristow, Michael Sperberg-McQueen, and the members of CHUG, the Brown University Computing in the Humanities Users' Group. It also, like so much similar work being done today, owes an enormous debt to the Text Encoding Initiative, and the TEI's sponsoring and funding organizations -- in the course of pursuing its principal goal of making machine-readable texts more useful and influential the TEI has along the way created a wonderfully rich and exciting environment for thinking about technology, text, and the humanities. [Back to text]

Note 2 Among the encoding projects to which we are are particularly indebted for empirical insights into the methodology of encoding are the Perseus Project, the Electronic Peirce Consortium, the Bergen Wittgenstein Archives, and the Brown Women Writers Project. In what follows we assume a basic familiarity with text encoding and, more specifically, some understanding of SGML and the TEI Guidlines (TEI 1990). [Back to text]

Note 3 Sperberg-McQueens's other two axioms also help put the controversy of the 'interpretative' nature of markup into perpective: Axiom 2: 'One's understanding of text is worth sharing'. Axiom 3: 'No finite markup language can be complete' (Sperberg-McQueen 1991). This paper is a notable exception to the lack of theoretical work on the methodology of text encoding. [Back to text]

Note 4 In the terminology of graph theory ordered hierarchies are 'ordered, rooted trees'. In linguistic theory the ancestral and ordering relations are often separately described as 'dominance' relations and 'precedence' relations. [Back to text]

Note 5 Elsewhere one will find 'editorial', 'logical' and 'sense' used to mean more or less the same thing we mean here by 'content'. Words used in place of 'object' by other authers include 'part', 'component', , and 'element'. 'Element' is in fact the technical SGML term that corresponds to our use of 'object' -- however we continue to use the word 'object' as name for a pre-theoretical notion that may or may not be adequately captured in the technical vocabulary of SGML. [Back to text]

Note 6 For further discussion of these alternative models, and the comparative advantages of the OHCO model, see Coombs, et al. 1987 and DeRose, et al. 1990. [Back to text]

Note 7 The classic presentation of the view that our ontological committments, whether metaphysical or scientific, are to be determined by examining the denoting phrases of the relevant theoretical statements is in Willard van Orman Quine's, 'On What There Is' (Quine 1953). 'To be' quips Quine, 'is to be the value of a bound variable'. In one form or another this criterion for 'ontological committment' has been adopted by many philosophers of science. [Back to text]

Note 8 Obviously a thorough discussion of this argument will require distinguishing and relating a variety of objects: work and edition, type and token, etc. But the crude presentation given here should be adequate to suggest the intuitions behind the argument from variation. [Back to text]

Note 9 There is a third alternative of course: one might claim that the project -- saying 'what text is' -- is somehow muddled or incoherent, an improper question of some sort. [Back to text]

Note 10 This view does not imply that these structures are, or are not, "absolute" or "objective" in any significant sense philosophical sense. [Back to text]

Note 11 Two objects, A and B, overlap when neither of the objects contains all of the contents of the other object. [Back to text]

Note 12 In a talk at the Brown University Computing in the Humanities Users Group in January 1991, Claus Huitfeldt, Director of the Bergen Wittgenstein Project, stopped at one point in his presentation and wrote three words on the board. He drew a strike-out line through the first and second words, and then a second strike-out line through the second and third words. The text encoders present sighed -- this was a vivid example of the vulnerability of the OHCO view of text. [Back to text]

Note 13 Initially we could dismiss cases where, for instance, sentences might break across paragraphs or chapters (as they might in some sort of avant garde writing) as being too cooked-up to be considered. Now, in light of the above examples, they seem only one more piece of evidence that the relevant logical structures of a text are not without exception hierarchical. Yet we note that the force and effect of such a device comes from its being an anomaly in what otherwise tends to be a hierarchical structure. In fact, the vivid effect of the cross-perspective overlaps cited earlier (metrical, dramatic, linguistic) seems to owe much to the way they work against the tendency of the logical elements of text to arrange themselves in hierarchies. [Back to text]

Note 14 The existence and importance of non-hierarchical text-descriptions have been noted by Paul Rohr, in an as yet unpublished paper (Rohr 1991). Rohr, starting without the hierarchical bias imparted by traditional text encoding schemes, and basing his work on deconstruction and modern literary theory, proposes a completely non-hierarchical notion of textual markup. [Back to text]

Note 15 In fact, the genre-based analysis of a text is probably best treated as a special case of analysis from an analytical perspective. [Back to text]