Balisage: The Markup Conference 2009

Proceedings

Documents Cannot Be Edited

Allen H. Renear

Associate Dean for Research and Associate Professor
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
<renear@illinois.edu>

Karen M. Wickett

Doctoral Student
Graduate School of Library and Information Science,
University of Illinois at Urbana-Champaign
<wickett@illinois.edu>

Abstract

Most definitions of document current in the document processing and digital publishing communities would, if taken literally, imply that documents are extensional entities that cannot undergo changes such as editing or revision. In other domains as well, such as textual criticism and library science, one can also find notions of text or document that are similarly difficult to reconcile with modification. We describe the problem and sketch some possible resolutions. Although the issues are conceptual and foundational the practical significance is real. Formal representation in logic-based ontology languages, increasingly important in information management, requires that familiar idioms, however serviceable and entrenched, be converted to expressions that support literal interpretation.

Table of Contents

The Problem

Document modification seems to be routine and widespread. Editing is a familiar practice to almost everyone, and revision a fundamental feature of publishing workflows. Yet document modification would appear to be an illusion. Common accounts of what documents are seem to imply that documents cannot undergo genuine modification.

The W3C XML specification defines an XML Document as a string that meets certain formal constraints. Strings are mathematical constructs that are defined, ultimately, in set theoretic terms. They are therefore purely extensional entities constrained by the axiom of extensionality, common to all standard set theories. As a consequence, strings have all of their non-relational properties essentially and cannot be altered or modified in any way. Although we can of course identify functions that map one string to another, the existence of such functions does not, by itself, provide an explanation of what constitutes the modification of a document.

This consequence is not unique to the definition of a document as a string. Familiar alternative definitions fare no better. Documents (in the relevant sense) have been defined as graphs, relations, Ordered Hierarchies of Content Objects (OHCO) [derose90], and sentences in formal logic [renear06]. But these are all also extensional entities: graphs, relations, tuples, and strings all have mathematical definitions in virtue of which they are kinds of sets. Nor is the problem particular to formal definitions that make explicit use of mathematical constructs. Many of the concepts of document (or text) used in library science, textual criticism, aesthetics, and other fields appear, upon examination, to be similarly problematic.

A Simple Example: The Verona Sentence

The problem can be introduced with a simple example that makes no explicit use of mathematical or philosophical notions [renear08].

Consider the sentence “I remember Verona.” Let this be the first sentence of the first chapter of a draft of a novel.

Suppose that the author decides to edit that sentence and revises it to read: “I remember, but dimly, Verona.”

It is natural to say that the first sentence of the chapter has been changed, that it is now longer. But exactly *what* has been changed? *What* has become longer? The new first sentence, “I remember, but dimly, Verona”, has not changed. That sentence has never consisted of fewer than five words. The original sentence, “I remember Verona”, has not changed either. It is not now longer than it was; it still consists of three words. It is true that “I remember, but dimly, Verona” is a longer sentence than “I remember Verona”, but it did not become a longer sentence than “I remember Verona” — it has always been a longer sentence than “I remember Verona”.

One might attempt to address the problem by shifting the scope of attention and propose that it is the text of the chapter that has changed. However, the chapter as a whole is simply another, if longer, textual entity, and has the same identity conditions. So the puzzle will arise again. The chapter has been revised, but neither the new text nor the original text undergoes any change during this process. And the same argument may be made for the entire text of the novel.

In short: While it is natural to speak of sentences or other textual entities changing when they are edited or revised, it appears that these entities themselves do not really change. So what does change?

The Verona puzzle may feel like a parlor trick, but the significance is real enough: we do not have a clear conceptual understanding of what is happening when documents are modified. Of course truth values can be correctly assigned to modification sentences such as "The sentence was changed", but only if we treat these sentences as idioms, as we do sentences like "The average plumber has 3.2 children". Such sentences do not have compositional semantics or support existential instantiation and there is often little systematic guidance for their interpretation. Up until now we have largely avoided these problems, relying on ad hoc solutions and human intervention. But the increasingly formal nature of new semantic approaches to information management and publishing, and the continuing minimization of human intervention, will inevitably require us to more systematically develop robust literal interpretations of fundamental concepts.

An Inconsistent Triad

Consider the following three assertions:

> All documents are strings.
>
> Strings cannot be modified.
>
> Documents can be modified.

As any two of these assertions will together logically imply the negation of the remaining assertion it is not possible for all three to be true. And yet each has some initial plausibility.

In favor of the first assertion, that all documents are strings, we begin by observing that in the definition of an XML Document in the W3C XML specification the first clause has as a consequence that XML Documents are strings:

> Definition: A textual object is a well-formed XML document if: Taken as a whole, it matches the production labeled document. It meets all the well-formedness constraints given in this specification. Each of the parsed entities which is referenced directly or indirectly within the document is well-formed.

Understanding documents (or a relevant sense of "text") to be sequences of characters or words is also consistent with approaches in library cataloguing [frbr98] and textual editing [tanselle89].[1]

In favor of the second assertion it may be argued that modification of an entity necessarily involves the loss of a property and that strings have no properties which it is possible for them to lose and survive the loss. [We assume losing a property and gaining the complement of that property are equivalent characterizations of the same event.] The string "13571" has the property of having a length of five tokens, the property of having one token type occur twice, and the property of having the substring "35". But these are all properties that "13571" cannot lose. That is, the string in question, "13571", which has property of having "35" as a substring, cannot at some point in the future lose the property of having "35" as a substring. There is no plausible entity that will serve as the reidentifiable persistent object that could undergo such a change. Cf. the Functional Requirements for Bibliographic Records: "...if a text is revised or modified, the resulting expression is considered to be a new expression, no matter how minor the modification may be" [frbr98] [2]

In favor of the third assertion, that documents can be modified, we simply observe that this is an assumption that for most of us is so deeply entrenched in our common understanding that the title of this paper probably seems more senseless than provocative.

Responses

One sort of response to a triad claimed to be inconsistent is to deny the inconsistency. Typically this takes the specific form of claiming that one of the assertions in the triad has two possible interpretations. Interpreted one way the assertion is true but the triad consistent. When the assertion is interpreted in the other way the triad becomes inconsistent, but the assertion is no longer plausible. When this is the situation the problematic nature of the triad is an illusion created by trading on this ambiguity.

As modal locutions well-known for generating ambiguity and paradox are evident in our triad we shall now indicate exactly how the English sentences are to be interpreted and confirm that given the intended interpretation the triad is in fact inconsistent. We do this by expressing the assertions in elementary predicate logic.

(x)[(isaDocument(x) -> isaString(x)]

(x)[(isaString(x) -> ~isModifiable(x)]

(Ex)[(isaDocument(x) & isModifiable(x)]

On the standard interpretation of quantifiers and connectives these three formulas clearly form an inconsistent set.

Once the inconsistency of a triad is granted the remaining responses are typically classified according to which assertion in the triad is rejected. We will consider responses that reject the first and third assertion; we do not here consider responses that deny the second assertion.[3]

Responses that Deny Documents are Strings

These responses reject the standard definitions of a document as a kind of string or relevantly similar entity, such as graph, relation or OHCO. For this to be a plausible response, a suitable alternative definition must be proposed.

The Materialist Strategy: This strategy holds that a document is not a string, but a concrete arrangement of a quantity of matter and energy. On this view modification of a document does literally occur: modification consists in physical changes to the material document, with the document preserving its identity across these changes (otherwise it would be destroyed rather than modified).

Against the Materialist Strategy: The identification of a document with a particular concrete arrangement of matter and energy instead of a string appears inconsistent with many of the things that we say about documents. For instance we speak of the document even when there may be many physical instances, intending not to refer to any one of them, or the set, but rather to something they all represent or instantiate. Whether this apparent reference to an abstract object can be paraphrased away remains promissory. Documents certainly have physical representations, and these representations are often materially involved in scenarios of putative document modification. But whether a document can be defined as one of, or even the class of, these separate individual representations remains to be seen.[4]

The Social Object Strategy: This response posits as the modifiable document a social object which is constituted by (but not identified with), one string at one time and another string at another time, the changes being determined by social (including institutional and linguistic) circumstances. This approach abstracts from the physical and is consistent with the common belief that documents can be modified. The document presumably maintains a coherent identity across various textual changes and may be associated as well with any number of different physical representations. A theoretical basis for this strategy can be found in John Searle's work [searle95] although Searle's theory appears to be primarily one of natural facts that in certain social circumstances "count as" social facts rather than natural objects that in certain social circumstances count as social objects. Barry Smith has defended a theory of social objects that allows social objects to exist even without natural objects that "count as" or constitute them [smith03]. Another related perspective is Brian Cantwell Smith's notion of holding objects in the "middle distance" [cantwellsmith96]

Against the social object strategy: The social object strategy preserves the intuition that documents are modifiable, but at a considerable cost. First, it requires an ontologically challenging entity, social objects that cannot be identified with physical objects, abstract objects, or even mental states. There is also a corresponding new distinctive metaphysical relation as well: constitution, the relationship that obtains between the social object and the different strings that constitute that object at different times. That this relation cannot be simple identity is evident from the fact that it is not transitive and symmetric: while the document qua social object may be constituted by one string at one time and a different string at another time, this does not imply that the two strings are themselves identical. Searle's notion of "counting as", if applied to social objects, may avoid some of the traditional problems with constitution, but it will still be hard to reconcile social objects with a naturalistic view of the world.

Responses that Deny Documents can be Modified

These strategies accept the first assertion of the triad, that a documents is a string, and reject the third assertion, that documents can be modified. If such an approach is to be plausible it must provide a convincing alternative account of what is happening in the situations which we would ordinarily describe as "modifying a document".

The title of this paper notwithstanding, denying document modification does not necessarily require claiming that sentences like "John edited a document" never express facts about the world. These sentences may be considered idioms. As such they sometimes communicate true assertions, but they are not literally true. Consider again the sentence "The average plumber has 3.2 children". This sentence might indeed be used to make a true assertion. And if it does make a true assertion then it is certainly reasonable to say that the sentence is a true sentence. However we would not conclude from the truth of the sentence "The average plumber has 3.2 children" that there is therefore an entity in the world which is the average plumber and which actually has some fractional number of children, as a naive Russellian interpretation of the definite description would entail. What is being denied in the rejection of the third assertion of the triad is not the truth of sentences like "John edited a document", but rather that such sentences imply a claim such as the one suggested by the first order formula:

$$(Ex)[(isaDocument \ \& \ edited(john,x)]$$

The New Document Strategy: This response maintains that the modification of a document is actually the creation of a new document.

Against the New Document Strategy: The new document strategy has new strings created with each document modification. This seems peculiar, contrary to the general notion of strings, and generally difficult to reconcile with a naturalistic view of the world. Consider the string “13571”. Was that string created? How did that happen? And when? Can it be destroyed? Can it be re-created? Unless strings are material objects the conceptual apparatus of creation and destruction seems entirely metaphorical.

The Selection Strategy: This approach holds that in a typical scenario of alleged modification a new but already existing string is selected for attention. When we say that a document has been modified we mean that a different string has been selected for the purpose at hand by some particular person or persons. The physical infrastructure of analog or digital document development and publishing, in combination with social conventions and practices, is a system for recording and communicating which string is currently distinguished in this way.

Against the Selection Strategy: The selection strategy identifies documents with already existing strings. But this seems strange. Either these already existing strings came into existence at some point in the past or they have always existed. If the former then this strategy has no advantages over the new document strategy; it must still somehow account for how strings can come into existence, from what materials and in what causal circumstances. But if the latter, if strings have always existed, then assuming only that the further requirements for being a document are non-contingent (such as matching a production) we will have documents, and not just strings, existing eternally — apparently even before cognitive agents. This seems peculiar.

A Strategy that both Redefines Document and Denies Modification

Strictly speaking the strategy we are about to take up denies only the third assertion of the triad (documents can be modified). However because it proposes a substantially new definition of document it has much in common with strategies that deny the first assertion and so we are placing it in a separate category.

The String-in-a-role Strategy: This strategy holds that a document is a string in a particular communicative role. The string itself may be an uncreated and pre-existing entity, but the strings which are documents need not always have been documents. Being a document is a property that strings have only in particular contingent social/linguistic situations. So on this account documents have not always existed, even though documents are strings, and strings have always existed. This is because while a string which is a document has always existed, that string has not always been a document — a string becomes a document only in the appropriate social circumstances.

Although this strategy, like the strategies that reject the first assertion of the triad (documents are strings) involves a new approach to the definition of document, it affirms rather than rejects the first assertion of the triad. After all, if a document is a string in a communicative role, then a document is a string. However, the string-in-a-role definition of document is unlike definitions such as the one in the XML specification in that it places a contingent rather than necessary constraint on strings satisfying the definition. If a document is defined as a string with certain purely formal constraints (as it is in the XML specification, where it must match a particular production in a grammar), then the things which are in fact documents are documents necessarily. This is because not only is it impossible for a string to cease to be a string or have been anything other than a string, it is also impossible for a string that matches a particular production to have ever failed to match that production or to fail to match that production in the future.

According to the string-in-a-role strategy the property of being a document is what Guarino and Welty refer to as a "non-rigid"[1] property [guarino00]. Guarino and Welty define rigidity using modal logic and model theory, but the basic idea is simple: a property is rigid if and only if nothing that has that property could have failed to have that property, or could come to lose that property (and still exist). For example, being a person is rigid because the things that are persons could not have been anything but persons and cannot cease to be persons (although they can cease to be). But being a student is not rigid because the things that are students (i.e. persons) might not have been students and can cease to be students (without ceasing to be). According to Guarino and Welty rigid properties indicate types, fundamental kinds of things, while non-rigid properties indicate roles that things of some particular type may enter into. On this view a document is not a type of thing, but a role that things of some type or other have in particular contingent circumstances.[5]

The redefinition of document as a "string-in-a-role" is not itself a response to the inconsistency of the triad; if a document is a string-in-a-role then it is still a string, and strings cannot change. The string-in-a-role strategy rejects the third assertion, and denies that, strictly speaking, documents change.

Concluding Remarks

There is much more to be said pro and con on the strategies we have described here, and there are possibly better strategies to consider. We have intended only to suggest some of the possibilities. Formal representation and inferencing is increasingly widespread and increasingly important, and systematically making information computationally available in logic-based ontology languages requires literal interpretation. Human beings may effectively communicate with natural language sentences such as "The document was edited" or "The average plumber has 3.2 children". But to support inferencing in the semantic web environment these idioms must be represented formally in languages that rely on compositional semantics, existential instantiation, and valid deductive consequences.

Or... *perhaps not.*

One has to wonder whether all this logic-chopping subtlety is going to be worth it. It isn't clear exactly how to finish the job, and yet it is clear that some of our most familiar — and effective — ways of conceptualizing our domain will be revised if we continue along this path. Particularly troublesome is the prospect that the revisions anticipated will add not only complexity in design and use, but increase computational complexity as well. As we have suggested elsewhere "denormalized" ontologies may be more appropriate for much of the practical work ahead.

Acknowledgements: This paper draws on work carried out collaboratively at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Much of the early work was organized by Dave Dubin and sustained by various research groups.

Bibliography

[buckland97] Buckland, M. K. "What is a "Document*?" Journal of the American Society for Information Science*, 48 (1997),804-809. doi:10.1002/(SICI)1097-4571(199709)48:9%3C804::AID-ASI5%3E3.0.CO;2-V

[1] Correction: "rigid" to "non-rigid". June, 2010.

[cantwellsmith96] Cantwell Smith, B. *On the Origin of Objects*. Cambridge: MIT Press, 1996.

[derose90] DeRose, S. D., Durand D. G., Mylonas, E., Renear, A. H. “What is Text, Really?” *Journal of Computing in Higher Education* 1 (1990), 3-26. Reprinted in ACM/SIGDOC *Journal of Computer Documentation* 21 (1997).

[guarino00] Guarino, N. and Welty, C. A. “A Formal Ontology of Properties.” In *Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management*. Lecture Notes In Computer Science. Springer-Verlag, (2000), 97-112.

[frbr98] International Federation of Library Associations (IFLA). *Functional Requirements for Bibliographic Records: Final Report*. UBCIM Publications-New Series. Vol. 19, München: K.G.Saur, 1998.

[renear03] Renear, A. and Dubin, D. “Towards Identity Conditions for Digital Documents.” In *Proceedings of the 2003 International Conference on Dublin Core and Metadata Applications* (Seattle, Washington, September, 2003). International Conference on Dublin Core and Metadata Applications. Dublin Core Metadata Initiative, 1-9. 2003.

[renear06] Renear, A. H. “Is an XML Document a FRBR Manifestation or a FRBR Expression? — Both, Because FRBR Entities are not Types, but Roles.” in *Extreme Markup Languages 2006 Proceedings*, (Montreal, Canada. 2006).

[renear07] Renear, A. H., Dubin, D. “Three of the Four FRBR Group 1 Entity Types are Roles not Types.” In *Proceedings 70th Annual Meeting of the American Society for Information Science and Technology* (Milwaukee, WI, October 2007).

[renear08] Renear, A. H. and Dubin, D., and Wickett, K. M. “When Digital Objects Change — Exactly What Changes?” In *Proceedings 71st Annual Meeting of the American Society for Information Science and Technology* (Columbus, OH, October 2008).

[searle95] Searle, J. R. *The Construction of Social Reality*. New York: The Free Press, 1995.

[smith03] Smith, B. “John Searle: From Speech Acts to Social Reality,” in *John Searle*, B. Smith (ed.) Cambridge University Press, 2003.

[tanselle89] Tanselle, T. G. *A Rationale of Textual Criticism*, Philadelphia: University of Pennsylvania Press, 1989.

Notes

[1] In this analysis we are using string based definitions of documents as a proxy for a broader class of definitions, including those that define a document as a kind of graph, a kind of relation, or an “Ordered Hierarchy of Content Objects.” So what follows from identifying documents as a kind of string is intended to also follow from definitions that imply that a document is a graph, a relation, an OHCO, or any other relevantly similar entity.

[2] It might be argued that “13571” does have the property of “being thought about by someone” (at this moment, as you read this paper), and that this is a property that it can lose. But the loss of properties of this sort (relational properties) does not seem to be genuine modification. We do not say that the second-tallest person in the room has undergone a modification when he becomes the tallest person in the room not in virtue of getting any taller, but in virtue of the previously tallest person in the room leaving the room. Genuine modification requires the loss or gain of a non-relational property. But strings, sets, relations, graphs, and such things have no non-relational properties they can lose: all of their non-relational properties appear to be in some sense essential to their identity.

Furthermore, even if we allowed that the loss or gain of a non-relational property was a genuine modification, this would not appear to be much help with the larger problem, as editing changes would in any case seem to be changes in the inherent and non-contingent properties of a string, and not its relational properties.

[3] Responses to an inconsistent triad of plausible assertions usually reject just one assertion. It is true that combinatorially there are seven possible combinations of assertions that might be rejected, but solutions rejecting more than one assertion bear a prima facie higher burden of defense and consequently are rare. We do not consider any such response here, although the final response we discuss qualifies, without actually rejecting, the first assertion as well as denying the third assertion.

[4] Although we describe this response as “materialist” because it asserts that documents are material things rather than strings, materialism *per se* neither entails, nor is entailed by, the materialist strategy described here. The materialist strategy can be adopted by the non-materialist who holds that while there are non-material things, documents are not among them. More significantly a materialist may decline the materialist response, allowing that documents are strings, but holding that strings are material things, choosing another assertion from the triad to reject. This last observation reveals that the materialist response requires the elimination of strings (from the definition). Mere reduction of strings to material things does not provide, in itself, any response at all to the inconsistency.

[5] A result in some respects similar to one that was reached in response to a different puzzle about XML documents. [renear06] [renear07].

Author's keywords for this paper: document; text; XML; ontology