Towards Identity Conditions for Digital Documents

Allen Renear, David Dubin
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
{renear, ddubin }@uiuc.edu

Abstract

By "identity conditions" we mean a method for determining whether an object x *and an object* y *are the same object. Identity conditions are arguably an essential feature of any rigorously developed conceptual framework for information modeling. Surprisingly, the concept of* same document, *which is fundamental to many aspects of library and information science, and to digital libraries in particular, has received little systematic analysis. As a result, not only is the concept of a* document *itself under-theorized, but progress on a number of important practical problems has been hindered. We review the importance of document identity conditions, demonstrate problems with current approaches, and discuss the general form a solution must take. We then describe our own strategy, based on the BECHAMEL XML Semantics Project — we propose to reduce the relatively elusive and undefined general problem of determining document identity to the comparatively well-understood problem of proving logical equivalence in predicate logic. This approach should also enable the determination of semantic relationships other than identity, including similarities and partial identities of various kinds, and will support new strategies in various areas of digital information management, such as preservation, conversion, integrity assurance, retrieval, federation, metadata, and identifiers. Our results complement and extend discussions of the IFLA/FRBR entities (particularly "expression" and "manifestation") taking place in the cataloguing communityt discussions of "resource" taking place in the W3C and Dublin Core communities, and the analysis of similar notions in ontology development for digital libraries, museums and archives. Although our project is still in a preliminary phase, a working inferencing environment, in object-oriented Prolog, has been completed and initial results that confirm our logic-based strategy.*
Keywords*: document, semantics, identity, expression, XML.*

1. Introduction

By "identity conditions" we mean a method for determining whether an object *x* and an object *y* are the same object. Identity conditions thus provide a way to identify and re-identify, to count, and to distinguish. Identity conditions are also referred to as identity "criteria", or, more rarely, "individuation conditions". (The concept should not be confused with the related but distinct notion of "object identity" found in object-oriented programming and modeling.) The specification of identity conditions for any entity intended as more than just a ***façon de parler*** has long been considered a fundamental requirement in logic and traditional formal ontology — as summarized by Quine's slogan "No entity without identity"[26]. More generally identity conditions have recently been advanced as an important feature of any rigorously developed conceptual model or information ontology[14][15].

Surprisingly then, there has been little research on developing formal identity conditions for an entity which is fundamental in library information science: the document, in the sense of an abstract symbolic expression that may be physically instantiated repeatedly and in various media. As a result, not only is this critical concept under-theorized, but progress on a number of important problems — including preservation, conversion, integrity assurance, retrieval, federation, metadata, identifiers — has been hindered.

The development of identity conditions for a particular kind of entity is not something separate from, let alone subsequent to, defining that entity, so we cannot *begin* our development of identity conditions with an explicit definition of what we mean by "document". Nevertheless, a definition of an entity in an ontology is generally a refinement and formalization of a pre-existing informal or "pretheoretic" notion, and so it is appropriate to begin by reiterating our informal characterization of "document" in order to identify just what concept it is that we will be formalizing. By ***document***, then, we refer to the abstract symbolic expression which may be physically instantiated repeatedly and in various media. This use corresponds more or less to the FRBR term "expression"[17] and has the colloquial synonym "text". Although now fairly common, this sense of "document" does compete with another well-established and closely related use of the term (particularly common in the library, archival, and legal communities) to refer to the physical carrier with its instantiated inscription.[5]. Terminological choices in this area are difficult, but we believe our usage is consistent with common and emerging practice in publishing and information science.

In what follows we focus on documents that are represented in an XML vocabulary and digitally encoded, although strictly speaking nothing in our general strategy is particular to either XML or to digital formats. (Because, "XML document," the technical term for a conformant

combination of markup and content — corresponding to SGML's "document instance" — can be easily confused with "document" in our sense, we will sometimes use the phrase "XML representation" to refer to an "XML document".)

We begin our discussion by giving examples of the role that the notion of *same document* has in a number of digital library and document management activities. We then describe a succession of plausible techniques for precisely and reliably specifying document identity. Although each of these techniques is shown to fail, each is progressively more adequate (has a smaller set of counterexamples) then the preceding strategy, and this progressive improvement, corresponding to increases in level of abstraction, suggests the general form of our solution. Unfortunately there is a substantial obstacle to making the final and necessary step in abstraction — the lack of a standard method for providing *semantics* for SGML/XML markup. Without such a semantics the document carried by an XML representation, and all lower levels of abstraction, cannot be reliably recognized or tested for identity.

We then describe an effort to remedy this defect, the BECHAMEL XML Semantics Project, and show how this approach will (i) support of the development of the markup semantics necessary to make the final step in abstraction and (ii) reduce the relatively elusive and undefined general problem of determining document identity to the comparatively well-understood problem of proving logical equivalence in first order predicate logic. This strategy should also enable the determination of semantic relationships other than identity, including similarities and partial identities of various kinds, and will support new strategies in various areas of digital information management. Finally, this approach should illuminate the concept of a document in general, contributing to a long-recommended but only slowly evolving "program of basic research into the nature of documents"[20].

Although this project is still in a preliminary phase, a working inferencing environment, in object-oriented Prolog, has been completed and is producing results that confirm this logic-based strategy.

2. Related Work

Traditional cataloguing theory has a rich history of relevant efforts to develop and refine concepts such as *book* and *work*[39][44]. More recently IFLA, as mentioned above, has presented a compelling and influential framework (work/expression/manifestation/item) for bibliographic entities[17]; other similar, and perhaps competing, ontologies in publishing, digital libraries, and cultural informatics are also important and relevant[11][16][18][19][22]. At present we do not yet respond directly to these discussions, but are working out the consequences of our own approach. However, to take one example, our notion of document seems sufficiently close to FRBR's notion of "expression" that, if we are successful, we will have in effect have provided explicit identity conditions for the FRBR *expression* entity.

Within the W3C efforts to develop document surrogates have produced standards for normalizing the XML representation, identifying the data structure independently of any particular serialization[4][9]. These also are important and relevant, but as we demonstrate below, this strategy will not provide document identity conditions.

The lack of a semantics for XML has been criticized since the 1980s[28] and has even been presented as a kind of a crisis in the effective use of XML/SGML vocabularies in textual criticism and literary computing[6] (we disagree with this extreme assessment[33] even though arguing the benefits of a formal semantics for XML in a wide range of applications[32]). Recently a number of technologies, standards, and research projects have recognized and responded to the challenges and problems of developing an XML semantics; particularly promising among these are [36][37][38][45][46]. We note that Raymond, Tompa, and Wood were early and eloquent both in their criticism of the failure of SGML markup systems to achieve an optimal level of abstraction in general[28][29], and in identifying the need for identity conditions specifically[29][35].

Standards such as W3C Schema, ISO Topic Maps, the DOM and HyTime architectural forms address limited aspects of some of the problems we are concerned with, but they don't provide complete or systematic solutions.

The W3C's "Semantic Web Activity" is obviously relevant and indeed has aspects in common with our project. But the overall agenda of that effort is to equip the web with knowledge representation technologies that model the semantics of document *contents*, while our work investigates the semantics of document *structures,* focusing on understanding existing document markup vocabularies rather then developing new general purpose knowledge representation systems. Among other things this difference makes our approach more appropriate for addressing the specific problem of document identity. However we can exploit Semantic Web formalisms (e.g. DAML/OIL, OWL, RDF, RDFS) for general interoperability.

Over the last few years there has also been discussion in the Dublin Core and W3C communities, and elsewhere in the digital library research community, about the nature of "document-like-objects," "digital objects", and "resources", and related entities[1][2][3][4][25]. Again, while we do not directly address these discussions we believe that our work may eventually contribute to illuminating these difficult topics.

The general approach presented in this paper is an application of the BECHAMEL XML Semantics Project, led by Sperberg-McQueen (W3C/MIT) and co-hosted at the HIT Center, University of Bergen (local lead Claus Huitfeldt) and the Electronic Publishing Research Group, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. BECHAMEL's general goals are to explore representation

and inference issues in document markup semantics, survey properties and relationships in popular markup languages, and develop a formal, machine-readable declarative representation scheme in which the semantics of a markup language can be expressed. The early Prolog inferencing system[41] has been developed into a prototype knowledge representation workbench for representing facts and rules of inference about structured documents.[12] Preliminary findings from BECHAMEL have been reported elsewhere[13][31][41][42][43].

3. Identity Problems

In this section we take a look the role that document identity plays in several areas of digital libraries and information management, and indicate the problems, practical and theoretical, that arise from an adequate conceptualization of document identity conditions.

Preservation. Digital preservation strategies obviously require the concept of *same document*. A conceptual framework for digital preservation must indicate *what* document has been archived or preserved, distinguishing it from other similar documents, but identifying it across various preservation structures each of which presumably preserves *that* document. Specifically any preservation scheme must answer the practical question whether or not a particular document presented for preservation has already been preserved. In the absence of a theory of document identity, preservation strategies typically (although not always[23][24]) fall back on treating the bitstream as a surrogate for the document. Some of the problems with this approach are well-known. The subsequent recovery and presentation of the document from an archived bitstream is problematic and there is no theoretically sound way to tell *what* document has been preserved or whether two preservation structures preserve the *same* document. In the case of migration-based preservation strategies, which require regular conversion (see below), these problems are permanent features of the preservation environment. This situation is at least in part a consequence of not having an adequate conceptualization of what a document is, and more specifically, not having practical method for determining document identity.

Conversion. A conversion is a re-encoding, into a different encoding scheme, of the same document. However, without a system for the explicit representation of the document *per se* it is not possible to rigorously design conversion systems, to define their success conditions, or to verify correctness, confirming whether or not the input and output encodings do indeed carry "the same document". In addition, without a system for representing the document itself it is not possible for the conversion process to take advantage — without ad hoc analysis and programming — of the high-level structure of document objects, which can be shown to provide unique opportunities for effective conversion strategies[13].

Cataloguing. The history of cataloging has included attention to not only the physical description of items, but also the abstract entities those items directly or indirectly carry. Within some areas of cataloguing theory there has been a progressive movement towards foregrounding the relatively abstract entities carried by physical items, whether the carried entities be texts, pure notation-independent works, or "textual works". While explicit identity conditions for notation-independent works (*works* in the FRBR sense) is obviously a daunting problem, identity conditions for texts (*expressions* in the FRBR sense) should not be out of reach.

Identifiers. Determining what exactly our identifiers and descriptors univocally refer to, or should refer to, is a well-known problem in the cataloguing and metadata communities. There are difficulties in developing an adequate theoretical scheme, and there are practical difficulties in applying schemes, in determining whether or not, for instance, two streams of bits are, or identify, the *same resource*.

Copyright. Copyright law makes essential reference to the expression of ideas as the proper object for copyright ownership. While there has been considerable discussion as to what is essential to expression-identity in this sense, there still is no comprehensive theory or rigorous method for determining identity.

Integrity Assurance. Although typically considered specific to information security, assuring document integrity is a central problem in any activity involving text, from digital libraries to routine office work, and is independent of whether or not the changes are malicious ("tampering") or benign (editing). Yet it may plausibly be argued that no contemporary document integrity assurance system actually verifies the integrity of *documents* — they instead take a low-level abstraction, such as a bit sequence, as a surrogate for a document and then verify (perhaps with a system of signatures and keys) the original or anticipated structure of that surrogate. Because, as we will show below, the surrogate-to-document relation is not one-to-one, but (at least) many-to-one, document integrity control by low-level surrogates over-reports possible unauthorized changes, substantially compromising its contribution to information assurance, and undermining any claim to be based on a theoretically sound notion of document identity.

4. Solution Strategies at Inadequate Levels of Abstraction

To illuminate the problem and motivate our particular solution let's consider some common approaches to determining document identity, arranged sequentially so that each is progressively more adequate (having fewer counterexamples) and the direction of the progression as a whole indicates the sort of solution that will ultimately be required.

4.1. Bit Stream Strategies

A common approach is to take the bit sequence as a surrogate for the document it presumably carries — different bit sequence, different document. This approach would yield satisfactory results if the bitstream-to-document relation were one-to-one. However, as we now show, the bitstream-to-document relation is not one-to-one, but rather (or at least) many-to-one. This strategy will therefore indicate as diverse documents that are in fact identical.

In the normal course of document management there are frequent changes to the bitstream that clearly do not affect what document is carried by that bitstream. Such changes have sometimes been called "meaning-preserving" in digital authentication and watermarking research, but we believe "document-preserving" is more appropriate. Obvious examples of document-preserving changes include changes in character encoding (e.g., UTF-8 vs. UTF-16, ASCII vs. EBCDIC), and other transcoding for compression, encryption, transport optimization and so on. Such "document-preserving" changes to the bitstream take place frequently, routinely, and often without users even being aware of them (simply "dragging" a document from one disk to another can silently alter the bitstream). Document identity conditions that rely on bitstream surrogates will therefore under-report identity.

4.2. Character Stream Strategies

Conceptualizing the document as a *character string* immediately improves empirical results and secures substantial practical advantages in many situations. The character sequence "abc" retains its identity regardless of how the characters are encoded.

But it can be easily seen that this higher level of abstraction will still falsely flag document-preserving changes as document alterations. Consider for instance the serialization artifacts of an SGML/XML document, such as attribute order, declaration order (for unique declarations), and nonsignificant "whitespace". Such alternative reserializations of a document occur routinely and should obviously be treated as document-preserving. In addition there are redundancies of various kinds (such as redundant namespaces prefixes and redundant attribute values) and also a variety of logically equivalent expressions (e.g. character references in decimal or hexadecimal); these also are all clearly document-preserving.

In short, changes in the bitstream, the character stream, and the serialization conventions are clearly not changes in the *document* and treating them as if they were compromises functionality and confuses our understanding of our systems and tools.

4.3. Normalized Serialization Strategies

Much more promising is the prospect of exploiting normalized (or "canonical") serializations of the data structure. This can certainly deliver improved results, successfully tolerating even more document-preserving changes, and in many cases providing an adequate test of document identity for a given practical purpose. But there will still be classes of document-preserving changes that are falsely flagged as document-alterations by this approach as well. For instance, many SGML/XML markup vocabularies have alternative markup constructs that clearly (and often explicitly) "mean the same thing"; varying from simple equivalent uses of the element/attribute mechanism (e.g. `<div type=p>` vs. `<p class=div>`) to the TEI's various alternatives for encoding "overlapping" elements — each of which creates a different normalized serialization, and a different corresponding data structure, but all of which still, by definition, represent identical document structures.[40]. So we see that although they are yet another further improvement, normalized serializations, and their corresponding data structures, also fail as document surrogates.

Obviously what is needed is a fresh start, one that is not based on the defensive strategy of progressively eliminating each class of new counterexamples with additional *ad hoc* constraints or incremental improvements in the surrogate representation, but rather an approach based on a *general theory of document identity*.

5. The Way Forward, and an Obstacle

We have seen that improved document identification strategies cannot be developed without (i) a method to identify the document carried by structures at lower levels of abstraction (such as bitstreams, character streams, serializations, or normalized serializations) and then (ii) a method compare the documents themselves, rather than use the lower level abstractions as surrogates. But it is not at all clear how to identify a document carried by a lower level abstraction, or to compare identified documents.

The rudiments of a solution can be found in the late 1980s when, based in part on an analysis of the effectiveness of descriptive markup[8], the notion of a document as a particular kind of conceptual abstraction emerged. This view, which was expressed informally, was that a text (or in our present terminology, a document) is "an ordered hierarchy of content objects" (an "OHCO")[10]. But how can we identify this abstract conceptual document in a way that makes it available to computational processing?

The SGML metagrammar for descriptive markup languages seemed, at least at first, to accomplish this. SGML/XML provides a rigorous technique for expressing machine-readable definitions of descriptive markup languages, languages that are designed to explicitly and generically identify the underlying meaningful structure (the OHCO) of a document, apart from any intended processing. The nature and degree of the superiority of descriptive markup over other publishing and document

processing strategies did indeed seem to suggest that SGML and XML made the "text itself" computationally available.

However from the beginning some researchers felt that formal metalanguages and descriptive markup alone could not deliver the functionality and interoperability expected[28]; it is now clear that this is indeed the case. An XML representation rigorously defines a structure, but the structure it defines (and serializes) is not itself the conceptual document, but rather a *data structure*, specifically a directed graph with ordered branches and nodes labeled with element names and further decorated with attribute/value assignments — in other words a structure at the same level of abstraction as the W3C Document Object Model. But, as we have seen, the data structure determined by an XML document, cannot be identified with the conceptual document itself. For instance, the data structure may vary (e.g. names for nodes may change), while the document remains the same.

Or to make the point differently, the data structure identified by an XML representation consists of things such as nodes, labels (strings), attribute/value assignments (more strings); and untyped parent/child relationships. But the conceptual document consists of such things as chapters and paragraphs, not nodes and labels, and the relationship between a chapter and a paragraph is a part/whole relationship, not an untyped parent/child relationship. Again, data structures don't have *paragraphs*, they have nodes labeled "p"; documents don't have nodes labeled "p", they have *paragraphs*. So the final step up the ladder of abstraction we have been climbing is getting from the serialized data structure to the conceptual document it represents. If we can make this final upward advance in abstraction we will be where we need to be in order to specify identity conditions for documents.

How can this final step be taken? The bridge that is required would be a formal system for connecting XML document markup to the conceptual structures that markup signals or expresses. That is what is needed. And that is what is missing. XML DTDs provide a mechanism for specifying the *syntax* of an XML vocabulary, but there is no formal mechanism for specifying the *semantics* of that vocabulary — where semantics simply means the basic facts and relationships that are represented by the occurrence of XML constructs. This is the remaining obstacle to the development of robust and theoretically sound identity conditions for documents.

6. Removing the Obstacle: XML Semantics

Semantics in our sense refers simply to the facts and relationships indicated by XML markup, not processing behavior, machine states, linguistic meaning, business rules, or any of the other things that are sometimes meant by "semantics". Consider the markup `<p lang="english">`. Its semantics might be informally expressed by saying that the markup asserts that its content has the property of being a paragraph and being in the English language. (However, there are some reasons to prefer saying that the markup "licenses" rather than "makes" those assertions[41]).

The example suggests that a semantics for XML markup might be given by providing rules for a translation into predicate logic, along with appropriate axioms — and this is indeed the BECHAMEL approach. The example also suggests that the translation will be trivial, which turns out not to be the case at all, as these aspects of markup meaning show[31]:

Propagation: Often the properties expressed by markup are understood to be propagated, according to certain rules, to child elements. For instance, if an element has the attribute/value notation `lang="de"`, indicating that the text is in German, then all child elements have the property of being in German, unless the attribution is defeated by an intervening reassignment. Language designers, content developers, and software designers all depend upon a common understanding of such rules. But XML DTDs provide no notation for specifying which attributes are propagated or what the rules for propagation are. The property of being a paragraph, for example, is not propagated at all (children of a paragraph aren't necessarily paragraphs), being-in-German is propagated until defeated, and being-in-Helvetica will be defeated by a subsequent rendition assignment of being—Helvetica, but *not* by a subsequent rendition assignment of being-in-times. There is no way to specify in a DTD which properties propagate, and what the logic of that propagation is, although of course such relationships are regularly intended by markup language designers, and assumed or inferred by markup language users and software engineers.[41].

Class Hierarchies: XML contains no general constructs for expressing class membership or hierarchies among elements, attributes, or attribute values — nor are there mechanisms for expressing full or partial synonymy, within and across markup languages; although, again, markup users intend and assume these relationships.

Ontological variation in reference: XML markup might appear to indicate that the same thing, is-a-noun, is-a-French-citizen, is-illegible, has-been-copyedited. But obviously either these predicates really refer to different things, or must be given non-standard interpretations. While human readers are not confused by such familiar ambiguities, they are an obstacle to accurate representation and automatic processing.

Arity and Deixis: Some properties expressed by markup are monadic, some polyadic — a title that is the immediate first child of a section for instance is probably the title *of* that section. But property arity is not evident from the markup itself, nor are there explicit "deictic" mechanisms for reliably locating and identifying the arguments[41].

Parent/Child Overloading: The untyped parent/child relations of the XML tree data structure are ambiguous, supporting a variety of implicit substantive relationships. A paragraph might have page break, sentence, and footnote as

child elements, but in each case the parent/child relation represents a different substantive relationship[12].

Fortunately, although the complexities described above do make the development of XML semantics non-trivial, they all appear to be easily handled with familiar knowledge representation devices.

(We emphasize that in saying that there is no standard mechanism for providing SGML/XML vocabularies with a formal semantics we are not saying that these vocabularies do not have semantics at all[33]. Obviously these vocabularies are meaningful (our arguments depend on this observation) and they do fairly effectively identify and describe documents, as they are designed to do. Markup language designers, content developers, and software designers in fact all depend upon a common understanding of the meaning of the XML markup vocabularies that they, respectively, design, apply and exploit. The problem is that this common understanding, even when supported by prose documentation, cannot deliver the reliability, functionality, and interoperability desired — and, in particular, cannot provide rigorous identity conditions.)

7. Using XML Semantics to Represent and Compare Documents

The BECHAMEL system processes an XML document along with the semantic rules for its XML vocabulary, converting the XML representation into a set of assertions in first order predicate logic. These assertions taken together are (or, more accurately, are logically equivalent to) the assertions "licensed" by the serialized XML representation; they are the *meaning* of that representation, the conceptual *document* that is represented. At this level of abstraction not only are all serialization artifacts gone, but so are the artifacts of the data structure: instead of a tree with untyped arcs and labeled nodes decorated with attribute/value pairs, we now have objects such a *paragraphs*, with properties such as *being in German*. The parent-child relationship has been unpacked into various n-place relations and axioms that govern propagation and class relationships[12][31].

This should provide, among other things, a reliable method for determining whether or not two different XML representations are representations of the same document: we convert the XML documents into a BECHAMEL representation (using the semantics associated with the relevant XML vocabularies) and solve for logical equivalence. At this point any cross-walking, application of partial of full equivalences in an interlingua, or other heterogeneity management strategies may also be applied. And in addition to determining identity or non-identity, one should also be able to discover something about the particular semantic relationships between non-identical documents as well. Finally, we can explore the effects of changing selected semantic rules, generalizing predicates, relaxing constraints, etc. — in effect determining under what semantic circumstances a particular XML representation would carry a particular document. This latter possibility suggests that our earlier claim that there is a many-to-one relationship between low level surrogates and the documents they carry is itself still an oversimplification — the relationship in fact appears to be many-to-many.

8. Example

To motivate the issues discussed above, we offer an example of a typical metadata problem. The following is a fragment from a paper, marked up using the tag set provided by the conference to which it was accepted:

```
    <PAPER SECNUMBERS="0"><FRONT
    ><TITLE>Object Mapping for Markup
Semantics</TITLE
    ><AUTHOR CONTACT="1"
    ><FNAME>David</FNAME
    ><SURNAME>Dubin</SURNAME
    ><ADDRESS
    ><AFFIL>University of Illinois</AFFIL
    ><SUBAFFIL>Graduate School of Library and
Information Science</SUBAFFIL
    ><ALINE>501 E. Daniel Street</ALINE
    ><CITY>Champaign</CITY
    ><STATE>IL</STATE
    ><POSTCODE>61820</POSTCODE
    ><CNTRY>USA</CNTRY
    ><EMAIL>ddubin@uiuc.edu</EMAIL
    ><PHONE>217-244-3275</PHONE
    ><FAX>217-244-3302</FAX></ADDRESS
    ><BIO><PARA>David Dubin is a senior research
scientist on the staff
    of the Information Systems Research Lab at the
University of
    Illinois Graduate School of Library and
Information Science. He is a
    member of the Electronic Publishing Research
Group.</PARA
    ></BIO></AUTHOR>
```

In the next example, the same metadata has been retagged using DocBook, an XML application with very similar tags:

```
    <ARTICLE
    ><ARTHEADER
    ><TITLE>Object Mapping for Markup
Semantics</TITLE
    ><AUTHOR><FIRSTNAME>David</FIRSTNAME
    ><SURNAME>Dubin</SURNAME
    ><AFFILIATION><ORGNAME>University of
Illinois</ORGNAME
    ><ORGDIV>Graduate School of Library and
Information Science</ORGDIV
    ><ADDRESS FORMAT="LINESPECIFIC"
    ><STREET>501 E. Daniel Street</STREET
    ><CITY>Champaign</CITY
    ><STATE>IL</STATE
    ><POSTCODE>61820</POSTCODE
    ><COUNTRY>USA</COUNTRY
    ><EMAIL>ddubin@uiuc.edu</EMAIL
    ><PHONE>217-244-3275</PHONE
    ><FAX>217-244-3302</FAX
    ></ADDRESS></AFFILIATION
```

```
    ><AUTHORBLURB><PARA>David Dubin is a senior
research scientist on the staff
    of the Information Systems Research Lab at the
University of
    Illinois Graduate School of Library and
Information Science. He is a
    member of the Electronic Publishing Research
Group.</PARA
    ></AUTHORBLURB></AUTHOR>
```

A comparison of the markup makes it clear that neither bit stream, character stream, nor normalized serialization strategies will correctly identify these documents as identical. Not only are there differences in element names (e.g., “orgdiv” vs. “subaffil”), subtle differences in the parent/child relationships are also in evidence. In the first example, “affil” and “subaffil” are the first two children of the “address” element, whereas in the second “orgname,” “orgdiv,” and “address” are the three children of the “affiliation” element.

One could, of course, solve this problem with a language like XSLT, by transforming one or both of the instances into a normalized form. This approach seems ad hoc to us, since the mapping from syntactic to semantic relationships is actually fairly complex. Consider:

1) A whole/part relationship holds between the organization and the division. But in both documents, the elements naming the organization and its division are siblings, not parent and child.
2) In the examples, the affiliation relationship holds directly between the author and the organizational division. The affiliation between the author and organization can be inferred via the
3) whole/part relationship noted above. However, that relationship would be direct if the division name were absent.
4) The situation is similar for the location relationship that holds between the address and the organization. In these examples, 501 E. Daniel street is the address of GSLIS, not the address of theUniversity of Illinois. But a different inference would be licensed if the “orgdiv”/”subaffil” elements were absent.
5) The email address, phone, and fax numbers aren't really part of the postal address. They represent alternate methods for contacting the author (not the organization or its division!).

Our approach to unifying the information in these fragments is to seek a mapping from the syntactic structures emerging from the parse of the document to statements of the substantive relationships expressed in logical form. These are statements such as:

1) that the author of the paper is David Dubin,
2) that David Dubin is affiliated with the Graduate School of LIS,
3) that GSLIS is part of the University of Illinois,
4) that GSLIS is located on Daniel Street in Champaign,
5) that Dubin can be reached by surface mail at the postal address or by email at the email address or by fax at the fax number or by telephone at the phone number.

A description of our strategies for mapping syntactic to semantic structures can be found in [12][13].

9. Future Directions

This research is still very much at an early stage and there is much work to do developing and testing semantics for popular document vocabularies (XHTML, TEI, ISO12083, DocBook, etc.) and determining what knowledge representation devices will be necessary to do justice to the semantics that markup language designers and users actually intend and rely on, as well as exploring specific applications to real-world problems.

In addition, identifying the consequences of expressiveness requirements for metatheoretical features (such as decidability and completeness) and the computational complexity of the system is also an important item on the agenda. For instance, to represent the semantic content of popular document vocabularies will a language of predicate constants, individual constants and conjunction (i.e. an “existential conjunctive” language) be enough? Or will other devices be needed, such as variables, negation and disjunction, identiy, functions, and universal quantification? Will modal operators be needed as well, and if so which ones? Will we need to represent alethic, epistemic, illocutionary and perhaps even deontic relationships[30][34]? If highly expressive languages are needed will the documents themselves nevertheless present only relatively manageable logical expressions? Necessarily or only in general?

In answering these questions we believe we will be not only developing the foundation for solving current practical problems, but also discovering fundamental facts about the nature of documents as communicative objects.

Acknowledgements

Obviously this work owes much to our BECHAMEL collaborators Michael Sperberg-McQueen and Claus Huitfeldt. We also thank Pat Lawton, Kevin Hawkins and other members of the GSLIS Electronic Publishing Research Group for help on this particular paper. Errors and confusions are ours alone.

References

[1]Arms, W. Y. (2000). *Digital Libraries*. MIT Press.
[2]Berners-Lee, T. (2003). *What do HTTP URIs Identify?* Retrieved February 15 2003 from http://www.w3.org/DesignIssues/HTTP-URI.

[3] Booth, D. (2003). *Four Uses of a URL: Name, Concept, Web Location and Document Instance*. Retrieved January 28 2003 from http://www.w3.org/2002/11/dbooth-names/dbooth-names_clean.htm.
[4] Boyer, J. (Ed.). (2001). *Canonical XML*. W3C Recommendation, 15 March 2001.
[5] Buckland, M. K. (1997). What is a "document"? *Journal of the American Society of Information Science 48*, 804-809.
[6] Buzzetti , D. (2002). Digital representation and the text model. *New Literary History 33*, 61-88.
[7] Clark, K. G. (2002). *Identity Crisis*. Retrieved May 11, 2002 from http://www.xml.com/pub/a/2002/09/11/deviant.html (XML.com).
[8] Coombs, J.H., Renear, A.H., DeRose, S.J. (1987). Markup systems and the future of scholarly text processing. *Communications of the Association for Computing Machinery 30*(11), 933-947.
[9] Cowan, J., Tobin R. (Ed.). (2001). *XML Information Set*. W3C Recommendation, 24 October 2001.
[10] DeRose, S.J., Durand, D., Mylonas, E., Renear, A.H. (1990). What is text, really? *Journal of Computing in Higher Education 1*(2), 3-26.
[11] Doerr M., Hunter J., Lagoze C. (2003). Towards a core ontology for information integration, *Journal of Digital Information. 4*(1).
[12] Dubin, D., Sperberg-McQueen, C.M., Renear, A., Huitfeldt, C. (2003). A logic programming environment for document semantics and inference. *Journal of Literary and Linguistic Computing 18*(1), 39-47
[13] Dubin, D. (2003). *Object mapping for markup semantics*. To be presented at Extreme Markup 2003. Montreal. (Scheduled to be published in the electronic proceedings).
[14] Guarino, N. (1999). The role of identity conditions in ontology design. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-99) Workshop on Ontologies and Problem-Solving Methods* (pp. 221-234). Stockholm, Sweden.
[15] Guarino, N., Welty, C. (2000). Conceptual modeling and ontological analysis. AAAI-2000 Tutorial. Retrieved May 11 from http://www.cs.vassar.edu/faculty/welty/presentations/aaai-2000/.
[16] ICOM/CIDOC Documentation Standards Group (1998). *CIDOC Conceptual Reference Model.* http://cidoc.ics.forth.gr/.
[17] IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). *Functional Requirements for Bibliographic Records.* München: K.G. Saur.
[18] ISO. (2003). *Draft Standard ISO/CD 21047, International Standard Textual Work Code* (ISTC).
[19] Le Boeuf, P. (2002). About IFLA's comments on ISTC. ISO TC 46/SC 9/WG 3 N 42 2002-03-19.
[20] Levy, D. M. (1988). Topics in Document Research. In *Proceedings of the ACM Conference on Document Processing Systems* (pp. 87-193). New York: ACM.
[21] Levy, D. M. (2001). *Scrolling Forward: Making Sense of Documents in the Digital Age*. New York: Arcade.
[22] Lagoze, C. and J. Hunter. (2001). The ABC ontology and model. *Journal of Digital Information. 2*(2).
[23] Lorie, R. A. (2002). A methodology and system for preserving digital data. In *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries.* New York: ACM.
[24] Lynch, C. (1999). Canonicalization: A fundamental tool to facilitate preservation and management of digital information. *D-Lib Magazine 5*(9).
[25] Pepper, S., Schwab, S. (2003). *Curing the Web's Identity Crisis.* Retrieved from http://www.ontopia.net/topicmaps/materials/identitycrisis.html on May 1, 2003.
[26] Quine, W.V. (1969). Speaking of objects. *Ontological Relativity and Other Essays*. New York: Columbia University Press.
[27] Ramalho, J.C., Henriques, P.R. (1998). Beyond DTDs: constraining data content. In *Proceedings of SGML/XML Europe 98*. GCA.
[28] Raymond, D.R., Tompa, F.W., Wood. D. (1993). *Markup Reconsidered*. Technical Report 356, Department of Computer Science, The University of Western Ontario. (Presented at *the First International Workshop on the Principles of Document Processing*, Washington DC, October 21-23 1992; earlier version circulated as "Markup Considered Harmful" in the late 1980s.)
[29] Raymond, D.R., Tompa, F.W, Wood, D. (1996). From Data Representation to Data Model: Meta-Semantic Issues in the Evolution of SGML, with Frank Wm. Tompa and Derick Wood, *Computer Standards and Interfaces 18*, 25-36.
[30] Renear, A. (2000). The Descriptive/Procedural distinction is flawed. *Markup Systems Theory and Practice 2*(4), 411-420. 2000.
[31] Renear, A., Dubin, D., Sperberg-McQueen, C.M., Huitfeldt, C. (2002). Towards a semantics for XML markup. In R. Furuta, J. I. Maletic, and E. Munson, (Eds.), *Proceedings of the 2002 ACM Symposium on Document Engineering* (pp. 119-126). McLean VA. ACM.
[32] Renear, A., Dubin, D., Sperberg-McQueen, C.M., Huitfeldt, C. (2003). XML semantics and digital libraries. In C. C. Marshall, G. Henry, and L. Delcambre, (Eds), *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 303-305). IEEE.
[33] Renear, A. (2003). *Text markup -- data structure vs. data model*. Presented at the Joint International Conference of the Association for Computers and the Humanities Association for Literary and Linguistic Computing (ACH/ALLC 2003). Athens, Georgia.
[34] Renear, A. *First thoughts on modal logic for document processing*. Presented at Extreme Markup 2003, Montreal

Canada. (Scheduled to be published in electronic proceedings).
[35] Salminen, A., Tompa, F.W. (2001). Requirements for XML document database systems. *Proceedings of the ACM Symposium on Document Engineering*. ACM.
[36] Simons, G.F. (1997). Conceptual modeling versus visual modeling: A technological key to building consensus. *Computers and the Humanities, 30,* 303–319.
[37] Simons, G.F. (1999). Using architectural forms to map TEI data into an object-oriented database. *Computers and the Humanities, 33,* 85–101. (Originally presented in 1997 at TEI 10, Brown University, Providence, RI.)
[38] Simons, G.F. (2003). *Developing markup metaschemas to support interoperation among resources*. Talk presented at the Joint International Conference of the Association for Computers and the Humanities Association for Literary and Linguistic Computing (ACH/ALLC 2003), Athens, Georgia.
[39] Smiraglia, R. (2001). *The Nature of a Work.* Scarecrow Press.
[40] Sperberg-McQueen, M., Burnard, L. (Eds). (1994). *Guidelines for Text Encoding and Interchange* (TEI P3). Chicago, Oxford: ACH/ALLC/ACL Text Encoding Initiative.
[41] Sperberg-McQueen, C.M., Huitfeldt, C., Renear, A. (2000). Meaning and interpretation of markup. *Markup Languages: Theory and Practice, 2,* 215–234.
[42] Sperberg-McQueen, C.M., Renear, A., Huitfeldt, C., Dubin, D. (2002). *Skeletons in the closet: Saying what markup means.* Presented at Joint International Conference of the Association for Computers and the Humanities Association for Literary and Linguistic Computing (ACH/ALLC 2002), Tübingen, Germany.
[43] Sperberg-McQueen, C.M., Dubin, D., Huitfeldt, C., Renear, A. (2002). Drawing inferences on the basis of markup. In B. T Usdin and S. R. Newcomb (Eds.), *Proceedings of Extreme Markup Languages 2002,* Montreal, Canada.
[44] Svenonius, E. (2000). *The Intellectual Foundation of Information Organization*. MIT Press.
[45] Welty, C., Ide, N. (1999). Using the right tools: Enhancing retrieval from marked-up documents. *Computers and the Humanities, 33:* 59–84. (Originally delivered in 1997 at the TEI 10 conference, Brown University, Providence, RI.)
[46] Wuwongse, V., Anutariya, C., Akama, K., Nantajeewarawat, E.: XML declarative description: A language for the semantic web. *IEEE Intelligence Systems, 16:* 54–65.