The Effect of XML Markup on Retrieval of Clinical Documents

Catherine Arnott Smith, PhD
Syracuse University

Abstract

Objective: To determine the effect on clinical information retrieval of structuring typical clinical documents in XML, according to the general guidelines of Health Level Seven's Clinical Document Architecture. **Methods:** One thousand clinical documents of eight frequently occurring types were deidentified and marked up in XML for access using a Web browser. Fifty information-seeking tasks were posed to subjects. The tasks were comprised of two typical clinical question types—individual patient results reporting and cohort identification. A control group of physician subjects could perform only free-text, keyword searching. The treatment group's interface permitted field-based searching of particular sections within each document. Differences in precision and other measures of search success across and between question types were investigated for statistical significance. **Results:** No statistically significant differences were found between the control and treatment conditions in mean time elapsed or the mean number of records in the final result set. In fact, tasks performed in the treatment condition required a mean number of *more* steps in the search sequence to a degree that was statistically significant. Tasks performed in the treatment condition had a statistically significant lower rate of mean precision. There was no statistically significant difference between the means of relevance of the individual patient and cohort identification tasks. **Conclusion:** These findings are in line with Tange et al.[1] who found that coarser granularity of clinical narrative gave better results. The results of this experiment also have implications for automatic text processing. Complex tag sets cannot ultimately resolve problems of unstandardized structure; the lack of existing structure within clinical documents is itself a significant limitation.

Introduction

Essin and Essin[2] were the first to propose "loosely structured documents" as the ideal medical record implementation. Loosely structured, or semistructured, documents have been defined simply as documents that have much in common—enough in common that a general statement can be made about their components[3]. Essin's strategy proposed to capitalize on loose structure first by understanding the data elements themselves, and second by developing a "meta-level" that contained knowledge *about* the data elements. Lincoln and Essin[4] later formally proposed this as an SGML solution and set in motion the chain of events culminating in Health Level Seven's Clinical Document Architecture (CDA; ANSI HL7 CDA R1.0-2000) (for a review and description of Health Level Seven's SGML/XML SIG's history and activities, see Dolin, Alschuler, Biron, et al.[5])

The CDA provides standards for electronic document exchange by using features of semistructured clinical documents for enhanced semantic processing. Wolff, Flörke, and Cremers[6] point out that the principal defining feature of structured documents is the presence of these explicit semantics for their structural parts. The benefit is that the meaning of the structural components—the sections—can be exploited, as can the meaning of the text they contain, thus comprising Essin's "meta-level".

The body of existing information retrieval work that most closely resembles that originally proposed by Essin is that called "passage retrieval". Defined as "the task of identifying and extracting fragments from large, or short but heterogeneous full text documents"[7], passage retrieval is a subset of research into "corpus-based" text processing, in which the text collection itself is used to derive information needed for analysis and for characterization[8]. Passage retrieval attempts to address the significant problem of full-text document retrieval which rests in the sheer size of the documents. This characteristic of full text may have a confounding effect: It may be large and difficult to manage, and relevant information may be widely scattered, and therefore hard for the user to extract.[9] Callan[10] points out that all information retrieval can be viewed as a passage retrieval task—or, at least, a task of retrieving documents that have an internal structure: "Each element is a source of evidence that can be used in retrieval".

The precise definition of a "passage" has varied in the literature. In general, passages have been defined as being "some semantic structural feature" of the

document[11]. Clinical documents can be classified as discourse passages, which are based on units of textual discourse such as sentences, paragraphs, and sections. If text is neither highly structured nor edited, passage retrieval is more difficult than document retrieval, because, according to Melucci[12], "Any pre-defined segmentation of the text is absent, unless the text author has provided the text itself with a structure reflecting the organization of the topic which might support the retrieval of passages relevant to the topics (p. 44)."

Both subtopics and sections are understood to be visible, explicit structural features of the text that are available for semantic processing. Their presence is alerted by strings of text that are legible labels—subheadings, or section headings: "[T]ext structure can be a good approximation of topic organization"[13].

The primary advantage of passage retrieval rests in its ability to enhance the relevance of results returned to the user. An understanding of the document's structure has been shown to help the user determine relevance of a passage[14]. Passage retrieval helps because it concentrates the reader's attention on those parts of the text that have a "high density" of relevant information and also gives the reader an "intuitive overview" of the way in which those relevant subsections are distributed throughout the corpus[15 16 17]. O'Connor[18] found, for example, in work with CANCERLIT, that answers to bibliographic questions tended to be located in particular places within bibliographic citations (another form of passage), and proposed that this knowledge could be exploited for ranking purposes.

Within the domain of healthcare, Huibert Tange and collaborators at Maastricht University have assessed the effect of structuring clinical documents on their retrieval [19 20 21]. Tange, Hasman et al. have proposed that the "search structure", or information retrieval process, in the domain of clinical information has two main aspects: the "granularity" of the paragraphs, and the relationship of those paragraphs to each other. The number of paragraphs being searched—an important aspect of granularity--is found to be inversely related to the ease of searching them, which relates obviously to the work of Salton and his colleagues ascribing passage understanding to relevance. The same proposition was put forward earlier by Lincoln and Essin[22], again with only implicit acknowledgment that relevance had anything to do with the desired result: "An ability to specify text searches as narrowly as necessary using additional tags [SGML] avoids *secondary parsing or sorting to eliminate unwanted material*" (italics mine; p. 229). In one experiment, Tange et al.[23] found that high granularity of clinical documents (meaning documents with large numbers of sections) was associated with increased speed of information retrieval for progress notes only; certainly, a finding of high value in a high-need clinical situation. However, this finding did not hold for other types of documents, specifically Medical History or Physical Examination documents, where excessive partitioning caused more problems than it solved. Because of these conflicting results, the Maastricht group concluded that more experimental investigation is necessary and recommended[24].

Application of results from the relatively small body of passage retrieval literature to retrieval of structured documents in medicine needs to be considered in light of the nature of the documents being retrieved. In retrieval of clinical documents by passages, one important difference from highly structured texts such as encyclopedias is that since each document represents only one patient and/or one clinical event (e.g., Mrs. Smith's family history; a radiology procedure), similarity of passages will probably seldom occur *within* documents, but could more easily occur *across* aggregations of documents that are of the same or similar type, e.g., all of Mrs. Smith's statements of her family history for the past 10 years, or all radiology reports for all patients in the system.

Clinical documents are also extremely short and have unique and nonredundant text. Unlike magazine articles, such as those investigated for passage retrieval by Hearst and Plaunt[25], the section heading labels in clinical documents are not content summarizations of the paragraphs they introduce. In fact, the section headings used in clinical documents bear more relationship to the fields of a database than they do to the discourse-structured text of passage retrieval experiment. These "section headings", "labels", or "segment labels", as they are variously called in the literature, serve as the means by which readers navigate the documents. Nygren, Johnson, and Henriksson[26] identified three reading techniques of medical records: first, skipping over irrelevant sections; second, skimming sections identified as possibly relevant; and third, reading needed information carefully. Tange, Dreessen, et al.[27] collapsed the first two filtering steps into one, portraying a user who searches through the record "guided by the internal structure" to select relevant sections, then reading the content. By alerting the reader to content, labels serve both to denote the structure and define the domain of knowledge: "The structured representation acts as an intensional

definition, in the particular vision of a world embedded in a structure.[28]"

Health Level Seven's RIM in concert with its Clinical Document Architecture proposes such an embedded world. This paper reports on a passage retrieval experiment using clinical documents structured according to the CDA. It assessed the effect of structuring clinical documents on their successful retrieval in a simulated clinical situation.

Methods

One thousand clinical documents from the MARS system in place at the University of Pittsburgh Medical Center (UPMC) were randomly selected and automatically deidentified. These documents comprised the electronic medical records of patients attended by UPMC oncologists. The thousand selected were evenly distributed among the 8 most frequently occurring types found in a pilot study within the same medical center[29]: radiology reports; progress notes; physician letters; operating room notes; history and physical notes; surgical pathology reports; discharge summaries; and emergency room visits.

Considerable variation was found in this pilot study, even within document types, as to the number and content of section headings in MARS documents. MARS documents contain no fielded information, nor are any stylistic conventions used to convey meaning (e.g., upper-case letters, bold face, underline, etc.). Thus little imposed standardization was thought to exist in these documents and in fact none was found. Document processing for the final experiment revealed, for example, 18 different variations on "Laboratory Test Results (see Table 1).

Markup in XML was done manually according to the following rules: (1) A "field" within a document was defined as "any string of characters concluding with a colon and a line break." (2) The labels used for section headings always replicated as much as possible the string of characters used in the original. Normalization of lexical variants (e.g., *Problem* vs. *Problems*) was accomplished through the principle of "literary warrant," in which the predominant usage in the texts being analyzed becomes the term denoting the concept. (3) With the exception of the conditions noted above, no inference was made by me as to the synonymity of a concept expressed in one section heading with a concept expressed in another.

Table 1
18 variations on "Laboratory Test Results"

Laboratories
Laboratory and Diagnostic Findings*
Laboratory and Radiographic Data
Laboratory and Radiographic Findings
Laboratory and Radiology Data
Laboratory Data
Laboratory Data/Radiographic Findings
Laboratory Evaluation
Laboratory Findings
Laboratory Results
Laboratory Studies
Laboratory Values
Laboratory Work
Laboratory
Laboratory, Radiographic, and Other Diagnostic
Study Findings
Labs
Labs on Admission
Radiographic and Laboratory Findings

A database was constructed using an open source XML product, Xindice, supported by the XML Apache Project (www.xml.apache.org). The search engine, accessible via a Web browser, was built using Java servlets and Java server pages. It permits use of Boolean operators AND and OR and partial string matching to search and display full-text XML documents.

Subjects for this experiment were 10 physicians (9 M, 1 F, ages 28-45) drawn from a convenience sample, all located within the UPMC system and ranging in clinical experience from a medical school graduate through attending faculty members. Six of the 10 subjects reported that they never used MARS to access patient data; three accessed it weekly or less frequently, and one reported daily use. Eight subjects had formal coursework in computer science or a related field, while 9 of the 10 had additional self-guided learning experiences involving computers. All subjects self-reported as "sophisticated" (3) or "very sophisticated) (7) computer users.

The two experimental groups were composed as follows. Subjects in the control group had access to clinical documents which were flat files, marked up only to the Clinical Document Architecture's Level One, or "document information": identified by document type only, as "radiology report" or "history and physical." The search interface in the control condition permitted free-text searching of the full text of the clinical document.

Subjects in the treatment group had access to clinical documents marked up in XML to Clinical Document Architecture Level Two: this is the level of semantic markup at which the sections of the document are specified, but the content of those sections is not. The search interface in the treatment condition permitted searching within specific fields of the document. For example, in a radiology report, subjects could search for specific text within the History or Impressions sections by selecting the fields "History" or "Impressions" from a pull-down menu. These fields were created from the section headings used in the clinical documents and normalized where necessary according to the rules delineated above.

Fifty queries (5 per subject x 10 subjects) were developed for this experiment and assigned as tasks to all subjects in both control and treatment groups. Two general types of queries were created based on the typology proposed by Safran and Chute[30]. Type I, "Results Reporting", focused on identification of an individual patient; Type II, "Cohort Description," required identification of a group of patients with one or more common attributes. All queries were back-generated from the clinical documents containing the answers, so that the gold standard answers were known prior to the beginning of the experiment. An example of each type of question appears in Table 2, below.

Table 2. Question Types

Type I: Results Reporting
One Progress Note documents the case of a patient whose skin has changed color over the past few months. What are the features of this change in pigmentation?
Type II: Cohort Identification
Using the Discharge Summaries document set, please identify all patients with a diagnosis of colon cancer.

Outcome measures were both subjective and objective. Each subject completed a post-task questionnaire designed to obtain opinions about each search task and asking specifically for subjects' relevance judgments of the document that fulfilled their information-seeking task. Objective measures were: time elapsed during search session; number of steps in search sequence; number of records in the final result set; and the total number of unique search strings used. Finally, precision measures were obtained using the standard formula for precision: A = Number of relevant and retrieved documents; B = Number of nonrelevant and retrieved documents; Precision = A/(A + B). As noted above, "relevant" documents were those known *a priori* to be the answer to specific search tasks.

Research hypotheses were these:

I. Tagging document elements will result in enhanced retrieval as measured by time elapsed during the session; number of steps in the search sequence; the number of records in the final result set; and the total unique search strings used.

II. Tagging document elements will result in a difference in relevance rates between Type I question tasks and Type II question tasks.

Results

Hypothesis I was not supported. No statistically significant differences were found between the control (339.5 seconds ± 287.8) and treatment conditions (400 seconds ± 311.9) in mean time elapsed or the mean number of records in the final result set (control, 4.1 ± 7.8; treatment, 5.2 ± 10.5).

However, tasks performed in the tagged treatment condition required a mean number of *more* steps in the search sequence to a degree that was statistically significant (control, 3.6 ± 2.9; treatment, 21.2 ± 13.2). Differences in precision were additionally investigated. The treatment condition tasks had a statistically significant lower rate of mean precision (control, .92 ± .231; treatment, .79 ± .383).

Hypothesis II was also not supported. There was no statistically significant difference between the means of relevance of Type I and Type II questions tasks (Type I, .72 ± .388; Type II, .67 ± .345).

Discussion

The findings of this study are in line with those of Moffat et al.[31], who found that breaking documents down into passages did not improve precision, and also to some extent with those of Tange et al.[32] in which coarser granularity of clinical narrative gave better results.

Theorists interested in the medical record have argued that the record serves to transfer embedded data in context[33] and furthermore that this context equates to a context of production that is consciously perceived by the reader[34]. However, analysis of the data showed no significant differences between the results of MARS users and non-users. The extreme lack of structure and standardization of the MARS data itself may have proved intractable. MARS

documents contain so many variants of section headings that the simple choice of fields presented a cumbersome obstacle to the clinician subjects. This situation might have been avoided if the data model had been developed by more than one person: in a distributed, democratic fashion by a committee of representative MARS users and developers. Finally, the user interface itself imposed limitations. This interface was purposely designed to permit selection of single but not multiple search fields so that it most closely emulated the interface currently available to UPMC clinicians. Based on subject comments, however, it is clear that an interface permitting "clustering" of related fields in a multi-field search would have enhanced the search experience.

Conclusion

What are the implications of this experiment for developers of the Clinical Document Architecture? The implications are primarily for the data model. If the inadequacy of section headings—labels—for representation of clinical content has contributed to the negative results displayed here, it may be a signal that complex tag sets cannot ultimately resolve problems of unstandardized structure; the lack of existing structure is itself a significant limitation. Or, more colloquially: garbage in, garbage out.

Future research will more effectively utilize clinicians' mental maps by involving the users themselves in refining the data model used for the present experiment. Once more normalization of fields is achieved, it should be possible to develop an improved user interface, particularly one allowing the selection of multiple search fields. Rerunning the experiment with the new interface should allow a more precise determination of the interface's contribution to these negative results. Finally, users' insight into the concepts used to label and aggregate clinical document fields—for example, a clear definition of what fields might usefully be clustered together as "related"—ought to be generally beneficial to designers of electronic medical record systems, particularly those who are interested in wholesale document conversion from paper.

Acknowledgments

Prof. Arnott Smith was supported by a U.S. National Library of Medicine (NLM) Medical Informatics Trainee Grant # 5-T15 LM07059.

References

[1]Tange HJ, Schouten HC, Kester ADM, Hasman A. The granularity of medical narratives and its effect on the speed and completeness of information retrieval. *JAMIA* 1998; *5*: 571-582.

[2]Essin DJ, Essin CD. Computerizing medical records: Software criteria for systems to document patient encounters. *Crit Care Med* 1990*; 18*(1): 100-102.

[3]Essin DJ. Intelligent processing of loosely structured documents as a strategy for organizing electronic health care records. *Meth Inf Med* 1993; *32*:265-268.

[4]Lincoln TL, Essin DJ. A document processing architecture for electronic medical records. *Medinfo* 1995; 227-230.

[5]Dolin RH, Alschuler L, Beebe C, Biron PV, Boyer SL, Essin D, Kimber E, Lincoln T, Mattison J. The HL7 Clinical Document Architecture. *JAMIA* 2001; *8*(6): 552-569.

[6]Wolff JE, Flörke H, Cremers, AB. Searching and browsing collections of structural information. In *Proc IEEE Adv Digit Libraries*, 2000; 141-150.

[7]Melucci M. Passage retrieval: A probabilistic technique. *Inf Proc Mgmt* 1998; *34*(1):43-67.

[8]Salton G, Allan J. Selective text utilization and text traversal. *Hypertext '93*; 131-144.

[9]Tombros A, Sanderson M. Advantages of query biased summaries in information retrieval. *Proc ACM SIGIR 1998;* 2-10.

[10]Callan JP. Passage-level evidence in document retrieval. *Proc ACM SIGIR* 1994; 302-310.

[11]Kaszkiel M., Zobel J. Passage retrieval revisited. *Proc ACM SIGIR* 1997*;* 178-185.

[12]Melucci, 1998, *op. cit.*

[13]Melucci, 1998, *op. cit.*

[14]Salton G, Buckley C. Automatic text structuring and retrieval: Experiments in automatic encyclopedia searching. *Proc ACM SIGIR* 1991; 21-30.

[15]Tombros & Sanderson, 1998, *op. cit.*

[16]Salton G, Allan J, Buckley, C. Approaches to passage retrieval in full text information systems. *Proc ACM SIGIR 1993* ; 49-55.

[17]Salton & Allan, 1993, *op. cit.*

[18]O'Connor J. Answer-passage retrieval by text searching. *JASIS* 1980*; 31*(4): 227-39.

[19]Tange HJ, Dreessen VAB, Hasman A, Donkers HHLM. An experimental electronic medical-record system with multiple views on medical narratives. *Comp Meth Prog Biomed* 1997;*54*:157-172.

[20]Tange et al., 1998, *op. cit.*

[22]Lincoln & Essin, 1995, *op. cit.*

[23]Tange et al., 1998, *op. cit.*

[24]Tange et al., 1997, *op. cit.*

[25]Hearst MA, Plaunt C. Subtopic structuring for full-length document access. *Proc ACM SIGIR* 1993; 59-68

[26]Nygren E, Johnson S, Henriksson P. Reading the medical record: I. Analysis of physicians ways of reading the medical record. *Comp Meth Prog Biomed* 1992; *39:* 1-12.

[27]Tange et al., 1997, *op. cit.*

[28]Rossi Mori A., Galeazzi E., Consorti F, Bidgood, WD. Conceptual schemata for terminology: A continuum from headings to values in patient records and messages. *Proc AMIA*,1997;650-654.

[29]Arnott Smith C, Lowe HJ. A preliminary analysis of XML's potential role in representing the semantics and structure of the oncology patient record. *Proc AMIA* [poster], 2001.

[30]Safran C, Chute CG. Exploration and exploitation of clinical databases. *Int J Biomed Comput*, 1995; *39*(1): 151-156.

[31]Moffat A, Sacks-Davis R, Wilkinson R, Zoebel J. Retrieval of partial documents. *Proc TREC-2*, 1994 ;181-190

[32]Tange et al., 1998, *op. cit.*

[33]Kluge E.-H. W. The medical record: Narration and story as a path through patient data. *Meth Inf Med* 1996; *35*:88-92.

[34]Berg M, Goorman E. The contextual nature of medical information. *Int J Med Inf* 1999; *56*:51-60.