A Native XML Database Design for Clinical Document Research

Stephen B. Johnson, PhD, David A. Campbell M. Phil., Michael Krauthammer MD M.Phil,
P. Karina Tulipano, Eneida A. Mendonça, MD PhD, Carol Friedman PhD, George Hripcsak, MD
Department of Medical Informatics, Columbia University, New York NY

INTRODUCTION

Health-care institutions are gaining an increasing interest in exploiting the data that are gathered through electronic medical records. Narrative data, generated by transcription or direct entry, represents a far greater challenge for analytic tasks. Moreover, a small number of institutions are beginning to explore deeper structuring of narrative data using natural language processing (NLP). The data produced by NLP systems has a complex, nested structure. Current electronic medical records do not have the ability to store and retrieve data of this complexity in a suitable way.

Technologies based on the Extensible Markup Language (XML) offer an intriguing alternative to conventional databases. Complex documents marked up using the XML standard constitute what is known as "self-describing data" or "semistructured data", which is defined as data that has an irregular structure not known in advance, and which can change frequently and without notice. This abstract describes the development of a new database design based on XML for the purpose of accessing clinical reports in a research environment. The primary focus of this database will be to facilitate research, especially research on clinical text. There are a broad range of user needs which must be accounted for, primarily:

1. Ability to rapidly process queries against text (key word searches, etc)
2. Ability to rapidly process queries against annotations or 'mark up' added to the text
3. A standard method for querying the documents
4. The ability to select documents along many different axes of interest (within or across patients, over time, etc)
5. The ability to deliver the correct level of granularity of information
6. A flexible schema to adapt to new data without changing or hindering query formulation
7. A flexible schema that can adapt to new annotation styles (from using new parsers, , vocabularies, etc)

RESULTS

We considered both commercial and free database systems for managing our documents. We chose to use the purely native XML Tamino DBMS. The adoption of this large robust vendor system gives us speed (goals 1 and 2), and a standard means of query (goal 3). Tamino supports these goals, and offers better support and security options than other XML database systems we explored. The newest version of Tamino uses the newly developed XQuery language which is the current WC3 standard for XML document queries. This language includes the join operations necessary for goal 4.

The schema for our database requires structure for the document's meta-information as well as structure for the document itself. For the structuring of this meta-information, we chose to adhere to the evolving HL-7 Clinical Document Architecture.[1] The standard is defined in XML and therefore instantly compatible with our desire to use an XML database system. Furthermore, it is likely that this standard will be more widely adopted in the future, making the database forward compatible. It includes standardized definitions for data such as patient and provider identifiers, demographics, and document type, which are required for the selections in goal 4.

Our schema for the document itself allows for structuring the document on a sentence by sentence level. This a level of granularity which the newest version of the CDA incorporates. However, we wish to allow more than one structural system simultaneously. For example, one NLP system may link words in a sentence through semantic relationships like 'part-of'. A second system may link words by grammatical relationships such as subject and predicate. Both structure are tree-like and could not be naturally incorporated in a single XML tree. We resolve this conflict by linearizing and flattening the structures so they may coexist.[2] This 'peaceful coexistence' of structures also gives of the flexibility to satisfy goal 5. A library of XSLT stylesheets allows the structures to be hidden or removed as the user desires satisfying goal 6.

Finally, our schema allows the annotation of individual words with vocabulary codes (UMLS, LOINC, etc), semantic tags from an NLP system and even part-of-speech tags from a tagger. Here too, an XSLT library is used to filter out unwanted level of tagging satisfying our final goal 7.

CONCLUSIONS

We have proposed an XML database for the storage of clinical documents to meet the document research needs outlined. This database is currently being implemented to support the research needs at Columbia University. We believe that the architecture will facilitate our research needs by providing researchers a stable and standard way to collect data sets for investigation.

REFERENCES

1. Dolin RH, et.al. "The HL7 Clinical Document Architecture" J Am Med Inform Assoc. 2001 Nov-Dec;8(6):552-69.

2. Krauthammer M., et.al. "Representing nested semantic information in a linear string of text using XML" *Proceedings/AMIA Annual Symposium*, 2002