Document Ontology: Supporting Narrative Documents in Electronic Health Records

Jason S. Shapiro, MD, Suzanne Bakken, RN, DNSc, Sookyung Hyun, RN, MS, Genevieve B. Melton, MD, Cara Schlegel, RN, MS, Stephen B. Johnson PhD
School of Nursing and Department of Biomedical Informatics, Columbia University, New York, New York

Electronic health records (EHRs) are beginning to manage an increasing volume of narrative data, such as clinical notes pertaining to admission, patient progress, shift change, follow-up, consultation, procedures, etc. These documents fall into a wide variety of classes, based on who is writing them, for what purpose, and in which location, suggesting the need for a document ontology (DO) to model our knowledge of health care documents and their properties. This paper focuses on one aspect of the Health Level 7 (HL7)/ Logical Observation Identifiers, Names, and Codes (LOINC) DO, the Subject Matter Domain (SMD). We created a new polyhierarchical structure for the SMD that combines the current value lists from the LOINC database with another value list from the American Board of Medical Specialties (ABMS). We refined and evaluated the new structure through expert review of the ontology, a survey of medical specialty boards, and specification of SMDs for a corpus of clinical notes.

INTRODUCTION

Clinicians rely on narrative for multiple purposes, including documentation in the medical record and communication of their findings and impressions to patients, patients' family members and other clinicians in formal letters and with the scientific community as a whole in the form of journal publications. Narrative allows us to share complex ideas in an efficient and often seemingly effortless manner. Its use in the medical record is extremely important for clinicians because it allows them to synthesize disparate facts and data elements and to paint a picture rich with meaning that is easily interpreted by other clinicians.[1] Many current systems that provide EHRs for hospitals and physician practices use template-based systems with check boxes and drop-down menus in order to capture structured data elements in databases. Structured data entry does not support the expressiveness and flexibility to which clinicians are accustomed, and it can be difficult to interpret and reconstruct meaning from structured data due to loss of contextual information.[2] Much of the meaning and inference that can be gleaned by the clinician through the use of narrative is lost when a rigidly structured template is used, and the ability to communicate complex ideas in an efficient and fluid manner diminishes. Structured data, however, is especially useful to support other important activities including computer-based decision support and alerts, clinical research, and billing and business practices.

Clinicians use narrative documents for many different purposes in a large variety of clinical situations, and these processes have become fairly conventional so that narrative documents fall into certain types or classes, based on who is writing them, for what purpose, and in which location. One of the current problems with using and maintaining a medical record based on narrative documents is that there are a very large number of document classes. This makes it difficult for clinicians to find a given document in the medical record, or to specify what kind of document they would like to create.

The work described in this paper is part of a greater effort to create a generalizable DO that organizes narrative clinical documents into a polyhierarchical taxonomy of names. The immediate purpose of this work is to help organize the collection of documents in the New York Presbyterian Hospital (NYPH) system for use in the eNote EHR, which is currently under development.[3] The eNote user interface utilizes selection criteria derived from the DO's structure to facilitate highly selective document searching and retrieval. Additionally, this allows eNote to present the user with a refined set of semi-structured, user-specific templates on which to create new documentation. Continuing work on the DO may also lead to a standardized structuring of fields and field content within each of these document template types.

BACKGROUND

eNote

The objective of the eNote project is to create a new kind of EHR that integrates different types of information across the record, harmonizes information across disciplines and maintains the continuity of information as our knowledge of the patient evolves. eNote uses an XML database to store documents represented using the Clinical Document Architecture (CDA) of HL7.[4,5] In addition, eNote employs natural language processing (NLP)[6,7,8] techniques to acquire the structured data elements of interest for the purposes mentioned above.

The DO serves two purposes in eNote: First, it will help further work toward widely-distributed document exchange between institutions by developing a naming standard for document organization and transmission.[9] Second, it will support the management of a large variety of clinical documents, allowing the creation of an EHR that is based on the narrative documents to which clinicians are already accustomed.

Previous DO Projects

In order to address user dissatisfaction with the time required to find documents among hundreds of irregularly structured titles in the Computerized Patient Record System (CPRS) in Veterans Affairs (VA) medical centers, Brown et al. created a Document-naming Nomenclature (DNN).[10] The DNN used a nomenclature to specify three document identifier categories: characteristics of author, health care event, and organizational unit providing care. These categories were then populated with allowable values, and a syntax was created for combining them to produce standardized document names. Their analysis demonstrated significant coverage with the DNN of document titles from three non-VA hospital systems, but the authors had difficulty fully specifying "sections" or subspecialty values in the care unit category.

Other work done initially by the HL7 Document Ontology Task Force (DOTF)[11,12] was later continued in a joint effort with the LOINC committee.[13,14] Together they have developed a DO that uses LOINC codes for clinical documents, concentrating initially on "clinical notes" as defined by the HL7 CDA standard.[15] Their model uses a system of 5 axes with multiple classes under each axis:

1. Kind of Document – Using the document's general structure, describes on a macro level the kind of document being considered (i.e. clinical note, letter, consent, etc). The purpose of this is to define distinct document headers.
2. Type of Service – Characterizes the actual kind of service that is provided to or for the subject of the note, usually the patient (i.e. evaluation and management, communication, interventional procedure, etc.). Time sequence is also subsumed by this axis.
3. Setting – An extension of CMS's coarse definitions (Home, Hospital, Nursing Home, and Outpatient) with synonyms and multiple subclasses in the hierarchy. This is not equivalent to location, which is often more locally defined and can be included within the message itself when documents are sent between institutions.
4. Subject Matter Domain – Characterizes the subject matter and/or discipline that is relevant to the document being considered, and is the main focus of the current study (see below).
5. Training/Professional Level – Characterizes the training or professional level of the document (e.g., Attending, Resident, Nursing Student, Nurse Practitioner).

The naming convention operates by using the Kind of Document and at least one of the values from the other four axes. See Figure 1 below for an example using a surgery intern's admission note. In this example the first row shows the document fully specified with values in each of the 5 axes. The second row shows how the same document could be less well specified, and be located in a higher place in the DO's hierarchy as simply as an "admission evaluation clinical note." This satisfies the minimal requirements of having the <Kind of Document>, plus one other axis, which in this example is <Type of Service>.

The work described in this paper augments the SMD with additional values and organizes it into a multiple hierarchy. This is a first step to customize the DO for local use with eNote.

Figure 1: Surgery Intern's note

<Subject Matter>	<Training/Professional Level	<Setting>	<Type of Service>	<Kind of Document>
Surgery	Intern	Hospital	Admission Evaluation	Clinical Note
			Admission Evaluation	Clinical Note

METHODS

Creation of Subject Matter Domain Model

The value list from the SMD axis from the HL7/LOINC effort was merged with a list of Approved Specialty Boards & Certificate Categories from the ABMS.[16] This merged SMD value list for the DO was then modeled in Protégé,[17] a tool for creating knowledge bases. The ABMS list is organized as a strict hierarchy with each subspecialty listed under the corresponding specialty, such that if a subspecialty is offered under more than one specialty, it is listed separately under each of them and the concept of multiple hierarchy is not modeled. The HL7/LOINC value list is flat with all specialties, subspecialties and other related areas within the SMD listed together without hierarchical organization. These two lists were merged by aligning the uppermost part of the ABMS list, namely the specialty terms, with the flat HL7/LOINC list. Initially this was done without revision.

Evaluation: Expert Review

Following the creation of the revised SMD model, a group of domain experts in Internal Medicine, General Surgery, Pediatrics, Preventative Medicine and Emergency Medicine were supplied with the list in Protégé format and asked for their input regarding the following 5 questions as applied to the SMD:

1) Does the hierarchy as presented seem to be modeled correctly?
2) Do you find any unintended redundancy, and if so where is it?
3) Is there anything missing that you believe should be added?
4) Do you think the hierarchy is generalizable outside of NYPH?
5) Any additional comments?

Evaluation: Input from Medical Specialties

Following the results of the expert review, an email survey of medical specialty boards was conducted. Additional input regarding medical specialty values was requested from 18 of the 24 specialty boards listed on the ABMS Web site. If the specialty board did not list an email address on its Web site, it was excluded from the survey. The email requested a list of additional subspecialties in the boards' respective areas based on 1) availability of a certification exam, 2) availability of fellowship training, or 3) general consensus among practitioners of the field.

Evaluation: Refinement and Validation

Once the results of the expert review and survey were obtained and reviewed, the SMD was further refined by adding new values and placing other values at the proper place in the hierarchy. This version of the SMD was validated in Protégé using the Racer tool.[18] Initially all the super-classes (specialties) were set to disjoint (no individual document could be listed with multiple specialties simultaneously), and subspecialties were allowed to have multiple parents.

Evaluation: Test of Document Coverage

All 163 document titles available in the Medical Entities Dictionary at NYPH at the time this study began were used to test the ability of the SMD model to fully specify documents based on their type and name.

First the documents were compared to the original HL7/LOINC SMD value list. If a document name or type did not suggest a specific subject matter (e.g., "Clinic Summary Report"), it was classified as "not specified." If a document name or type did suggest a specific subject matter, but it could not be specified because the subject matter was not included in the value list (e.g., "Electromyography Report Note"), it was classified as "other." In all other instances, the document was specified with the SMD value list.

This procedure was repeated for the ABMS hierarchy, and then for the merged and validated version of the SMD value list.

RESULTS

Expert Review

The expert review revealed the following: 1) that although the way in which the hierarchy was modeled is generally correct, numerous subspecialties needed to be pushed down in the hierarchy from the top level in order to model a more consistent level of abstraction; 2) that there was no redundancy; 3) that there were multiple specialties and subspecialties missing from the hierarchy that needed to be added; and 4) that the hierarchy was generalizable beyond NYPH.

Input from Medical Specialties

Of the 18 emails sent, 14 responses were received. Responses included an average of 5 subspecialties, with a range from 0 to 17. Certain subspecialties, namely Surgery, Plastic Surgery and Neurosurgery, listed multiple subspecialties that were not listed on the ABMS Web site but are generally regarded as subspecialties within their fields either by consensus or the availability of fellowship training in an area.

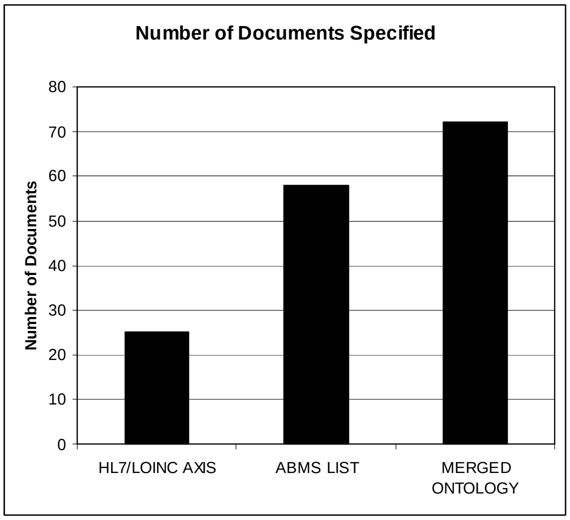
Model Validation

Initially a number of inconsistencies were picked up by the Racer tool. Once all disjoints from the super-class level were removed, the ontology was validated as consistent.

Test of Document Coverage

Of the163 documents, there were a total of 91 documents that were classified as "not specified" because the document type and name did not suggest a specific SMD. The remaining analysis relates to the 72 documents with explicit SMDs. Using the HL7/LOINC SMD value list, 25 documents' SMDs were specified and 47 were classified as "other." With the ABMS list, the SMDs of 58 documents were specified and 14 were classified as "other." The merged SMD had 100% coverage for the 72 documents. For two documents, the SMDs were re-specified to a lower level in the hierarchy. See Figure 2 below:

Figure 2



DISCUSSION

The validation problems initially experienced using the Racer tool were predictable, since many of the disjoint super-classes had common children, thereby making it impossible for them to be set as disjoint. This raises the question as to whether a strict hierarchy without multiple parents might be a more accurate model for the SMD. Because many of the subspecialties can be pursued from two or more specialty backgrounds, it seemed reasonable to model the SMD with multiple parents. However, in reality there is a potential difference in the implied subject matter if a note is written by a sub-specialist with one specialty background versus another. In other cases there is little if any implied difference in the subject matter of a note based on the individual's prior training.

Although the approach of merging the two value lists in our experiment improved specification of a list of sample documents, the SMD may still not be fully specified. Additions may need to be made in the future, but because of the architecture of the merged SMD in Protégé, this should be fairly straightforward. A formal process will need to be put in place to maintain the SMD and the DO as a whole.

Although the degree of granularity achieved by the merged and refined SMD is much greater than either of the two precursor value lists, the new SMD only allowed 2 of the documents from the sample list to be better specified, i.e., have their SMD names come from a deeper part of the hierarchy. It is reasonable to assume that with a larger set of documents, more documents could be specified with a greater degree of granularity. This will allow documents to be more accurately modeled in general, and will help support multiple consistent views of the SMD, allowing it to be used for multiple purposes.[19] These features should help the SMD and the DO become more easily generalized beyond one institution. Overall, the new SMD allows a given document to be specified in a more general category (e.g., internal medicine), or in a more specific category (e.g., thryroidology, which is a child of endocrinology, which is a child of internal medicine), depending on what is known about a document being entered into an EHR, or what criteria are being used in a document search.

CONCLUSIONS

The analysis methods should be applied to each of the other four axes in order to validate their value sets and to prove that existing documents can be properly specified and modeled when they are represented using the axes.

The primary goal of this endeavor is to standardize clinical document names so clinicians can easily find the documents and templates they need. If some form of DO is widely accepted, it will lead to standardization of document names across multiple institutions. This would allow clinicians to easily recognize document titles from locations other than their own and will help fuel the sharing of clinical documents between institutions.

This paper also hopes to bring more attention to the importance of the work being done by the HL7 DOTF and LOINC toward widely-distributed sharing of documents. More studies will need to be conducted to ensure that the DO they develop and approve is generalizable across multiple institutions and in multiple healthcare settings. The vision of a

National Health Information Infrastructure that allows patient records to be quickly and easily transferred between clinicians and institutions across the country will require such an ontology.

Acknowledgments

This research was supported by NIH/NLM 1R01 LM 07268-01 (S. Johnson, PI), T15 NLM 007079-13 (J. Shapiro & G. Melton) and T32 NR007799 (C. Schlegel). The authors thank the domain experts for their participation and the other members of the eNote working group for their advice and support, as well as the HL7 DOTF and LOINC for providing the foundation for the current study and for their continuing work in this important area.

References:

1. Hyun S. Bakken S. Friedman C. Johnson SB. Natural language processing challenges in HIV/AIDS clinic notes. Annual Symposium Proceedings/AMIA Symposium. :872, 2003.

2. Patel VL. Arocha JF. Kushniruk AW.. Patients' and physicians' understanding of health and biomedical concepts: relationship to the design of EMR systems. Journal of Biomedical Informatics. 35(1):8-16, 2002 Feb.

3. eNote project website. Columbia University Department of Biomedical Informatics Website. Available at http://lucid.cpmc.columbia.edu/enote/. Accessed: February 28, 2005.

4. Dolin RH, Alshuler L, Beebe C, et al. The HL7 clinical document architecture. J Am Med Inform Assoc. 20018:552-569.

5. HL7 Clinical Document Architecture, Release 2.0. [homepage on the Internet]. Ann Arbor: Health Level Seven, Inc. c2004 [updated 2004 Dec 12; cited 2005 Feb 28] Available from: http://www.hl7.org/v3ballot/html/infrastructure/cda/cda.htm.

6. Friedman C. A broad coverage natural language processing system. In Overhage M, editor. Proc AMIA Symp 2000: 2000:270-274.

7. MedLEE – A Medical Language Extraction and Encoding System. [homepage on the Internet]. New York: Columbia University. [updated 2004 Dec 7; cited 2005 Feb 28]. Available from: http://lucid.cpmc.columbia.edu/medlee/.

8. MEDical Language Extraction and Encoding System. [homepage on the Internet]. New York: Columbia University Department of Biomedical Informatics. [cited 2005 Feb 28]. Available from: http://www.dbmi.columbia.edu/homepages/campbed/Topics/TopicsWebSite/Introduction.htm.

9. The National Health Information Infrastructure. [homepage on the Internet]. Washington, D.C.: US Department of Health and Human Services. [cited 2005 Feb 28] Available from: http://www.aspe.hhs.gov/sp/nhii/index.html.

10. Brown SH, Lincoln M, Hardenbrook S, et al. Derivation and evaluation of a document-naming nomenclature. J Am Med Inform Assoc. 2001;8:379-390.

11. Huff SM, chair. Document Ontology Task Force. Proposal for an Ontology for Exchange of Clinical Documents. [monograph on the Internet] Ann Arbor: Health Level Seven, Inc. c1997-2005 [2000 July; cited 2005 Feb 28]. Available from: http://www.hl7.org/Special/dotf/docs/DocumentOntologyProposalJuly00.doc.

12. Document Ontology Task Force. [homepage on the Internet]. Ann Arbor: Health Level Seven, Inc. c1997-2004 [cited 2005 Feb 28]. Available at: http://www.hl7.org/special/dotf/dotf.htm.

13. Huff SM, Rocha RA, McDonald CJ, et al. Development of the Logical Observations Identifiers, Names, and Codes (LOINC) vocabulary. J Am Med Inform Assoc. 1998;5:276–92.

14. Logical Observation Identifiers Names and Codes (LOINC) website. [homepage on the Internet] Indianapolis: Regenstrief Institute, Inc. c2004 [cited 2005 Feb 28]. Available from: http://www.loinc.org.

15. Huff SM, Frazier P, Dolin RH. HL7 LOINC Document Type Vocabulary Domain Paper. [monograph on the Internet] Indianapolis: Regenstrief Institute, Inc. c2004 [2003 Sep 24; cited 2005 Mar 9]. Available from: http://www.regenstrief.org/loinc/discussion /.

16. American Board of Medical Specialties list of Approved ABMS Specialty Boards & Certificate. [homepage on the Internet]. Evanston: American Board of Medical Specialties. C2005 [cited 2005 Feb 28]. Available from: http://www.abms.org/member.asp.

17. Protégé 2000. [homepage on the Internet]. Stanford: Stanford Medical Informatics. C2003 [updated 2005 Mar 10, cited 2005 Mar 14]. Available from: http://protege.stanford.edu/.

18. Racer System. [homepage on the Internet]. Montreal: Concordia University. [cited 2005 Feb 28] Available from: http://www.cse.concordia.ca/%7Ehaarslev/racer/.

19. Cimino JJ, Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. Meth Inform Med 1998; 37: 394-403.