

An Update on HL7's XML-based Document Representation Standards

Robert H. Dolin, MD¹; Liora Alschuler²; Sandy Boyer, BSP; Calvin Beebe³
for the Kona Editorial Group

¹ Kaiser Permanente (Robert.H.Dolin@kp.org); ² The Word Electric; ³ Mayo Clinic

Many people know of HL7 as an organization that creates healthcare messaging standards. But HL7 is also developing standards for the representation of clinical documents (such as discharge summaries and consultation notes). These document standards comprise the HL7 Clinical Document Architecture (CDA). Last year we presented a high-level conceptual overview of the CDA¹. Since that time, CDA has entered HL7's formal ballot process (which when successful will make the CDA an ANSI-approved HL7 standard). This article delves into the technical details of the current CDA proposal. Note that due to space limitations, only a subset of CDA details can be described. Also, because the ballot process elicits considerable feedback, it is likely that the material presented here will undergo evolution prior to becoming a final standard. The most up-to-date information is available on HL7's web site (www.hl7.org).

INTRODUCTION

The HL7 Clinical Document Architecture (CDA) (previously known as the Patient Record Architecture, PRA) is a document markup standard that specifies the structure of exchanged clinical documents. A CDA document is a defined and complete information object that can include text, images, sounds, and other multimedia content, and that can be transported within an HL7 message.

Key aspects of the CDA include: (1) CDA documents are encoded in Extensible Markup Language (XML); (2) CDA documents derive their meaning from the HL7 Reference Information Model³ (RIM)* and use RIM data types; (3) The complete CDA will include a

* The RIM is HL7's information model of the healthcare domain. All future HL7 message and document specifications will be derived from the RIM. The definition of every field in every future HL7 specification can be traced back to the RIM, which is the ultimate defining reference.

hierarchical set of document specifications. This hierarchy is referred to as an "architecture".

CDA CONCEPTS

The CDA is envisioned as an extensible and hierarchical set of document specifications, that taken together comprise an "architecture". Each specification is represented by a distinct XML DTD (or in the future by a distinct XML Schema⁴). Use of an architecture avoids imposition of unacceptably rigid constraints on local document implementation and facilitates interoperability.

"Levels" within the architecture represent a quantum set of specializations, to which further constraints can be applied. These levels establish baselines for conformance claims, baselines that standardize at stepwise levels of greater shared meaning thereby enabling the computer-processable expression of richer shared semantics. CDA Level One is the specification at the root of the architecture, and specifies the semantics of the header. Codes for the document type (such as "discharge summary", "consultation note") and for the sections contained in the document body (such as "chief complaint", "review of systems") can be expressed. CDA Level Two uses the same codes for document types and sections, and will enable the ability to constrain the set of allowable sections based on document type. CDA Level Three defines observations and services within the document body, with a granularity as rich as is possible with the RIM.

CDA TECHNICAL SPECIFICATIONS

Introduction

The current ballot proposal contains the conceptual framework for all of CDA, and the technical specifications for CDA Level One. Technical specifications for CDA Level Two and CDA Level Three are currently in development.

CDA Level One conceptually consists of three components: The CDA Header, the CDA Level One Body, and the HL7 Version 3 data types.

Figures 1 and 2 show a sample CDA Level One document instance. Figure 1 shows how the report might typically appear when viewed or printed, and Figure 2 shows the underlying XML representation. Note that the samples are modified for illustration purposes, and some required components may be missing.

Coded CDA components have associated "vocabulary domains" which represent allowable value sets for the component. These domains can include HL7-defined concepts or can be drawn from HL7-recognized coding systems such as LOINC or SNOMED. For example, line 21 in Figure 2 states the patient's gender code:

```
<gender_cd v="M"/>
```

In this case, the value "M" is drawn from the vocabulary domain associated with the gender_cd component.

HL7 Data Types

HL7 is currently balloting a set of data types (including familiar types such as STRING, INTEGER, TIME STAMP, CODED VALUE and some new types such as CONCEPT DESCRIPTOR, ENCAPSULATED DATA) for use with their Version 3 and CDA standards. The XML representation of these data types models the data type subcomponents as XML attribute values.

For example, Figure 2, line 4 shows document.service_cd, which is declared to be of type CODED VALUE:

```
<document.service_cd v="11488-4"  
  S="LOINC" PN="Consultation note"/>
```

The CODED VALUE data type has a component "V" for the code value, a component "S" for the source of the code value, and a component "PN" for the print name corresponding to the code value.

Line 49 uses the ENCAPSULATED DATA data type to reference an external image that is logically part of the CDA document:

```
<observation_media.value  
  MD="image/jpeg" ADR="rash.jpeg"/>
```

The "MD" component stands for media descriptor, and draws its vocabulary domain from the set of MIME types. The "ADR" component is for a valid URL.

CDA Header

The purpose of the CDA Header is to enable clinical document exchange across and within institutions; facilitate clinical document management; and facilitate compilation of an individual patient's clinical documents into a lifetime electronic patient record.

Document metadata identifies the document, defines confidentiality status, and describes relationships to other documents and orders. Encounter data describes the setting in which the documented encounter occurred. Service actors include those who authenticate the document, those intended to receive a copy of the document, document originators and transcriptionists, and health care providers who participated in the service(s) being documented. Service targets include the patient, other significant participants (such as family members), and those devices that may have originated portions of the document.

An example CDA header instance is in Figure 2, lines 2-25. XML elements in the header map back to structures in the RIM. The process of deriving XML DTDs from the RIM is described in the HL7 Message Development Framework⁵.

Locally-defined markup must be used when local semantics have no corresponding representation in the CDA specification. CDA seeks to standardize the highest level of shared meaning while providing a clean and standard mechanism for tagging meaning that is not shared. This is achieved with the CDA "local_header" element (see line 24 in Figure 2).

CDA Body

The CDA Level One body is comprised of nested containers. Containers (including sections, paragraphs, lists, and tables) have captions and contents, and may be coded. In addition, each container has a "confidentiality" and an "origination" attribute. The language of character data can be specified using the "xml:lang" attribute as defined in the XML 1.0 Recommendation. The value(s) specified for these attributes are assumed to be inherited by nested containers, unless overridden.

A CDA <section> can contain "structures", nested <section> elements, and <coded_entry> elements. CDA structures include the <paragraph>, <list>, and <table> elements. These structures contain CDA "entries", which include the <content>, <link>, <coded_entry>, <observation_media>, and <local_markup> elements in addition to plain character data.

Line 27 of Figure 2 shows a section containing a LOINC code that represents "History of Present Illness":

```
<section V="8684-3" S="LOINC">
  <caption>
    History of Present Illness
  </caption>
```

The CDA uses a modified XHTML table model⁶ by removing formatting tags and by setting the content model of cells to be similar to the contents of other CDA containers.

The CDA element <coded_entry> (e.g. Figure 2, line 62) enables the insertion of codes from HL7-recognized coding schemes into CDA documents. Where there are no suitable HL7-recognized codes available, locally-defined codes can be used. The domain of allowable codes is unspecified, and the primary intent of <coded_entry> is to facilitate document indexing, search and retrieval.

The <link> element is currently based on the HTML anchor tag. Several groups are actively developing formal link specifications (such as Xlink⁸). When a suitable open standard is available, it will be reviewed with the intent to incorporate it into the CDA specification.

The CDA element <observation_media> is derived from the RIM "Observation" class, and is used to represent media (such as images) that is logically a part of a CDA document, but is stored outside the document and referenced. Multimedia that is integral to a document, and part of the attestable content of the document, requires the use of <observation_media>. Multimedia that is simply referenced by the document and not an integral part of the document should use <link>.

The CDA <content> element can nest recursively, which enables wrapping an arbitrary string of plain text down to the individual character. These <content> tags serve as anchors,

and <coded_entry> elements can reference these anchors to indicate the original text that supports the use of the code, in a fashion similar to that described by Friedman, et al⁷.

Kaiser Permanente Consultation Report

Dictated By: Robert Dolin, MD
Date: December 7, 1999
Patient: Henry Levin, the 7th
MRN: 12345
Sex: Male
Birthdate: September 24, 1932


History of Present Illness
 Henry Levin, the 7th is a 67 year old male referred for further asthma management. Onset of asthma in his teens. He was hospitalized twice last year, and already twice this year. He has not been able to be weaned off steroids for the past several months.

Medications

- Proventil inhaler 2puffs QID PRN
- Prednisone 20mg qd

Physical Exam

- Vital Signs :: BP 118/78; Resp 16 and unlabored; HR 86 and regular.
- Skin :: Erythematous rash, palmar surface, left index finger.



- Lungs :: Clear with no wheeze. Good air flow.

Labs

- Peak Flow today: 260 l/m.

Assessment

- Asthma, with prior smoking history. Difficulty weaning off steroids. Will try gradual taper.

Plan

- Decrease prednisone to 20qOD alternating with 18qOD.

Signed by: Robert Dolin, MD on December 8, 1999

Figure 1. Sample document instance.

```

1 <CDA_document>
2 <clinical_document_header>
3 <document.id V="a123" AA="100.12.92.81.5.7"/>
4 <document.service_cd V="11488-4" S="LOINC" PN="Consultation note"/>
5 <originator><tmr V="19971207"/>
6 <person><id V="KP00017" AA="100.12.92.81.5.7"/>
7 <nm><G V="Robert" CLAS="R"/><F V="Dolin" CLAS="R"/><SF V="MD" CLAS="PT"/></nm>
8 </person>
9 </originator>
10 <originating_organization>
11 <organization><id V="M345" AA="100.12.92.81.5.7"/>
12 <organization_nm V="Kaiser Permanente"/>
13 </organization>
14 </originating_organization>
15 <legal_authenticator><tmr V="19991208"/>
16 <person><id V="KP00017" AA="100.12.92.81.5.7"/></person>
17 </legal_authenticator>
18 <patient>
19 <person><id V="12345" AA="100.12.92.81.5.7"/>
20 <nm><G V="Henry" CLAS="R"/><F V="Levin" CLAS="R"/><SF V="the 7th" CLAS="R"/></nm>
21 <birth_dttm V="19320924"/><gender_cd V="M"/>
22 </person>
23 </patient>
24 <local_header ignore="all" descriptor="MyLocalTag">... extra stuff ...</local_header>
25 </clinical_document_header>
26 <body>
27 <section V="8684-3" S="LOINC"><caption>History of Present Illness</caption>
28 <paragraph>
29 <content>Henry Levin the 7th is referred for asthma..</content>
30 </paragraph>
31 </section>
32 <section V="10160-0" S="LOINC"><caption>Medications</caption>
33 <list>
34 <item><content>Proventil inhaler 2puffs QID PRN</content></item>
35 <item><content>Prednisone 20mg qd</content></item>
36 </list>
37 </section>
38 <section V="11384-5" S="LOINC"><caption>Physical Examination</caption>
39 <section V="8716-3" S="LOINC"><caption>Vital Signs</caption>
40 <list>
41 <item><caption>BP</caption><content>118/78</content></item>
42 <item><caption>Resp</caption><content>16 and unlabored</content></item>
43 <item><caption>HR</caption><content>86 and regular</content></item>
44 </list>
45 </section>
46 </section>
47 <section V="8709-8" S="LOINC"><caption>Skin</caption>
48 <paragraph>
49 <content>Erythematous rash, palmar surface, left index finger.
50 <observation_media.value MD="image/jpeg" ADR="rash.jpeg"/>
51 </content>
52 </paragraph>
53 </section>
54 <section V="11391-0" S="LOINC"><caption>Lungs</caption>
55 <paragraph><content>Clear with no wheeze. Good air flow.</content></paragraph>
56 </section>
57 <section V="11502-2" S="LOINC"><caption>Labs</caption>
58 <paragraph><caption>Peak Flow</caption><content>260 l/m</content></paragraph>
59 </section>
60 <section V="11496-7" S="LOINC"><caption>Assessment</caption>
61 <paragraph>
62 <content>Asthma...<coded_entry V="D2-51000" S="SNOMED"/></content>
63 </paragraph>
64 </section>
65 <section V="1234-X" S="LOINC"><caption>Plan</caption>
66 <paragraph>
67 <content>Decrease prednisone to 20qOD alternating with 18qOD.</content>
68 </paragraph>
69 </section>
70 </body>
71 </CDA_document>

```

Figure 2. Sample PRA document instance.

As a result, line 62 of Figure 2, which currently just has an embedded code:

```
<content>
  Asthma...
  <coded_entry V="D2-51000" S="SNOMED"/>
</content>
```

could be modified to have the code unambiguously reference the text:

```
<content>
  <content ID="S001">Asthma</content>
  with prior smoking history.
  Difficulty weaning off steroids.
  Will try gradual taper.
  <coded_entry ORIGTXT="S001"
    V="D2-51000" S="SNOMED"/>
</content>
```

CONCLUSIONS

Looking over Figure 2, it's easy to see areas where more detailed markup is possible (such as coding the components of the vital signs). In fact, the RIM does include a detailed model of services and observations, and therefore such functionality will ultimately be part of CDA Level Three. But we have to recognize the balance between standardizing the contents of clinical notes with highly detailed markup, and the cost required to enter such markup. Keeping provider-based documentation time unchanged while enabling a computer to understand more and more of the contents of a clinical note is an iterative process. CDA Level One standardizes at a level where broad participation is likely to be feasible. Later, we standardize more, especially as new data entry techniques are developed.

Ultimately, decision support applications will benefit from having an HL7 interface, which will enable them to interoperate automatically with electronic health records. The CDA document standards from HL7 take us a major step forward by standardizing the contents of health records. Standards for data representation will progressively enable two different computer applications to speak the same language and thereby hold an intelligent conversation, where each application correctly interprets the utterings of the other. But realistically, we still have a lot of work to do to achieve this.

Acknowledgements

The Kona Editorial Group (KEG) is a working group of HL7's Structured Document committee charged with reconciling and incorporating all of the comments and requirements into the CDA specification. The collaborative spirit and the enormous personal contributions made by the KEG members are greatly appreciated. Present and past KEG members include:

Carl Adler, Fred Behlen, Dean Bidgood, Paul Biron, Carol Broverman, Ron Capwell, John Carter, Michal Coleman, Don Connelly, Steve Doubleday, Joachim Dudeck, Dan Essin, Jasen Fici, Lloyd Harding, Juggy Jagannathan, Ed Jones, Eliot Kimber, Tom Lincoln, John Majerus, John Mattison, Bob Moe, Wes Rishel, Anil Sethi, Rachael Sokolowski, John Spinosa, Michael Toback, Jason Williams.

References

1. Dolin R, Alschuler L, Behlen F, Biron P, Boyer S, Essin D, Harding L, Lincoln T, Mattison J, Rishel W, Sokolowski R, Spinosa J, Williams J. HL7 Document Patient Record Architecture: an XML document architecture based on a shared information model. Fall AMIA, November 1999.
2. Extensible Markup Language (XML) 1.0. W3C Recommendation 10-February-1998. [www.w3.org/TR/1998/REC-xml-19980210.html]
3. HL7 Reference Information Model. [www.hl7.org/library/data-model/RIM/modelpage_non.htm]
4. XML Schema Part 1: Structures; Part 2 Data types. W3C Working Draft. [www.w3.org/TR/xmlschema-1/] [www.w3.org/TR/xmlschema-2/]
5. HL7 Version 3 Message Development Framework, 1999. [<http://www.hl7.org/Library/mdf99/mdf99.pdf>]
6. XHTML 1.0. A Reformulation of HTML 4 in XML 1.0. W3C Recommendation 26-January-2000. [www.w3.org/TR/xhtml1/]
7. Friedman C, Hripcsak G, Shagina L, Liu H. Representing information in patient reports using natural language processing and the Extensible Markup Language. JAMIA 1999;6(1):76-87.
8. XML Linking Language (XLink). W3C Working Draft.