SGML and XML as Interchange Formats for HL7 Messages

Robert H. Dolin, MD[1]; Wes Rishel[2]; Paul V. Biron, MLIS[1];
John Spinosa, MD, PhD; John E. Mattison, MD[1]
[1]Kaiser Permanente, Southern California; [2]Wes Rishel Consulting

OBJECTIVE: *To report on the use of SGML and XML (a proper subset of SGML) as transfer syntaxes for HL7 Version 2.3 and Version 3.0 messages.* ***METHODS:*** *The methodology has focused largely on two questions: Can it be done? How best to do it? The first question is addressed by attempting to build an SGML/XML representation of HL7 messages. The second question requires a consideration of several metrics: message length, speed of message creation and parsing, interversion compatibility, local customization, conformance determination, and the availability of software tools and skill on the format.* ***RESULTS:*** *Detailed specifications for expressing HL7 in SGML and XML have been developed. Some HL7 requirements are not readily expressed, while some ambiguous areas of the HL7 standard are made explicit in the SGML/XML representation. With the current design, an SGML/XML parser can extract any component of any data type from a message.* ***CONCLUSIONS:*** *SGML and XML can both serve as implementable message specifications for HL7 Version 2.3 and Version 3.0 messages. The ability to explicitly represent an HL7 requirement in SGML/XML confers the ability to validate that requirement with an SGML parser. The optimal message representation will be a balance of functional, technical, and practical requirements.*

INTRODUCTION

In 1996, the HL7 SGML initiative evolved as a special interest group (SIG) of HL7. This HL7 SGML SIG[1], which in 1997 was renamed the HL7 SGML/XML SIG, is interested in coordinating the development of a comprehensive document architecture for healthcare; educating the healthcare community in the capabilities and utility of SGML-based information; developing, coordinating, and maintaining a framework for the interoperability of healthcare documents in an open, structured manner; coordinating and cooperating with other SGML initiatives where appropriate; and investigating the use of SGML/XML as a messaging syntax. The architectures put forth by the HL7 SGML/XML SIG will be in conformance with the evolving HL7 healthcare data model.

Our objective here is to report on the use of SGML and XML as transfer syntaxes for HL7 Version 2.3 and Version 3.0 messages. In 1993, The European Committee for Standardization (CEN) studied several syntaxes (including ASN.1, ASTM, EDIFACT, EUCLIDES, and ODA) for interchange formats in healthcare[2]. A subsequent report extended the CEN study to look at SGML[3]. By using the same methodology, example scenarios, healthcare data model, and evaluation metrics, the report presented a direct comparison of SGML with the other syntaxes studied by CEN, and found SGML to compare favorably. None of the interchange formats were able to explicitly represent all functional requirements.

SGML (Standard Generalized Markup Language, ISO 8879:1986) reduces a document or message to a word in a known context-free grammar through a process of markup. The formal markup specification for a collection of documents is called a Document Type Definition (DTD). Documents or messages are then written to conform to a particular DTD, enabling them to be automatically parsed and validated against that DTD. Some introductory tutorials on SGML and some further background on the use of SGML in healthcare can be found in the references[4-6].

A recently emerging standard, XML (Extensible Markup Language, www.w3.org/TR/1998/REC-xml-19980210), is a proper subset of SGML (meaning that all XML documents are valid SGML documents) that excludes some of the more obscure or less-used features[7]. By enabling authors to define their own DTDs, XML greatly extends the types of documents that can be delivered over the Internet. As a result, next generation web browsers, including those from Netscape and Microsoft, will directly support the display and manipulation of XML documents. The discussions in this document are equally

1091—8280/98/$5.00

applicable to both SGML and XML unless stated otherwise.

HL7 is the leading healthcare messaging standard in the United States. Version 2.3 is in current use, while the draft Version 3.0 Message Development Framework is being continuously refined[8]. We have developed detailed draft specifications for expressing HL7 Version 2.3 and Version 3.0 messages in SGML and XML[9,10]. This report summarizes these detailed specifications and describes where further efforts are ongoing.

METHODS

The methodology has focused largely on two questions: Can it be done? How best to do it? Answering whether or not an SGML representation of HL7 messages can be done has been addressed by building an algorithm to do so. The next question of how best to represent HL7 messages in a DTD requires a consideration of several metrics, some of which can be addressed quantitatively, while others at this time are only addressed qualitatively.

Can it be done?

The approach taken for HL7 Version 2.3 was to explicitly express to the extent possible the constraints and requirements of the HL7 standard in a DTD. In Version 2.3, messages are defined based on their segments using the HL7 Abstract Message Syntax. An equivalent representation in an XML DTD is shown in [Figure 1]. Segments are then defined based on their fields. [Figure 2] shows a portion of an imaginary segment and how that segment definition is expressed as an XML element definition in a DTD. Each row of the table represents a field or slot within the segment. Column "DT" represents the HL7 data type of the particular field. Column "OPT" represents the optionality of the field. Column "RP" specifies whether or not the field may have more then one value. The values of "OPT" and "RP" together determine the XML occurrence indicators, as shown in [Figure 2]. The XML content model for a field is its HL7 data type. The content model for a compound data type is its data type components. [Figure 3] shows a sample message along with an equivalent XML instance representation.

HL7 Abstract Message Syntax		Equivalent XML
WRP	**Widget Report**	<!ELEMENT **WRP** (
MSH	Message Header	MSH,
MSA	Message Ack	MSA,
{ WDN	Widget Desc	(WDN,
WPN	Widget Portion	WPN,
{[WPD]}	Widget Detail	(WPD)*
}		)+
		)>

Figure 1. Example of an HL7 V2.3 message definition and an equivalent XML element definition expressed in a DTD.

WDN Widget Description Segment				Equivalent XML
DT	OPT	RP	NAME	<!ELEMENT WDN (
ID	R	N	Widget ID	WidgetID,
ST	R	Y	Widget Name	WidgetName+,
ST	O	Y	Widget Nickname	WidgetNick*,
ST	O	N	Widget Producer	WidgetProd?)>

Figure 2. Example of an HL7 segment definition and an equivalent XML element definition expressed in a DTD. ("N"=Cannot Repeat, "O"=Optional, "R"=Required, "Y"=Can Repeat)

The Version 3.0 methodology[8] is built upon 'use cases' or scenarios that demonstrate where a healthcare message is to be used, what information needs to be conveyed in the message, and who the message participants are. These use cases lead to the enhancement and validation of the HL7 Reference Information Model (RIM)[11], a global model that represents those objects and attributes used in healthcare messaging. Basing the information in all messages on the HL7 RIM helps ensure that the semantics of a message, as intended by the sender, are properly interpreted by the receiver. Use cases also lead to the development of message 'interaction models' and 'message information models', and ultimately culminate in the creation of a Hierarchical Message Descriptor (HMD) for each message to be sent.

The approach for Version 3.0 was to develop an algorithm (or "Implementation Technology Specification") to convert an arbitrary HL7 Version 3.0 HMD into a DTD. The DTD then represents an implementable message specification. [Table 1] shows how the components of an HMD are mapped into SGML structures. The "Information Model Mapping" section represents a mapping from the objects of the HL7 RIM into the HMD. "Multiplicity" refers to the number of times an object may repeat. "Data Type" is the HL7 data type. The "Message Elements" section represents the

```
MSH|...<cr>
MSA|...<cr>
WDN|wdgx.325.7890353|Widget Wonder|Widget Wadget~Wadget Widget Wonder|Widget Wonder Makers of America<cr>
WPN|...<cr>
WPD|...<cr>
WPD|...<cr>
```

```
<WRP>
<MSH>...</MSH>
<MSA>...</MSA>
<WDN>
  <WidgetID><ID v="wdgx.325.7890353"/></WidgetID>
  <WidgetName><ST v="Widget Wonder"/></WidgetName>
  <WidgetNick><ST v="Widget Wadget"/></WidgetNick>
  <WidgetNick><ST v="Wadget Widget Wonder"/></WidgetNick>
  <WidgetProd><ST v="Widget Wonder Makers of America"/></WidgetProd>
</WDN>
<WPN>...</WPN>
<WPD>...</WPD>
<WPD>...</WPD>
</WRP>
```

Figure 3. Sample message encoded per HL7 encoding rules and an XML instance representation.

segments and fields that can occur in a particular message, and how they correspond to RIM items. "Nesting" specifies segment containment. "Structure" specifies where optionality, lists, and groups may occur. "Segment Slot Type / Tag Value" and "Slot Name" name the segments and fields. "Shared Type?" references fields defined elsewhere. "Data Field Domain Spec" specifies table value restrictions. The "Message Elements" section differentiates messages that share information in the preceding sections. "Conditional Presence" defines the logic dictating the occurrence of values for a particular field. "Required Value" constrains a field to a particular value. "Inclusion" states whether or not null values are allowable. "Repetitions" is similar to "Multiplicity".

How best to do it?

There are many ways to map HL7 into either SGML or XML syntax. The optimal approach will be in part determined by the technical mapping requirements noted above, and in part determined by practical implementation issues. Some of these issues are explored more quantitatively (e.g., message length), while others remain more qualitative / needing further analysis (e.g., speed of message creation). The relative weight of each of these optimization considerations has not been addressed.

Message length minimization techniques are employed to decrease the total number of characters (including data and/or markup) comprising a message. The optimal techniques used to minimize SGML messages are not necessarily the same as those best suited to minimize XML messages. Techniques common to both SGML and XML include the use of abbreviations, assuming that a slot not sent represents a null value, and in some cases modeling components as SGML attributes as opposed to elements. Full SGML provides even greater minimization capacity with the use of SHORTTAG, OMITTAG, and SHORTREF techniques, resulting in very small messages that are not valid XML.

Processing speed considerations include the time for message creation, parsing, validation, filtering, augmenting (e.g., changing the format of a field or data type component), and routing. Each of the minimization techniques potentially has an impact on subsequent processing speed. As an example, one can minimize an SGML message by leaving double-quotes off an attribute's string value. However if the string contains internal white space, the double-quotes are required. Thus the processor has to scan an entire string before determining if double-quotes can be left off. The price of minimization is increased message creation time. Other metrics include conformance determination, compatibility with future versions of the HL7 standard, and the availability of software tools and skill on the format.

RESULTS

Detailed draft specifications for expressing HL7 Version 2.3 and Version 3.0 in SGML and XML have been developed and are available on the HL7 SGML/XML SIG web site[9,10]. The Version 2.3 mapping has been undergoing considerable

Table 1. Mapping the components of the HL7 Version 3.0 Hierarchical Message Descriptor into SGML.

HMD	SGML
Information Model Mapping	
Multiplicity	Combined with "Inclusion", determines SGML occurrence indicators (for Union messages only).
Data Type	Modeled as an SGML element whose content model is the data type components.
Message Elements	
Nesting	Specifies SGML element containment.
Structure	Choice, List, and Group structures are modeled as SGML container elements.
Segment Slot Type / Tag Value	These segment names become SGML elements.
Shared Type?	Common Message Element Definitions (CMEDs) for Segment Expressions are assumed to be fully expanded in the HMD. CMEDs for Data Types are modeled like other data types.
Slot Name	These (abbreviated) field names become SGML elements.
Data Field Domain Spec	SGML attribute value restriction.
Message Structures	
Conditional Presence	Not represented in the DTD.
Required Value	Can be modeled as an SGML attribute value restriction.
Inclusion	Combined with "Multiplicity" (in the case of a Union message) or "Repetitions", determines SGML occurrence indicators.
Repetitions	Combined with "Inclusion", determines SGML occurrence indicators.

scrutiny as it is the object of a prototype implementation project that was initiated in March 1998. The Version 3.0 mapping was revised in February 1998 to reflect the January 1998 revision of the HL7 Version 3.0 Message Development Framework. It too is undergoing considerable scrutiny and refinement thanks to broader input from the HL7 community.

We are not yet able to fully automate either DTD generation from the HL7 Standard or the transformation of standard-encoded HL7 messages into equivalent SGML-encoded messages. While the process of going from HL7 tables within the Standard to a DTD is automated, there are aspects of the Standard that are not formally expressed within tables. We are continuing to identify these aspects and determine how they should be represented. Some of our findings are described below.

The Abstract Message Syntax [Figure 1] can describe an ambiguous ordering of segments within a message. For instance, in the imaginary message MSG, the syntax may state that segment SEG1 may be followed by SEG2, and then optionally by SEG3, then optionally by SEG2 again. The SGML equivalent of this is: <!ELEMENT MSG (SEG1, SEG2?, SEG3?, SEG2?)>, which is invalid because the parser, upon seeing the presence of SEG2 occurring in a message following SEG1, is unable to determine which SEG2 is being referred to. In addition, some HL7 message definitions have to be gleaned in part by reading the textual description of the message in addition to the formal Abstract Message Syntax expression of that definition. For instance, message MFN (Master File Notification) contains " {MFE [Z..] } " in its definition. The "Z.." represents "one or more HL7 and/or Z-segments...".

Fields are defined based on their data types. The "CM" (Composite) data type is "a field that is a combination of other meaningful data fields", and that combination can be distinct for each use of the data type. Thus, we've made each use of "CM" into a distinct data type. Fields of "Variable" data type have their data type specified in another field of the message. "Variable" data types are currently modeled as 'or-groups' listing the possible data types allowed for that field.

There are instances where proper use of the HL7 Standard is described predominately in text. For example, field 3 of the OBX segment (OBX-3 Observation Identifier) is of data type CE (coded entry), thus it's first component is a ST (string) data type representing the identifier of the observation. Chapter 7 of the Standard describes "code suffixes for constructing observation IDs

for the common components of narrative reports". As a result, the string in the first component of OBX-3 can have an internal subcomponent delimiter separating the code from the suffix, which is not legal in the ST data type.

CONCLUSIONS

We have found that SGML and XML can both serve as implementable message specifications for HL7 Version 2.3 and Version 3.0 messages and that the ability to explicitly represent an HL7 requirement in SGML confers the ability to validate that requirement with an SGML parser. We have found both HL7 requirements that are difficult to express, and HL7 ambiguities that we have made explicit by the SGML representation.

SGML and XML messages will likely be longer then current HL7 messages. The greater the percentage of data characters (as opposed to markup characters) in an average message, the less important the overhead imposed by markup becomes. Data from the Duke HL7 production environment suggests that on average, data characters comprise 70% of overall message length. (Data from Duke courtesy of Al Stone, and posted to the HL7 SGML/XML SIG List Server 1/15/98 and 1/16/98.) Given this, and depending on the minimization techniques employed, XML messages may be approximately 40% to 100% larger then current messages, while SGML messages can actually be a little shorter. Because message length is a fairly straight-forward metric to quantify, there is a risk that it will assume significance out of proportion to the other metrics, all of which have to be factored together to determine the optimal SGML/XML representation.

As noted above, a prototype implementation of an SGML/XML representation of HL7 Version 2.3 is underway. The intention for Version 2.3 is to ultimately produce an HL7 SGML/XML SIG-sponsored draft that can be submitted to the broader HL7 community for further consideration and ultimately for balloting as an informative document. Recent discussions on the HL7 Control Query listserver suggest that XML may become the syntax of choice for Version 3 messages. As of this writing (June, 1998), interest in XML for Version 3 is growing rapidly, and the interested reader is likely to find the most up to date discussions taking place on the HL7 CQ listserver.

In conclusion, SGML and XML can serve as interchange formats for HL7 messages. The optimal message representation will be a balance of functional, technical, and practical requirements.

Acknowledgments

Special thanks to the HL7 SGML/XML Special Interest Group.

References

1. HL7 SGML/XML Special Interest Group web site. (www.mcis.duke.edu/standards/HL7/sigs/sgml/)
2. European Committee for Standardization / Technical Committee 251 - Medical Informatics. Investigation of Syntaxes for Existing Interchange Formats to be used in Healthcare. CEN/TC251. January 1993.
3. Dolin RH, Alschuler L, Bray T, Mattison JE. SGML as a message interchange format in healthcare. JAMIA Fall Symposium Supplement 1997: 635-9.
4. Goldfarb CF. The SGML Handbook. Y. Rubinsky, ed. Oxford Univ. press, 1990.
5. Text Encoding Initiative. A gentle introduction to SGML. (www.sil.org/sgml/gentle.html)
6. Alschuler L, Dolin RH, Spinosa J. SGML in healthcare information systems. Proceedings of the GCA SGML/XML Conference, 1997.
7. Khare R, Rifkin A. XML: A door to automated web applications. IEEE Internet Computing 1997; 78-87.
8. HL7 Version 3.0 Message Development Framework, Version 3.1 January 1998. (www.mcis.duke.edu/standards/HL7/pubs/version3/Version3.htm)
9. HL7 SGML/XML SIG. SGML/XML as an interchange format for HL7 V2.3 messages. DRAFT, April 1998. (www.mcis.duke.edu/standards/HL7/sigs/sgml/WhitePapers/hl7v2/hl7sgml2.rtf)
10. HL7 SGML/XML SIG. SGML/XML as an interchange format for HL7 V3.0 messages. DRAFT, February 1998. (www.mcis.duke.edu/standards/HL7/sigs/sgml/WhitePapers/hl7v3/hl7sgml3.rtf)
11. HL7 Reference Information Model (DRAFT) (www.mcis.duke.edu/standards/HL7/data-model/HL7/modelpage.html)