SGML as a Message Interchange Format in Healthcare

Robert H. Dolin, MD [1]; Liora Alschuler [2]; Tim Bray [3]; John E. Mattison, MD [1]

[1] Kaiser Permanente, Southern California Region (Robert.H.Dolin@kp.org); [2] The Word Electric; [3] Textuality

INTRODUCTION: In 1993, The European Committee for Standardization (CEN) studied several syntaxes for interchange formats in healthcare, but excluded SGML due to resource constraints. We sought to extend the CEN report and formally evaluate the use of SGML as a message interchange format. METHODS: We followed the methodology set forth by CEN, using their example scenarios and healthcare data model. General message descriptions based on this model set the functional requirements for the interchange format. These general requirements are then mapped into SGML to see how well they can be supported. RESULTS: Results follow the CEN format, enabling a direct comparison of SGML with ASN.1, ASTM E1238, EDIFACT, EUCLIDES, and ODA (those syntaxes studied by CEN). CONCLUSION: SGML compares favorably with other syntaxes investigated by CEN. None of the interchange formats support all functional requirements. Optimal and standard mechanisms of combining different formats through a modular approach to achieve greater overall functionality requires further study.

INTRODUCTION

In 1996, the HL7 SGML initiative evolved as a special interest group (SIG) of HL7. This HL7 SGML SIG[1] 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; and coordinating and cooperating with other SGML initiatives where appropriate. The architectures put forth by the HL7 SGML SIG will be in conformance with the evolving HL7 healthcare data model.

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). Some introductory tutorials on SGML can be found in the references[2,3].

In January, 1993, the European Committee for Standardization (CEN) issued a report, "Investigation of Syntaxes for Existing Interchange Formats to be used in Healthcare"[4]. Their objective was to "investigate syntaxes for existing interchange formats (IFs) to be used in healthcare and to define a strategy for selecting IFs". The 23 identified formats were prioritized, and those five formats with a priority one (ASN.1, ASTM E1238, EDIFACT, EUCLIDES, and ODA) were evaluated in detail. SGML was given a priority two, and was not evaluated due to resource constraints.

Since the time of the CEN report, the use of SGML in healthcare continues to grow. There are currently four primary areas of application for SGML in healthcare: medical publishing, new drug submissions, clinical practice guidelines, and patient health records. "MD Consult" (www.mdconsult.com) brings together the three largest medical publishers to present over 140 journals and reference texts from an SGML source. Several pharmaceutical companies have joined the Mulitagency Electronic Regulatory Submission Project, an effort to create standards for the use of SGML in international electronic drug submissions[5]. SGML is used for clinical practice guidelines at Kaiser Permanente, the National Library of Medicine (Reenie Prettyman, private correspondence), and other sites. Use of SGML for patient records is in its intial stage with experimental work reported by vendors and users[6-8].

Given the increasing popularity of SGML, we sought to study the use of SGML as a message
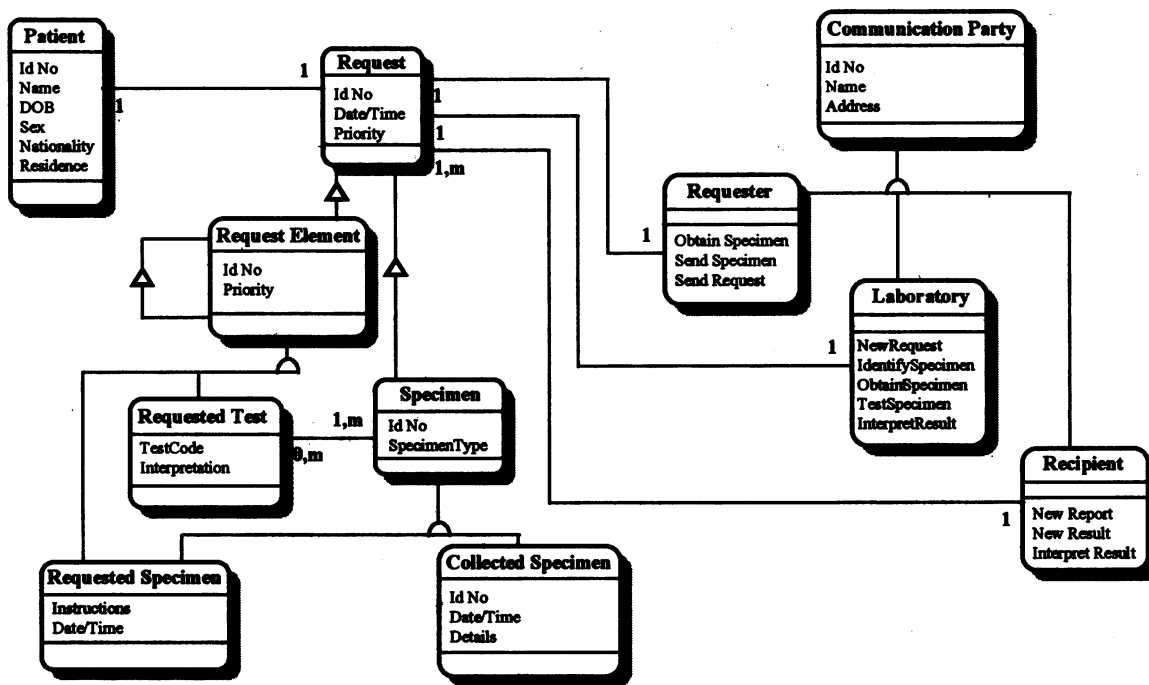
FIGURE 1. Object model of example scenario (modified from CEN[4] with permission).

interchange format, following the methods, example scenarios, and reporting format of CEN, enabling a direct comparison of SGML with ASN.1, ASTM E1238, EDIFACT, EUCLIDES, and ODA.

METHODS

The methods used by CEN in the formation of healthcare messages are similar to those being used in the HL7 Version 3 Message Development Framework draft[9]. Steps include: (1) Scope out the problem domain; (2) Determine parties and communication roles; (3) Scenario development; (4) Determine services associated with each communication role; (5) Develop a Domain Information Model (DIM) with specification of attributes and operations; (6) Formulate General Message Descriptions (GMDs). Through the use of this formal process, messages are kept consistent with the underlying healthcare data model. The GMDs are independent of, and set the functional requirements for the interchange format. The ability of the various syntaxes to support the functional requirements of the GMDs are then evaluated.

For the example scenario used by CEN and in this report, the scope is limited to the

interchange of information from a requester to a laboratory and from the laboratory to a result recipient. Communication roles include a Requester, a Laboratory, and a Recipient. The requester issues a Request to the laboratory in order to get a set of services performed. Later the laboratory sends the results of the services performed to the recipient. A model of the domain, including objects, classes, attributes, generalization / specialization structures, whole / part structures, and instance connections is shown in Figure 1. Not shown in Figure 1 are the data types of the attributes or the message connections. Model representation follows that described by Coad and Yourdon[10].

Figure 2 shows a portion of one possible SGML DTD that is based on the model in Figure 1 and the other functional requirements specified for this example scenario in the CEN report. An attempt is made in DTD creation to map structures from an object oriented model onto SGML structures. In general, objects become SGML Elements, attributes become SGML Attributes, and generalization / specialization structures are reflected as SGML Element containment. Whole / part structures are modeled on the one-side as SGML Element containment where the one contains many of the other object (SGML Element), and on the many-

```
<!-- ENTITY Definitions -->
<!ENTITY % priority "NL | Emer | Urgent">

<!-- SERVICE Definitions -->
<!ELEMENT SVC-NewRequest - - (Request,
    Requester, Laboratory, Patient,Recipient)>

<!-- GEN / SPEC STRUCTURES -->
<!ELEMENT RequestedSpecimen - -
    (RequestElement,Specimen,DateTime)>
<!ATTLIST RequestedSpecimen
    Instructions CDATA #IMPLIED>
<!ELEMENT RequestedTest - -
    (RequestElement,CodedValue)>
<!ATTLIST RequestedTest
    Specimen.IdNo NUMBERS #IMPLIED
    Interpretation (YES | NO) #REQUIRED>
<!ELEMENT (Requester|Laboratory|Recipient)
    - - (CommunicationParty)>

<!-- OBJECTS and ATTRIBUTES -->
<!ELEMENT Patient - -
    (Name,DOB,Residence*)>
<!ATTLIST Patient
    IdNo NUMBER #REQUIRED
    Sex (M|F|O|U) #REQUIRED
    Nationality CDATA #IMPLIED>
<!ELEMENT Request - -
    (DateTime,RequestElement*,Specimen*)>
<!ATTLIST Request
    IdNo NUMBER #REQUIRED
    Patient.IdNo NUMBER #REQUIRED
    Requester.IdNo NUMBER #REQUIRED
    Lab.IdNo NUMBER #REQUIRED
    Recipient.IdNo NUMBER #REQUIRED>
<!ELEMENT Report - - (#PCDATA)>
<!ELEMENT RequestElement - o
    (RequestElement*)>
<!ATTLIST RequestElement
    IdNo NUMBER #REQUIRED
    RequestElement.IdNo NUMBER #IMPLIED
    Request.IdNo NUMBER #REQUIRED>

<!-- DATATYPES -->
<!ELEMENT CodedValue - o EMPTY
    -- HL7 Coded Element datatype -- >
<!ATTLIST CodedValue
    ID CDATA #REQUIRED
    Text CDATA #IMPLIED
    CodingSystem CDATA #REQUIRED>
```

FIGURE 2. Portion of SGML Document Type
Definition for object model of Figure 1.

side as an SGML Attribute. Instance connections are modeled as SGML Attributes. Services are modeled as SGML Elements which contain those objects (SGML Elements) pertaining to the message. While the creation of this DTD is not required to determine the ability of SGML to fulfill the CEN evaluation criteria, fragments of the DTD will help illustrate how those criteria can be satisfied.

RESULTS

Interchange formats are evaluated along six axes in the CEN report: (1) Support of Information Structures; (2) Support of Datatypes, including numbers, text strings, images, and sounds; (3) Encoding, i.e., the transformation to and from a transfer syntax; (4) Evolution and Backwards Compatibility; (5) Conformance and Certification; (6) Support and Availability. Each of these six axes are further broken down into Evaluation Criteria. Table 1 shows the results for SGML along with those previously reported for ASN.1, ASTM E1238, EDIFACT, EUCLIDES, and ODA by CEN. Also shown are illustrative fragments of the DTD of Figure 2.

Some points to make regarding SGML include: (1) SGML defines few simple data types, but can support the construction of complex types via SGML Element declarations. For example, Figure 2 shows element RequestedTest containing element CodedValue, which is a complex data type. The element CodedValue is defined analogously to an HL7 Coded Element data type; (2) Bit-data is outside the scope of SGML, meaning that SGML parsers only parse and validate character data. Thus, SGML parsers will not, for instance, allow one to validate that a bit-mapped image is in conformance with the particular image format specifications. Through the use of SGML NOTATION declaration, one can provide a pointer to an external description of the format and to a local application which can process the bit-data; (3) SGML is not necessarily object-oriented, thus the initial mapping from an object model into SGML can be partially automated, but requires human input.

CONCLUSION

In conclusion, SGML compares favorably with the other syntaxes investigated by CEN in their

TABLE 1. Evaluation results. A shaded cell indicates support for a function. Data for columns labeled 1-5 are reproduced from CEN[4] with permission and stand for ASN.1, ASTM, EDIFACT, EUCLIDES, and ODA respectively.

#	Evaluation Criterion	1	2	3	4	5	SGML (including Sample DTD Fragment)
A. Supported Information Structures							
1.	Optional						`<!Attlist Patient ... Nationality CDATA #IMPLIED>`
2.	Choice						`<!Attlist Patient ... Sex (M | F | O | U) #REQUIRED>`
3.	Repetition (set of)						`<!Element Patient - - (...,Residence*,...)>`
4.	Unordered sequence						`<!Element ABC - - (X & Y & Z)>`
5.	Ordered sequence						`<!Element ABC - - (X,Y,Z)>`
6.	1-1 relationship						`<!Attlist Request ... Patient.IdNo NUMBER #REQUIRED>`
7.	1-n relationship						`<!Attlist RequestElement ... Request.IdNo NUMBER #REQUIRED>`
8.	m-n relationship						`<!Attlist RequestedTest ... Specimen.IdNo NUMBERS #IMPLIED>`
9.	Recursive structures						`<!Element Section - - (para | Section)+>`
10.	Single inheritance						`* <!Element RequestedTest - - (RequestElement,...)>`
11.	Multiple inheritance						`* <!Element RequestedSpecimen - - (RequestElement,Specimen,...)>`
B. Supported Data Types							
1.	Encapsulate other IFs						
2.	ASCII text [ISO 646]						`<!Entity % ISOlat1 PUBLIC`
3.	Latin alphabet No. 1 [ISO 8859-1]						`"ISO 8879-1986//ENTITIES Added Latin 1//EN"> %ISOlat1;`
4.	Extended character set						
5.	Layout support	†					`<!Element (B | I) - - (#PCDATA)>`
6.	Raster graphics	†					‡
7.	Geometrical graphics	†					
8.	Continuous-tone still image						`<!NOTATION GIF PUBLIC "-//MyCorp//NOTATION GIF Format Design//EN" "Show_GIF.EXE">`
9.	Moving image	†					
10.	Wave form	†					`<!ENTITY Logo SYSTEM "Logo.GIF" NDATA GIF>`
11.	Sound	†					
12.	Coded datatype						`<!Attlist Patient ... Sex (M | F | O | U) #REQUIRED>`
13.	Binary datatype						
C. Encoding							
1.	Bit oriented transfer syntax						
2.	Variable data length						`<!Element Report - - (#PCDATA)>`
D. Evolution and Backwards Compatibility							
1.	Msg version handling						
2.	Adding new attributes to existing object types						`<!Attlist Patient ... Title CDATA #IMPLIED>`
3.	Adding new object types to a GMD						`<!Element Medication - - (...)>` `<!Attlist Medication ... Patient.IdNo NUMBER #REQUIRED>`

TABLE 1. (continued)

#	Evaluation Criterion	1	2	3	4	5	SGML
E. Conformance							
1.	General syntax conformance						
2.	Message syntax conformance						
F. Support and Availability							
1.	International standard						
2.	Used in health care			§			
3.	Used in other sectors						
4.	Available off the shelf software tools						
5.	Available skill on the Interchange Format						
6.	Registration mechanism for message types						

* Mapping an object-oriented model into SGML can be approximated.

† ASN.1 does not support these data types directly but syntaxes which define complex data types may be built using ASN.1 as a construction language.

‡ SGML does not support these data types directly but can provide a pointer to a description of the format and to a local application which can process the data.

§ The use as an interchange format is limited.

1993 report. We concur with CEN's finding that none of the interchange formats support all functional requirements. "To meet the requirements of messages which include non-textual information, a combination of the functionality of different interchange formats may be needed. This can be achieved through a modular approach." CEN recommended the subject of combining interchange formats for further study.

A further recommendation by CEN was for the bodies developing and maintaining interchange formats to take into account the value to users of using an object oriented approach to message development. This is the approach being taken in the HL7 Version 3.0 Message Development Framework[9]. Preliminary work mapping the requirements specified in this HL7 draft to SGML suggests that HL7 messages pose

constraints and requirements that were not identified in the example scenarios created by CEN. SGML as an interchange format for HL7 messages will be the subject of a future report.

References

1. HL7 SGML Special Interest Group web site. http://www.mcis.duke.edu/standards/HL7/committees/sgml/
2. Alschuler L. ABCD...SGML. Boston, MA: International Thomson Computer Press; 1995.
3. Text Encoding Initiative. A gentle introduction to SGML. http://www.sil.org/sgml/gentle.html
4. 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. (http://miginfo.rug.ac.be:8001/centc251/fox/word0222.htm)
5. Dilks C, Kutlesa J. SGML-based Electronic Drug Submissions: A Case Study. In: Proceedings GCA SGML'96, 1996:583-7.
6. Alschuler L, Lincoln TL, Spinosa J. Medicine for SGML. In: Proceedings GCA SGML'96, 1996:181-9.
7. Kahn CE, Huynh PN. Knowledge representation for platform-independent structured reporting. AMIA Fall Symposium 1996:478-82.
8. Rosenthal D, Sokolowski R. Using SGML for Voice-Enabled, Structured Medical Reporting. In: Proceedings GCA SGML'96, 1996:191-6.
9. HL7 Version 3.0 Message Development Framework (DRAFT). http://www.mcis.duke.edu/standards/HL7/pubs/version3/Version3.htm
10. Coad P, Yourdon E. Object-oriented Analysis. 2nd Edition. Englewood Cliffs, NJ: Prentice Hall; 1991.