Model Formulation

XML Schema Representation of DICOM Structured Reporting

K. P. Lee, PhD, Jingkun Hu, MS

Abstract Objective: The Digital Imaging and Communications in Medicine (DICOM) Structured Reporting (SR) standard improves the expressiveness, precision, and comparability of documentation about diagnostic images and waveforms. It supports the interchange of clinical reports in which critical features shown by images and waveforms can be denoted unambiguously by the observer, indexed, and retrieved selectively by subsequent reviewers. It is essential to provide access to clinical reports across the health care enterprise by using technologies that facilitate information exchange and processing by computers as well as provide support for robust and semantically rich standards, such as DICOM. This is supported by the current trend in the health care industry towards the use of Extensible Markup Language (XML) technologies for storage and exchange of medical information. The objective of the work reported here is to develop XML Schema for representing DICOM SR as XML documents.

Design: We briefly describe the document type definition (DTD) for XML and its limitations, followed by XML Schema (the intended replacement for DTD) and its features. A framework for generating XML Schema for representing DICOM SR in XML is presented next.

Measurements: None applicable.

Results: A schema instance based on an SR example in the DICOM specification was created and validated against the schema. The schema is being used extensively in producing reports on Philips Medical Systems ultrasound equipment.

Conclusion: With the framework described it is feasible to generate XML Schema using the existing DICOM SR specification. It can also be applied to generate XML Schemas for other DICOM information objects.

normal operation of a healthcare enterprise. When judiciously combined with knowledge-based information processing, we can significantly enhance clinical decision-making. As with any structured data in healthcare, benefits exist in outcome analysis and point-of-care applications.

Integrating the Healthcare Enterprise (IHE)\(^8\) is an ongoing initiative aiming at promoting and supporting the integration of healthcare systems to improve the efficiency and effectiveness of clinical practice. IHE defines a number of Integration Profiles, which are real-world situations that can be effectively managed by careful and consistent implementation and assembly of functionality provided by standards such as DICOM and others. Most profiles involve the exchange of information from various sources, and interoperability is key to successful achievement.

The DICOM specification, including SR, is maintained in Microsoft Word format and published as a PDF file. This format is unfortunately not amenable to machine processing and has the potential for misinterpretation. We believe that it is essential to provide access to clinical reports across the health care enterprise by using technologies that facilitate information exchange and processing by computers as well as provide support for robust and semantically rich standards, such as DICOM. For example, the HL7\(^7\) standard is developing its Clinical Document Architecture (CDA) for exchanging and processing electronic healthcare documents and uses Extensible Markup Language (XML)\(^6\) technologies. XML is a set of technologies that define a universal data format for tree-based, hierarchically formed information. A number of specifications extending its range and power, such as Extensible Stylesheet Language (XSL), Document Object Model (DOM), and XSL Transformations (XSLT), have already been developed. XML offers the advantages of platform independence and web awareness, and many XML tools are readily available. Thus XML technologies can provide a solution for enterprise-wide access to clinical information including medical reports. Having SR documents encoded in XML allows them to be validated for conformance to the standard, exchanged easily across the healthcare enterprise, and processed by computers.

To facilitate a uniform understanding of an XML encoding of medical reports, it is necessary to define a document type definition (DTD). Such a DTD has been derived from a Unified Modeling Language (UML) model of the DICOM SR information model.\(^13\) However, a DTD cannot fully represent DICOM data types and value representations because both UML and DTD have inherent limitations. XML Schema\(^14\) from the World Wide Web Consortium (W3C) provides more expressive power and allows rich structure and data type definition (among others) in XML documents. We propose a framework to generate XML schemas for SR automatically from the DICOM SR specification to ease valid encoding of SR documents in XML.

The reader is assumed to have a working knowledge of the DICOM standard and XML technologies. D. Clunie’s book\(^1\) is an excellent source of information on SR.

**Limitations of XML DTDs**

A DTD is used to describe the permissible elements and attributes in an XML document, primarily in terms of structures and restrictions of “document-like” objects such as articles and books. DTDs work well for this purpose. However, XML is being applied more and more to nontraditional areas such as databases, interprocess communications, e-business, and medical information management.\(^11\) These applications have much richer data type requirements that cannot be met by DTDs. XML Schema is intended to address some of the shortcomings of DTDs. For a short comparison of schemas and DTDs, see Mertz.\(^10\)

DICOM data types are very rich in constraints. For instance, DICOM defines approximately 20 types of strings in terms of the maximum lengths they can have, but in contrast a DTD cannot even constrain the length of a string. DICOM supports numerous data types such as string, short, and float, whereas DTD supports only the string type. In addition, DICOM defines attribute* types with specific optionalities: Type 1, Type 1C, Type 2, Type 2C, and Type 3. Type 1 is mandatory and must have a value. Type 1C is treated as Type 1 when certain conditions are met, but is optional otherwise. Type 2 is mandatory but can be empty. Type 2C is treated as Type 2 under certain conditions. And Type 3 is optional and can be empty. DTD cannot specify such granularities of conditionality; only two possibilities exist: required or optional. Moreover, it cannot constrain an element or an attribute to always have a non-empty value.

*It is unfortunate that the same term “attribute” is used to mean different things in XML and in DICOM. However, both usage patterns are well established. The context will indicate which one is meant.
To illustrate such limitations, consider the DICOM attribute Patient’s Name (0010,0010). It is a Type 2 attribute with a value that is a string with maximum length of 64 characters. This attribute can be represented in an XML DTD as follows:

```xml
<!ELEMENT patients_name EMPTY>
<!ATTLIST patients_name
  CodingScheme CDATA #FIXED "DCM"
  CodeId CDATA #FIXED "(0010,0010)"
  ValueType CDATA #FIXED "string"
  Value CDATA #REQUIRED
  codeMeaning CDATA #FIXED "Patient’s Name" >
```

However, it cannot specify that the length of the value of the Value attribute of patients_name cannot be more than 64 characters.

Another example is Patient’s Age (0010,1010). In DICOM definition, the value of this attribute must consist of three digits followed by one of the letters ‘Y’, ‘M’, ‘W’ or ‘D.’ We can specify an XML element patients_age as follows in a DTD:

```xml
<!ELEMENT patients_age EMPTY>
<!ATTLIST patients_age
  CodingScheme CDATA #FIXED "DCM"
  CodeId CDATA #FIXED "(0010,1010)"
  ValueType CDATA #FIXED "string"
  Value CDATA #REQUIRED
  CodeMeaning CDATA #FIXED "Patient’s Age" >
```

But there is no standard way to define a pattern of three digits followed by ‘Y’, ‘M’, ‘W’ or ‘D’ in a DTD.

DTDs have other shortcomings that make them inconvenient to process programmatically. One is that they do not support namespaces, which are becoming more and more important as the use of XML increases. Another is that a DTD is written with syntax different from that of an XML document, which means that DTDs cannot be processed easily by the variety of XML tools available.

XML Schemas have more capabilities than DTDs for specifying the validity of XML documents with rich structure and data-type definitions.

### Features of XML Schema

XML Schema is an XML definition language for describing and constraining the content of XML documents. After two years of intensive work, XML Schema is now a W3C Recommendation. It extends the capabilities of DTD by providing more detailed descriptions of data types and constraints on valid entry values such as ranges. In addition, XML Schema is more object-oriented with support for the notion of inheritance and allows abstract types, super types, and subtypes.

Other features of XML Schema enhance its usability. It supports namespaces, which enable definitions and declarations from different vocabularies to be freely intermixed in an XML document. A schema can be divided into several schema documents when it becomes larger. This increases its readability and reusability and enables better access control and maintenance. XML Schemas are themselves valid XML documents, which makes it possible to use the same tools to validate both schemas and their instances.

The following sample XML Schema type definitions address the issues raised above. The first is a type definition that restricts the value of a Patient’s Name DICOM attribute to be a string of no more than 64 characters (Figure 1).

The next example specifies that the value of Patient’s Age must conform to a particular pattern (Figure 2).

More details on how these types are obtained can be found below.

### A Framework for Generating XML Schemas for DICOM SR

DICOM SR defines three information object definitions (IODs): Basic, Enhanced, and Comprehensive. Each IOD contains several normalized or composite object classes called information entities (IEs). An IE consists of a number of Modules, which contain a set of atomic attributes, sequence attributes, or Macro attributes; and sequence attributes or Macro attributes themselves may also contain atomic attributes, sequence attributes, or Macro attributes, and so on recursively. Each IE, Module, or attribute is defined by tables in the DICOM specification and may have different usage, type, or value representation. Our approach to generating XML Schema for SR is to represent each table in XML using a straightforward encoding process and then transform the resulting XML files to schemas using XSLT technology. DICOM attribute names are converted to schema element names using a fixed set of conventions. For

---

*†DICOM attributes are uniquely identified by a tag, which is a pair of numbers.*
example, the DICOM name Patient’s Name is converted to the schema name patients_name. The schema generation process has two passes:

1. The XML files are processed once to produce one file containing data type definitions for all the attributes needed. This file is then sorted and pruned for duplicates, resulting in a single definition for each DICOM attribute. An advantage is that these definitions are now sorted in alphabetical order, making it easier to locate any specific data type definition.

2. A second pass is performed over the XML files to generate the final schemas. These schemas reuse the data type definitions obtained in the first pass.

The schema generation process is illustrated with an example below, and the generated top-level schema for the SR IOD is given in Appendix A. The set of generated schemas collectively represent a schema for SR. There are obviously interdependencies among these schemas. Schemas that are used by others will have to be included in the target schemas (as illustrated in Appendix A). In addition, SR has constraints that cannot be easily expressed in schema notation. These will have to be enforced by other means, e.g., other constraining language, XSLT stylesheets or external functions written in a conventional programming language. Some preliminary results are reported by Lee.9

```
<xsd:element name="patients_name">
  <xsd:complexType content="empty">
    <xsd:attribute name="CodingScheme" use="fixed" value="DCM"/>
    <xsd:attribute name="CodeId" use="fixed" value="(0010,0010)"/>
    <xsd:attribute name="codeMeaning" use="fixed" value="Patient’s Name"/>
    <xsd:attribute name="Value" use="required">
      <xsd:simpleType>
        <xsd:restriction base="xsd:string">
          <xsd:maxLength value="64"/>
        </xsd:restriction>
      </xsd:simpleType>
    </xsd:attribute>
  </xsd:complexType>
</xsd:element>
```

**Figure 1.** Schema for patient’s name.

```
<xsd:element name="patients_age">
  <xsd:complexType content="empty">
    <xsd:attribute name="CodingScheme" use="fixed" value="DCM"/>
    <xsd:attribute name="CodeId" use="fixed" value="(0010,1010)"/>
    <xsd:attribute name="codeMeaning" use="fixed" value="Patient’s Age"/>
    <xsd:attribute name="Value" use="required">
      <xsd:simpleType>
        <xsd:restriction base="xsd:string">
          <xsd:pattern value="[0-9]{3}[DWMY]"/>
        </xsd:restriction>
      </xsd:simpleType>
    </xsd:attribute>
  </xsd:complexType>
</xsd:element>
```

**Figure 2.** Schema for patient’s age.
Root Element and Namespaces

The root element of an XML schema is an element called schema in the XML Schema namespace (Figure 3). Its attributes define the target namespace and other namespaces used in the schema (and additional qualifications if necessary).

In this example, the target namespace is http://www.philips.com/pms as given by the value of the targetNamespace attribute and denoted as pms. http://www.w3.org/2001/XMLSchema is the namespace for the XML Schema Recommendation and abbreviated to xsd. Other namespaces used are http://www.w3.org/1999/XSL/Transform, the namespace for XSLT and abbreviated to xsl, and http://www.w3.org/2001/XMLSchema-Instance, the namespace for schema instances and abbreviated to xsi.

A schema may include one or more external schemas. The above example shows the inclusion of a schema called patient_module.xsd.

Generation Example

We illustrate the schema generation process by showing how a DICOM macro is transformed. Table 1 is a sample macro—Numeric Measurement Macro (C.18.1–1 from reference 3). This is a typical macro that contains an atomic attribute, two sequence attributes, and a nested macro.

This macro is represented in XML in a straightforward way as shown in Appendix B. The resulting schema for a complexType called numeric_measurement_macro is listed in Appendix C and a portion is reproduced in Figure 4.

The following points should be noted:

1. A DICOM Sequence (such as the Measured Value Sequence) is represented as a sequence of XML elements (measured_value_sequence_item) in order to accommodate DICOM attributes with multiple values. In this case the measured_value_sequence_item type is defined in Appendix D.

2. The Measured Value Sequence is of Type 2. This is reflected in the attribute nillable="true" in the schema element measured_value_sequence. Such an element is required to be present in any valid XML instance document, but can have an empty value. As is noted in the specification: “The Measured Value Sequence (0040,A300) may be

Table 1

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Tag</th>
<th>Type</th>
<th>Attribute Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measured value sequence</td>
<td>(0040,A300)</td>
<td>2</td>
<td>This is the value of the Content Item. Shall consist of a Sequence of Items conveying the measured value(s), which represent integers or real numbers and units of measurement. Zero or one Items shall be permitted in this sequence.</td>
</tr>
<tr>
<td>&gt;Numeric value</td>
<td>(0040,A30A)</td>
<td>1</td>
<td>Numeric measurement value. Only a single value shall be present.</td>
</tr>
<tr>
<td>&gt;Measurement units code sequence</td>
<td>(0040,08EA)</td>
<td>1</td>
<td>Units of measurement. Only a single Item shall be permitted in this sequence.</td>
</tr>
</tbody>
</table>

Defined Context ID is 82.
empty to convey the concept of a measurement whose value is unknown or missing.” Such a case will appear in an instance XML document as `<measured_value_sequence xsi:nil="true"/>

3. The table specifies that “zero or one Items shall be permitted” in a Measured Value Sequence. In the generated schema this is represented with the minOccurs attribute set to 0 and the maxOccurs attribute set to 1 for the element measured_value_sequence_item.

4. We have added four schema attributes to all generated elements that identify uniquely the original DICOM attributes. They are VR, codingScheme, codeId and codeMeaning, corresponding to the

<!DOCTYPE--Definition of numeric_measurement_macro-->
<xsd:complexType name="numeric_measurement_macro">
  <xsd:sequence>
    <xsd:element name="measured_value_sequence" nillable="true">
      <xsd:complexType>
        <xsd:sequence>
          <xsd:element name="measured_value_sequence_item" type="pms:measured_value_sequence_item" minOccurs="0" maxOccurs="1"/>
        </xsd:sequence>
        <xsd:attribute name="VR" fixed="SQ"/>
        <xsd:attribute name="codingScheme" fixed="DCM"/>
        <xsd:attribute name="codeId" fixed="(0040,A300)"/>
        <xsd:attribute name="codeMeaning" fixed="Measured Value Sequence"/>
      </xsd:complexType>
    </xsd:element>
  </xsd:sequence>
</xsd:complexType>

Figure 4. numeric_measurement_macro complexType.

<xsd:complexType name="measurement_units_code_sequence">
  <xsd:sequence>
    <xsd:element name="measurement_units_code_sequence_item" type="pms:measurement_units_code_sequence_item" minOccurs="1" maxOccurs="1"/>
  </xsd:sequence>
  <xsd:attribute name="VR" use="fixed" value="SQ"/>
  <xsd:attribute name="codingScheme" use="fixed" value="DCM"/>
  <xsd:attribute name="codeId" use="fixed" value="(0040,08EA)"/>
  <xsd:attribute name="codeMeaning" use="fixed" value="Measurement Units Code Sequence"/>
</xsd:complexType>

<xsd:complexType name="measurement_units_code_sequence_item">
  <xsd:sequence>
    <xsd:element name="code_sequence_macro" type="pms:code_sequence_macro"/>
  </xsd:sequence>
</xsd:complexType>

Figure 5. measurement_units_code_sequence complexType.
Value Representation of a DICOM attribute, the coding scheme used, and the unique code ID as well as a human readable code meaning. All are obtained from Part 6—Data Dictionary. These attributes are strictly speaking not necessary but we have decided to include them to provide an extra level of validity checking in an XML instance document.

This example also shows how a macro included in a table is processed. The Numeric Measurement Macro has a Measurement Units Code Sequence which includes a Code Sequence Macro (Table 8.8-1 of reference 3). The type definition for measurement_units_code_sequence is given in Figure 5. The included macro becomes an element code_sequence_macro of type code_sequence_macro, which is defined separately and simply included (see the next subsection).

Including External Schemas

The definition of an IOD, IE, Module, or Macro frequently refers to other Modules or Macros. These included Modules and Macros are themselves defined by their own tables and therefore need not be defined further in the including table. All that needs to be done is to include the corresponding schemas. The schema for the SR IOD given in Appendix A is such an example. Types for all the included modules are defined elsewhere and simply included here.

To generate the appropriate include instructions in the schema, the transforming stylesheet must look through the XML file for an IOD, IE, Module or Macro and determine which schemas need to be included. A schema should be included only once even if there are multiple references in a table. If there is a recursive reference, then it has to be detected to avoid a schema trying to include itself. The template to process inclusion is called do_include and reproduced in Figure 6.

This is extracted from the stylesheet to transform modules and macros. There is a corresponding template for the stylesheet for transforming IODs.

SR XML Schemas and Validation

We have developed XSLT stylesheets to implement the framework described. The XSLT processor together with stylesheets generates an XML schema for SR automatically. The only manual intervention is choosing a VR when a DICOM attribute has a choice of VR (e.g. Pixel Padding Value (0028,0120) has a VR of US or SS).

The informative SR Content Tree example in Annex K of reference 3 has been encoded in XML and validated against the generated schemas. The schemas are currently used extensively in creating reports produced by Philips Medical System ultrasound equipment.

Conclusions and Future Work

We have proposed a representation of the DICOM SR IOD in XML and a framework for automatically generating XML schemas from the representation. It addresses processing rules for IODs, IEs, Modules, Macros, atomic attributes, sequence attributes, and
their usages/types. These schemas will benefit XML implementation of DICOM SR for any medical equipment supplier. The framework can also be used for other DICOM IODs such as Waveforms. However, there are aspects of this framework that need further refinement to enhance the usefulness of the generated schemas. Among these:

1. Representation of DICOM Value Representation currently takes into account only the length attribute and some patterns. Other constraints need to be incorporated into the schema. For example, certain control characters must be excluded from some string types.

2. Many DICOM constraints are difficult or impossible to specify with a schema. There are additional mechanisms to increase the expressive powers of constraint specification, but because of the nature of some DICOM constraints, a programming approach seems to be the only way to address all the constraints. Some preliminary results have been reported by Lee.

We are also extending this framework to allow automatic generation of XML schema for SR templates as described in reference 5.

References


Appendix A. Generated Top-Level Schema for SR

```xml
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema targetNamespace="http://www.philips.com/pms"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema"
    xmlns:pms="http://www.philips.com/pms">
    <xsd:annotation>
    <xsd:documentation>SR IOD (Section A.35.1.1)</xsd:documentation>
    </xsd:annotation>
    <!--Start of external schemas-->
    <xsd:include schemaLocation="dataTypes.xsd"/>
    <xsd:include schemaLocation="typeDefs.xsd"/>
    <xsd:include schemaLocation="general_equipment_module.xsd"/>
    <xsd:include schemaLocation="general_study_module.xsd"/>
    <xsd:include schemaLocation="patient_module.xsd"/>
    <xsd:include schemaLocation="patient_study_module.xsd"/>
    <xsd:include schemaLocation="sop_common_module.xsd"/>
    <xsd:include schemaLocation="specimen_identification_module.xsd"/>
    <xsd:include schemaLocation="sr_document_content_module.xsd"/>
    <xsd:include schemaLocation="sr_document_general_module.xsd"/>
    <xsd:include schemaLocation="sr_document_series_module.xsd"/>
    <!--End of external schemas-->
    <xsd:element name="sr_iod">
        <xsd:complexType>
            <xsd:sequence>
```
Appendix B. XML Representation of Numeric Measurement Macro

<?xml version="1.0" encoding="UTF-8"?>
<macro_attributes>
  <reference>C.18.1</reference>
  <attribute_name>NUMERIC MEASUREMENT Macro</attribute_name>
  <name>numeric_measurement_macro</name>
  <sequence_attribute minOccurs="0" maxOccurs="1"/>
Appendix C. Generated Schema for Numeric Measurement Macro

<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema targetNamespace="http://www.philips.com/pms"
 xmlns:xsd="http://www.w3.org/2001/XMLSchema"
 xmlns:pms="http://www.philips.com/pms">
  <xsd:annotation>
    <xsd:documentation>NUMERIC MEASUREMENT Macro (Section C.18.1)</xsd:documentation>
  </xsd:annotation>
  <!--Start of external schemas-->
  <xsd:include schemaLocation="code_sequence_macro.xsd"/>
  <!--End of external schemas-->
  <!--Definition of numeric_measurement_macro-->
  <xsd:complexType name="numeric_measurement_macro">
    <xsd:sequence>
      <xsd:element name="measured_value_sequence" nillable="true">
        <xsd:complexType>
          <!--Start of external schemas-->
          <xsd:include schemaLocation="code_sequence_macro.xsd"/>
          <!--End of external schemas-->
          <xsd:sequence_attribute minOccurs="1" maxOccurs="1">
            <xsd:attribute name="Numeric Value" type="xsd:string" use="optional"/>
            <xsd:include schemaLocation="code_sequence_macro.xsd"/>
          </xsd:sequence_attribute>
        </xsd:complexType>
      </xsd:element>
    </xsd:sequence>
  </xsd:complexType>
</xsd:schema>
Appendix D. Type for Measured Value Sequence Elements

```xml
<xsd:complexType name="measured_value_sequence_item">
  <xsd:sequence>
    <xsd:element name="numeric_value" type="pms:numeric_value"/>
    <xsd:element name="measurement_units_code_sequence">
      <xsd:complexType>
        <xsd:sequence>
          <xsd:element name="measurement_units_code_sequence_item" type="pms:measurement_units_code_sequence_item"/>
        </xsd:sequence>
      </xsd:complexType>
    </xsd:element>
  </xsd:sequence>
</xsd:complexType>
```