# Safe Software Standards and XML Schemas

**Robert C. Leif*[a]**

**[a]XML_Med, a Division of Newport Instruments,
5648 Toyon Road, San Diego, CA 92115
rleif@rleif.com; www.newportinstruments.com**

## ABSTRACT

The goal of this work is to develop a safe software construction means for an XML based data standard for a class of medical devices, cytometry instruments. Unfortunately, the amount of empirical evidence to archive this goal is minimal. Therefore, technologies associated with high reliability were employed together with reuse of existing designs.

The basis for a major part of the design was the Digital Imaging and Communications in Medicine (DICOM) standard and the Flow Cytometry Standard (FCS). Since the DICOM Standard is a Class II device, the safety of software should be maximized. The XML Schema Definition Language (XSDL) has been used to develop schemas that maximize readability, modularity, strong typing, and reuse. An instance and an instrument XML schema were created for data obtained with a microscope by importing multiple schemas that each consisted of a class that described one object. This design was checked by validating the schemas and creating XML pages from them.

**Keywords:** DICOM, schema, XSD, CytometryML, FCS, software engineering, reuse, health care cost

## 1. INTRODUCTION

Presently, there is significant overlap in the coverage of cytometry and/or cytology in the existing standards of two groups, the pathologists and the cytometrists. The activities of these two groups have significant overlap including the transfer of technology from the cytometrists to the pathologists. These two groups have different data file standards. The pathologists are in the process of adopting the Digital Imaging and Communications in Medicine (DICOM)[1] standard (http://medical.nema.org/) and the Cytometrists use the International Society for Advancement of Cytometry, ISAC (http://www.isac-net.org/) Data File Standard for Flow Cytometry, Version FCS3.0 (http://www.isac-net.org/ index.php?option=com_content&task=view&id=101&Itemid=150), which is commonly referred to as Flow Cytometry Standard, FCS. The syntaxes used for both FCS and DICOM are unique and require software interfaces to work with other applications. Both groups have started to create software in XML. The ISAC data standards task force has created Gating-ML: XML-Based Gating Descriptions in Flow Cytometry[2]. The DICOM Working Group 27 is creating schemas that are an extension to the capabilities of DICOM Part 18: Web Access to DICOM Persistent Objects (WADO) (ftp:// medical.nema.org/medical/dicom/2008/08_18pu.pdf)[3]. These schemas will permit transmission of some data via the Web Services Description Language, WSDL (http://www.w3.org/TR/wsdl20/), between DICOM data stores and XML. The XML metadata returned as part of the QueryResult is generalized in the form of a string and thus is weakly typed. However, this approach is acceptable because it has been designated for transfer only between computers without human intervention.

Since it would be useful to leverage this work on WSDL to produce XML that could be used for human comprehension, the DICOM hierarchy of patient, study, series, and instance[1] has been followed. two schemas instance.xsd and instrument.microscope.xsd that include strongly typed data have been created and tested. Significant development of a third schema, series.xsd, has already occurred[4]. Only work concerning the series, and instance will be discussed. A series contains metadata that indicates the locations of one or more measurements, images and/or list-mode files, produced by a single instrument on a specimen or specimens that came from a common ancestor. If the specimen is a solid, usually either one or more tissue sections or a dispersion of singles cells and small aggregates, it is located on a microscope slide, which serves as a container; or if it is a flow cytometry measurement, a cellular suspension is in a liquid container, which is usually a vial, evacuated tube, or microtiter plate. Usually, both the cellular suspensions and the solid material on a slide need to be prepared by somewhat complex procedures. The description of the protocol that is common to the preparation and analysis of all of the specimens prior to the subsequent treatment required for the individual measurements is part of the series. The description of the protocol that is specific for the creation and characterization of the individual or closely related measurements, such as images and list-mode files, is part of the instance.

Since there is only one instrument and it can have many channels, it is redundant to repeat the information contained in either instrument.flow.xsd or instrument.microscope.xsd with the data that describes each channel. The content of the protocol element of an instance is specific to that instance. The instance[4] container file[5] includes the binary data (list-mode, images, and index[6] files), which are referenced by the protocol element contained within an instance element. These individual protocol elements also reference the list-mode and image context files that contain or point to the metadata similar to that, which previously had been included in FCS and DICOM, such as FCS, JPEG, or TIFF files. The individual instance protocol elements also reference the individual channels settings for each parameter. Settings such as staining protocol and optical configuration can change for each instance XML document. The series container includes the description of entities that are the same for all instances that are part of the series. The instrument XML page that includes the elements that describe the fixed parts of the instrument (Flow Cytometer or Microscope) is part of the series or is pointed to by a URL, because these fixed elements have constant values or settings for all acquisitions of the data contained in all instances that are part of the series. These fixed parts include detailed descriptions of the manufacturer, serial number, and similar data for both the microscope and flow cytometer elements. Since the XML page that describes the instrument includes all of the major stable components, this XML page could be maintained by the manufacturer and/or vendor and located on their web site. The following elements are also only included in the series protocol element: the total number of instances and a brief summary of the information in each instance container file. It is anticipated that the user will often look first at the series file in order to select instances.

## 2. METHODS

### 2.1 Disclaimer

If a claim of adequate safety is to be truly valid, experimental data to substantiate the appropriateness of the development and testing techniques needs to be provided. The unfortunate facts are the quality of software development techniques has not been adequately measured and most programs of useful size cannot be completely tested. Hopefully, these upsetting comments will provide some small encouragement to improve this situation.

Although this author believes the nomenclature for type and element names employed in the XML pages below and the schemas from which they were derived are the safest choices, evidence to prove this level of safety does not appear to exist. Since most programming languages use strings that lack spaces as the way to name data-types and objects, there is a necessity to indicate the spaces that would be present had the name been allowed to include spaces. The two major ways to create these space-free strings are to replace spaces by the use of internal capitalization, camelCase, or the use of underscore characters'_'. Data at the level of an open label clinical trial on the reading comprehension of text formatted to adhere to of each of these solutions could not be found with a Google search. Other similar software engineering questions are:

1. Are standard rules and uses of capitalization helpful in the comprehension of software?

2. Are differences in case of one or more characters sufficient to describe different data-types?

3. What is the capacity of humans to remember abbreviations including two letter abbreviations?

4, Should suffixes, such as _Type, be used to designate specialized classes of strings, such as data-type names.

5. What is the error rate for each programming language for some unit of code?

6. What is the appropriate unit of code for question 5 above and questions 7, 8, and 9 below?

7. Which programming language has the longest mean time between failures (exceptions, errors, crashes, etc.)?

8. Which programming language has the shortest mean time to repair from a failure?

9. Which programming language has the greatest efficiency in extending programs.

### 2.2 Rational

Much of the information and data-types present in the XML schemas and subsequently XML pages were prepared by domain experts and this information was reused from Digital Imaging and Communications in Medicine (DICOM)[1] standard (http://medical.nema.org/) or Flow Cytometry Standard, FCS. New data-types were created and data-types from other CytometryML schemas[7] were reused.

Because the Digital Imaging and Communications in Medicine (DICOM)[1] standard (http://medical.nema.org/) is a FDA Class II device[8], the safety of the software developed as part of a standard should be maximized. The programming language that is claimed to be "Suitable for use in mission critical and high-integrity software development" and purported to have the highest reliability is SPARK Ada (http://libre.adacore.com/libre/tools/spark-gpl-edition/). Readability, modularity, strong typing, and reuse are four software engineering principles that are used in SPARK. These have been applied to the CytometryML[4,7] XML schemas. This is possible with the use of the XML Schema Definition Language (XSDL) structures (http://www.w3.org/TR/xmlschema11-1/) and data-types (XSD) (http://www.w3.org/TR/xmlschema11-2/). In fact, XSDL version 1.1 also includes assertions, which are one of the key parts of SPARK. The requirement of readability should satisfy the requirement that a "Detailed semantic shall be provided to prevent potential misinterpretations and misusages of the standard", which included in a requirement[9] of the ISAC Advanced Cytometry Standard (ACS) for a data file standard[9].and Many of the other requirements are met for a data file standard or facilitated by the use of XSDL, and the structure of CytometryML.

The XSDL schemas were validated by XMLSpy (http://www.altova.com/) and oXygen (http://www.oxygenxml.com/). Many of the schemas have also been tested with XSDL 1.1 with the Saxon-EE 9.2.0.3 parser. A new XML page was subsequently produced from each of the main schemas and then filled with the values from the original XML page, and validated.

## 3. RESULTS

The Protocol is one of the two parts of main metadata element of the instance XML file. The other is the Instance_Header. A section of the content of each of the CytometryML instance.xsd and instrument.microscope.xsd schemas will be described in terms of the XML pages generated from these schemas. The description of the optical path, because it can change, occurs as part of the Protocol element of the instance XML document. Each Protocol element contains one Channel_Reference that contains the elements present in Table 1.

**Table 1 : Channel_Reference Elements (simplified)**

| Elements | Example of Values |
|---|---|
| Analyte Reporter | Anti5Brdu |
| Parameter | FL1-A |
| Channel Number | 3 |
| Measurement | Fluorescence |
| Long Name | AlexaFluor |
| Optical Path | Described below |
| Statistics | CV= 3.0% |
| Quality Assurance | Bead-based alignment setup |

The Example of Values column of Table 1 includes in most cases only one of the values of one of the parts of each element.

The optical path element, which could be used during the acquisition of a fluorescence image or list-mode data, is described in detail in Figure 1, which shows an episcopic illuminated system. In the case of other optical configurations, either an objective or a condenser can be used for illumination. In the case of a flow cytometer, a condenser can be used for imaging the light on a detector or an aperture in front of a detector, such as a PMT. The order of the optical path has been defined, such that, if an optical element is being used for excitation, its position will have a negative value; similarly optical components used for imaging have a positive value in the optical train. The slide or flow cell that holds the specimen is 0. The optics then go in a positive direction towards the detector and a negative direction towards the light source.

### 3.1 Optical_Path element example

Figure 1 shows the optical path of a fluorescence microscope, which is described in the Optical_Path element of the Channel_Reference element that is located within the instance Protocol element. Once, parsers for XSDL 1.1 become readily available, the data-type for the Optical_Path and many other sequences will be changed to an all. The all data-type is a realistic model of an optical path in that it permits an unfixed order of and multiple instances of the elements in the XML page.

.

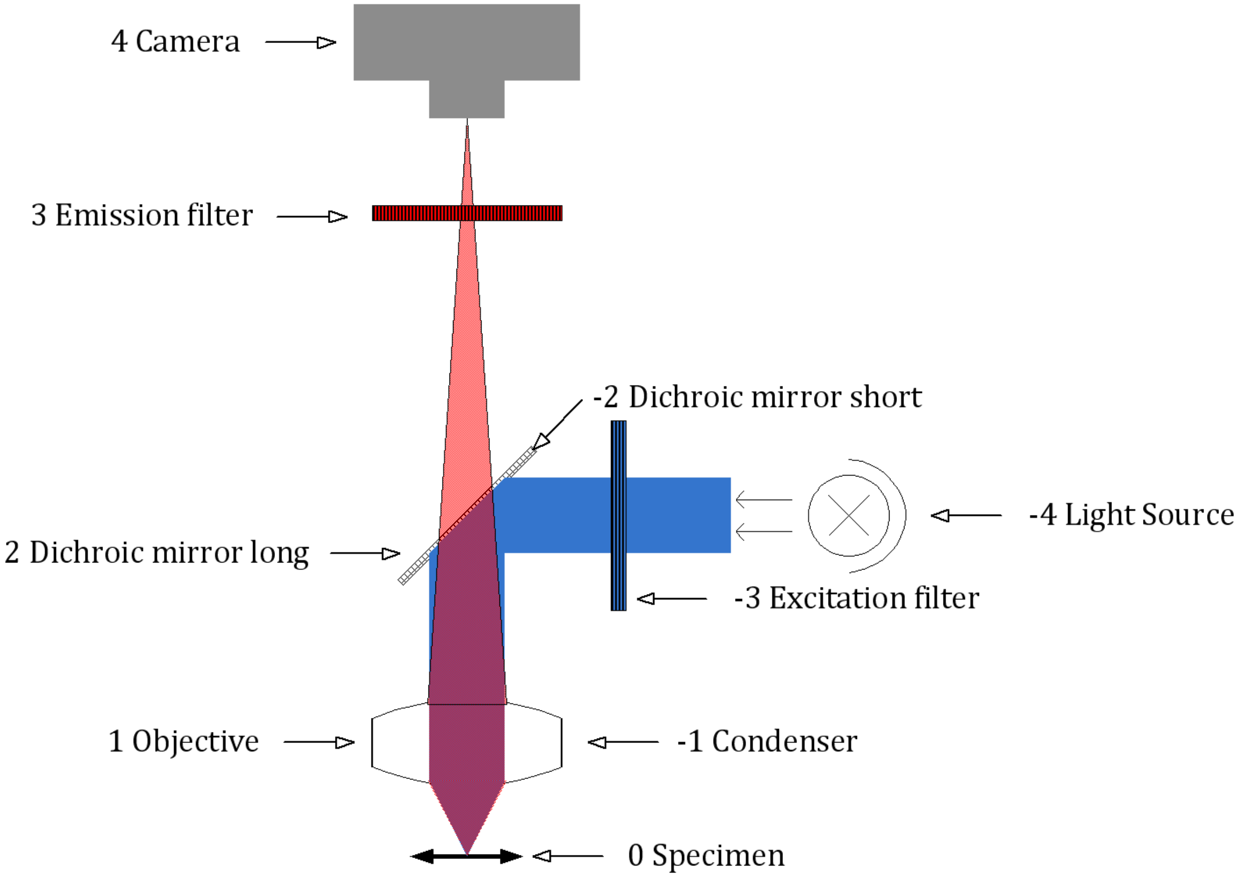

**Figure 1** is a cartoon of an epiiscopic fluorescence optical microscope. Each of the componants have been numbered in the order of their prescence in the excitation and emission paths.

The optical parts are numbered with the specimen being 0 and the part of the path that goes to the detector having positive values and the part that emanating from the light source having negative values. The order values of the different optical components are shown in Figure 1 and lines 3, 13, 15, 16, 19, 37, 39 of the Optical_Path Code Fragment. Because this is an epiiluminated system, where the excitation and emission beams are separated by a dichroic mirror, some of the components have two values. The objective (1) is also the condenser (-1). The dichroic mirror (lines 15 and 16) and the objective (line 19) are parts of both the excitation and emission paths.

**Code Fragment 1 Instance Optical_Path**

```
1<channels:Optical_Path>
2    <channels:Path>Episcopic_Illumination</channels:Path>
3    <channels:Light_Source_Info Order="-4"
```

```
    UID="1.111.11.112.11.2>
    <excite:Short_Name>LED365</excite:Short_Name>
    <excite:Light_Source>LED</excite:Light_Source>
    <excite:Wavelength Wavelength="365" Units="nm"/>
    <excite:Power Units="milliwatt" Power="200"/>
    <excite:Polarization>None</excite:Polarization>
    <excite:Description>Came with microscope
    </excite:Description>
</channels:Light_Source_Info>
<channels:Excitation_Filter Order="-3"
    Short_Name="Wide365" UID="001.001.0002.001.5"/>
<channels:Dichroic_Mirror Imaging_Order="2"
    Excitation_Order="-2" Short_Name="Pass365"
    UID="001.001.0003.001.5" Location="Parallel_Area"/>
<channels:Objective_Info Location="Object_Focal_Plane"
    Imaging_Order="1" Excitation_Order="-1">
    <optics:Magnification>40</optics:Magnification>
    <optics:NA>0.7</optics:NA>
    <optics:Contrast>None</optics:Contrast>
    <optics:Field_Flatness>Plan</optics:Field_Flatness>
    <optics:Immersion>Air</optics:Immersion>
    <optics:Chromat>Fluorite</optics:Chromat>
    <optics:Abbreviated_Info
        UID_Value="001.001.0003.001.5">
        <item:Identifier>ID_2</item:Identifier>
        <item:Manufacturer>Any microscope company
        </item:Manufacturer>
        <item:Model_Name>high dry</item:Model_Name>
        <item:Description>high dry that came with the
              microscope. </item:Description>
    </optics:Abbreviated_Info>
</channels:Objective_Info>
<channels:Detector_Emission_Filter
```

```
    Short_Name="Center530" Order="3"
    UID="001.001.0002.001.6"/>
<channels:Detector Order="4">
    <channels:Camera_Info>
        <cameras:Abbreviated_Info
            UID_Value="001.001.0004.001.7">
            <item:Identifier>ID_10</item:Identifier>
            <item:Manufacturer>Point Grey</item:Manufacturer>
            <item:Model_Name>Dragon2</item:Model_Name>
            <item:Description>Does analog time-gated
            illumination</item:Description>
        </cameras:Abbreviated_Info>
        <cameras:Columns>640</cameras:Columns>
        <cameras:Rows>480</cameras:Rows>
        <cameras:Technology>CCD</cameras:Technology>
        <cameras:Intensified>
            <cameras:Not_Intensified>
                true</cameras:Not_Intensified>
        </cameras:Intensified>
        <cameras:Binning>2</cameras:Binning>
        <cameras:Exposure_Duration
            Prefix="milli" Units="Seconds">
            1.0</cameras:Exposure_Duration>
        <cameras:Exposure_Off_Duration
            Prefix="milli" Units="Seconds">1.0
        </cameras:Exposure_Off_Duration>
        <cameras:Summation_Mtd>
            <cameras:Method>Analog</cameras:Method>
            <cameras:Num_Exposures_Summed>100
            </cameras:Num_Exposures_Summed>
        </cameras:Summation_Mtd>
        <cameras:Temperature_Centigrade>273
        </cameras:Temperature_Centigrade>
```

```
70      </channels:Camera_Info>
71    </channels:Detector>
72</channels:Optical_Path>
```

**Code Fragment 2 Series Microscope**

```
<instr:Light_Source>
    <excite:Light_Source>LED</excite:Light_Source>
    <excite:identifier>ID_4</excite:identifier>
    <excite:Emitter>GaAlAs</excite:Emitter>
    <excite:Wavelength Wavelength="365" Unit="nm"/>
    <excite:Max_Power Units="milliwatt" Power="250"/>
    <excite:Polarization>None</excite:Polarization>
    <excite:Object_Plane Units="µm" Width="0.50"
          Shape="Circular" Height="0.5"/>
    <excite:Description>Stock Nichia</excite:Description>
    <excite:General_Info UID_Value="1.111.11.112.11.2">
        <item:Identifier>ID_5</item:Identifier>
        <item:Manufacturer>Nichia</item:Manufacturer>
        <item:Model_Name>UV LED</item:Model_Name>
        <item:Model_Number>12345678</item:Model_Number>
        <item:Description>UV LED that can be pulsed at  kiloHz
        </item:Description>
        <item:Item_Serial-number>01234567
        </item:Item_Serial-number>
        <item:URI_Var>http://www.Nichia.com</item:URI_Var>
    </excite:General_Info>
</instr:Light_Source>
```

The content of the instance file substantially differs from that contained in the Instrument (Microscope) file, which is either located within the series container or pointed to by the series.XML document. The instance file describes the settings and configuration used to perform the measurement, which can and often does differ between measurements (instances). The Instrument file, which in this case describes a digital microscope, provides a detailed description of the instrument and its components. Since these details are unchanged between instances, the Instrument file or its URL or both are included in the series container.

Comparison of the Light_Source_Info element [lines 3 to 12] contained in the Code Fragment 1 Instance Optical_Path with the Light_Source element of the Microscope.XML page, Code Fragment 2 Series Microscope, from a series container shows: 1) Light_Source_Info contains only 8 items and that only four of these: UID, Light_Source, Wavelength, and polarization had the same name and values in the Light_Source element. If the light source provided multiple wavelengths and was polarized, this could have been only two elements with the same name. The UID provides the link between both elements. It appears that this type of link can be implemented by use of an XML Schema 1.1

assertion. Order numbers in the Light_Source_Info element refer to the actual configuration used for the measurement; whereas, order numbers used in the Light_Source element can refer to elements in a drawing etc.

The content of the Light_Source element of the instrument file is extensive. It includes: the emitting material, the maximum power output, the dimensions and shape of the image of the Light_Source at the object plane, and manufacturer related information including the manufacturer's URL.

It is anticipated that the instrument files for commercial medical devices will be under the control of the manufacturer. In the case of instrument developers and researchers, the instrument file in most cases will by necessity be under the researcher's control.

The instance Optical_Path and other elements in an instance.XML page serves two purposes. The first purpose is to permit the person who reads the XML page to understand how the measurement was made and what was measured. The second purposes to permit the measurement to be repeated. In order accomplish this, a record must be made, preferably at the time of the measurement, that describes the configuration of the microscope (instrument), which together with the detailed information of the instrument components (parts), provided by the microscope.XML page is sufficient to repeat the experiment.

## 4. DISCUSSION

This iteration of the code development for CytometryML has demonstrated the feasibility of applying the DICOM design specified organization of series and instances to cytometry data. It has also demonstrated that at least a significant part of DICOM series and instance metadata can be kept in the form of XML pages. This use of XML has the very significant advantages over the present DICOM standard of innate interoperability and being in a format that can be validated.

## 5. CONCLUSIONS

It has been possible with XSDL to maximize readability, create a modular structure, and strongly typed, reusable data-types. Maximizing reuse including reuse of designs and documentation, besides increasing safety and minimizing development costs, should significantly help to improve the US medical informatics infrastructure. This infrastructure improvement should benefit the patients while significantly decreasing health care costs.

## 6. ACKNOWLEDGMENT

This work was sponsored by Newport Instruments internal research funds.

## 7. REFERENCES

1] Pianykh O.S., [Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival Guide], Springer Publishers, Berlin & Heidelberg, (2008).

2] Spidlen J., Leif R.C., Moore W., Roederer M., International Society for the Advancement of Cytometry Data Standards Task Force, Brinkman R.R., "Gating-ML: XML-based gating descriptions in flow cytometry," Cytometry Part A 73A, 1151-1157 (2008).

3] DICOM Working Group 27, "DICOM Supplement 148: Web Access to DICOM persistent Objects by means of Web Services, Extension of the Retrieve Service (WADO Web Service), working draft in progress", (2009).

4] Leif R.C., "Toward the integration of cytomics and medicine," J. Biophoton. 2, 482-493 (2009).

5] Leif R.C., Spidlen J., Brinkman R.R., "A Container for the Advanced Cytometry Standard (ACS)," Proc. SPIE 7182, 71821Q (2009).

6] Leif R.C., "CytometryML, Binary Data Standards," Proc. SPIE 5699, 325-333 (2005).

7] Leif R.C., Leif S.B., and Leif S.H., "CytometryML, an XML Format Based on DICOM and FCS for Analytical Cytology Data," Cytometry 54A, 56-65 (2003).

8] Office of Device Evaluation, "Guidance for the Submission Of Premarket Notifications for Medical Image Management Devices", US FDA, Available at: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm073721.pdf (2000).

9] Spidlen J., Brinkman R., Leif R. C., and other members of the ISAC Data Standards Task Force., "Advanced Cytometry Standard (ACS) Requirements for a data file standard format to describe cytometry and related analytical cytology data, Version 0.070920," Available at: http://cdnetworks-us-1.dl.sourceforge.net/project/flowcyt/Analytical Cytology Standard/Analytical Cytology Standard/Requirements-v070920.pdf (2007).