CytometryML, a markup language for analytical cytology

Robert C. Leif, Stephanie H. Leif, Suzanne B. Leif

XML_Med, a Division of Newport Instruments, 5648 Toyon Road, San Diego, CA 92115-1022;

ABSTRACT

Cytometry Markup Language, CytometryML, is a proposed new analytical cytology data standard. CytometryML is a set of XML schemas for encoding both flow cytometry and digital microscopy text based data types. CytometryML schemas reference both DICOM (Digital Imaging and Communications in Medicine) codes and FCS keywords. These schemas provide representations for the keywords in FCS 3.0 and will soon include DICOM microscopic image data. Flow Cytometry Standard (FCS) list-mode has been mapped to the DICOM Waveform Information Object. A preliminary version of a list mode binary data type, which does not presently exist in DICOM, has been designed. This binary type is required to enhance the storage and transmission of flow cytometry and digital microscopy data. Index files based on Waveform indices will be used to rapidly locate the cells present in individual subsets. DICOM has the advantage of employing standard file types, TIF and JPEG, for Digital Microscopy.

Using an XML schema based representation means that standard commercial software packages such as Excel and MathCad can be used to analyze, display, and store analytical cytometry data. Furthermore, by providing one standard for both DICOM data and analytical cytology data, it eliminates the need to create and maintain special purpose interfaces for analytical cytology data thereby integrating the data into the larger DICOM and other clinical communities. A draft version of CytometryML is available at www.newportinstruments.com.

Keywords: CytometryML, Cytometry Markup, Cytometry Language, XML, DICOM, FCS

1. INTRODUCTION

In previous papers, we have discussed the deficiencies of Flow Cytometry Standard, FCS[1,2], suggested its replacement by DICOM[3,4,5], and suggested the use of XML schemas based on DICOM elements (data-types)[5]. In a companion paper[6], we have summarized and extended this material including a preliminary higher level description of Cytometry Markup Language, CytometryML, which is a collection of XML schemas[7,8] primarily based on the DICOM Waveform Information Object[9]. These schemas are primarily based on DICOM and FCS data-types. Backwards linkage to both DICOM and FCS was maintained by including the DICOM tags, DICOM value representations (VR), and FCS key words as fixed (constant) attributes. Since these fixed attributes are constants, their presence has been limited to the schemas and they need not be included in the XML documents. This should facilitate interoperation of DICOM and XML without cluttering the XML documents or forms with this information. The adherence of CytometryML to the previously published requirements has been demonstrated[6].

This paper provides a detailed description of CytometryML. Besides adhering to the previously published requirements[3], the development of CytometryML is based on the concept of standards' parsimony and the desire to easily interface with commercial off-the-shelf software. Standards' parsimony is the process of creating standards based on existing standards. It involves the minimization of 1) the creation of new data-types and the maximization of the reuse of existing data-types; and 2) the minimization of the creation of new syntactical constructs. We hope to demonstrate below that this parsimony has led to the successful implementation of CytometryML.

2. METHODS

The schemas[7,8] and XML documents[10] were developed with XMLSpy (www.xmlspy.com). The mapping of the DICOM data types to XML and their correspondence with FCS equivalents have been documented in these schemas. A preliminary mapping was performed with an Microsoft® Excel spreadsheet. These schemas, wherever reasonable, limit the

Robert C. Leif, XML_Med, a Division of Newport Instruments, 5648 Toyon Road, San Diego, CA 92115-1022; E-mail rleif@rleif.com, Tel. & FAX (619)582-0437; web HTTP://www.newportinstruments.com

acceptable ranges of data types. These limits increase the safety of the system by forcing the software to indicate an error whenever an element or attribute has a value that is out of the specified range[11]. Reusability and readability of the schemas were maximized by declaring only data types and one element based on a complex type for schemas that are referred to by XML documents. All of the XML documents were based on an automatic translation by XMLSpy of an individual schema into an XML document. These XML documents were subsequently edited to remove the constant attributes and other extraneous material.

Prior to the publication of this paper, the schemas used to produce this publication and any other supplemental materials were made available in color at www.newportinstruments.com. It should be cautioned, that these documents are only relevant to the status of CytometryML at the time of submission of the paper. The future versions will be posted at the Newport Instruments web site (www.newportinstruments.com) and hopefully, at some official web site.

The DICOM Waveform Information Object[9] is described by three main schemas: waveform, multiplex_groups, and parameters. These schemas were kept to a manageable size by employing object oriented design methodology to design multiple schemas, each associated with a single class or cohesive group of classes. Four of these helper schemas are dicom, fcs, num_types, and time. Since XML is case sensitive, all schema names were made lower case. The camel case convention of XML terms has been maintained. XML camel case has compound words starting in lower case with each subsequent word or abbreviation being delineated from its predecessor by having its first letter capitalized. Compound words that describe new types, elements, and attributes follow the common convention for case insensitive software. The first letter of every word is capitalized and words are separated by underscores. If a word is an abbreviation, such as FCS it is given as all capitals. This permits these new compound types to remain intact when they are formatted by the pretty-printers of case insensitive languages

3. Detailed Description of CytometryML

The DICOM schema includes: the Tag_Type, the VR_Type, and DICOM string types. The Registry of DICOM data elements (data-types) in the Data Dictionary[12] is a table with four fields: Name, Tag, Value Representation (VR), and Value Multiplicity (VM). The DICOM name maps to the XML element name. Each DICOM element (object) has both a specific type, which is described by a tag and a class, which is described by a value representation, VR. The Tag and VR are attributes each with a fixed (constant) value. The Value Multiplicity is handled by the XML minOccurs and maxOccurs attributes. The DICOM Tag for each data-type serves as a unique serial number. The code for the Tag_Type is shown below.

```
1<simpleType name="Tag_Type">
2  <restriction base="string">
3    <pattern value="[0-9a-fA-F]{4},[0-9a-fA-F]{4}"/>
4  </restriction>
5</simpleType>
```

All XML statements begin with the '<' character and end with the '>' character. The end of an XML statement includes the '/' character. Statement numbers have been added at the left. These will be abbreviated as Sn, where n is the number shown at the far left. XML is a nested language that, as shown in S1, employs the less than character '<' to begin a construct. simpleType is an element which has an attribute, name. The value of an attribute is set in quotation marks and follows the equals sign. In this case it is Tag_Type. When elements are composed of structures or have attributes, they are complexTypes. When they are only composed of an elementary type, such as a string or an integer, they are simpleTypes. All attributes are simpleTypes. Tag_Type is a simple type, which is based upon a string. Thus, an attribute can be based on it. S3 restricts the string to a pattern of four hexadecimal numbers followed by a comma and a second set of four hexadecimal numbers. S3 is terminated by the two character string, />. Statements (4) and (5) respectively end the restriction and the simpleType. The two character prefix, </, is used to show an ending statement. The VR_Type, value representation, is a simpleType that enumerates the DICOM classes.

The FCS schema only has one simpleType, FCS_Keyword_Type. The Num_Types schema includes: integers, unsigned

integers, and floating point types. The names of the types are based on the simple, concise ones provided by a public standard, ECMA[13]. These include: Int8, Int16, Int32, Int64, UInt8, UInt16, UInt32, UInt64, float32, float64, and decimal. ECMA data types are used in the Microsoft .NET architecture.

The time schema includes the DICOM time types[14] necessary for the description of cytometry data. These types include: the Acquisition Datetime, which is the starting time of the original analytical cytology measurement, and the Content Date and Content Time, which are respectively the date and time analytical cytology data was created. These latter two types can be applied to processed data. The Instance Creation Date and Instance Creation Time are respectively the date and time that the data was created for transmission or storage.

3.1 Waveform

The waveform schema imports data-types from the following schemas: DICOM, FCS, Num_Types and Time. Since the code in a schemas is much larger and harder to read than the code in an XML document, the latter will be presented for the waveform and parameters. Since the multiplex_groups is the smallest entity, the schema for this XML document is presented.

An example of a waveform XML document is shown below. It and the subsequent documents will be separated by sections of explanatory text.

```
1 <?xml version="1.0" encoding="UTF-8"?>
2<!--Sample XML file generated by XML Spy v5.2 U (http://www.xmlspy.com)and then edited
   by RCL. All rights to this and the subsequent code reserved.-->
3 <wave:Waveform
3a  xmlns:wave="file://C:\Stds\DICOM_XML\waveform.xsd"
3b  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
3c  xsi:schemaLocation="file://C:\Stds\DICOM_XML\waveform.xsd
          C:\Stds\DICOM_XML\waveform.xsd">
```

When parts of a statement need to be referred to separately, the first line will start with a statement number. Subsequent lines will have a lower case letter following the number. S1 tells the browser that this is an XML file and that the UTF-8 character set is used. UTF-8 is an 8 bit format Unicode character representation. Latin-1 encoding would be specified as ISO-8895-1. The next two lines S2 are a comment. Comments have a start with the string "<!--" and end in the string "-->".

Waveform (S3), which is a global element of the schema, requires the schema prefix, wave:, whose name space (xmlns) and location are given in S3a. Similarly 3b gives the name space and location of the parent XML schema. S3c specifies the schema location for the above XML document. This schema, which is too long to be shown here, specifies the XML document's types, elements and attributes.

```
4   <Modality>Flow</Modality>
5   <Waveform_Originality>Original</Waveform_Originality>
6   <Acquisition_Date_Time>2002-12-15T13:30:47-05:00</Acquisition_Date_Time>
```

S4 specifies the Modality (Flow, Sort, Slide_Image, or Plate_Image). The Originality of the data (Original or Processed) is specified in S5. Processed data is produced by some analytical process. Although FCS 3.0 employs an analysis section for this type of data, it does not provide an audit trail suitable for today's medical-legal environment. The third element (S6) is the start Date-Time of acquisition of the flow cytometry list mode data or image data from which list mode can be derived.

```
7   <Acquisition_Context>
8      <Acquisition_Context_Description>1 to 1024 Chars</Acquisition_Context_Description>
```

```
9       <!--Chars=Characters-->
10      <Triggers Trigger_Min_Value="100" Trigger_Max_Value="255">
11          <Trigger_Source>1 to 16 Chars</Trigger_Source>
12          <Trigger_Source_Long_Name>1 to 64 Chars</Trigger_Source_Long_Name>
13      </Triggers>
```

The Acquisition_Context, which includes multiple elements starts at S7 and ends at S23. It starts (S8) with an optional description of up to 1,024 characters. The Triggers start on S10 and end at S13. The range of values, 100 to 255, that trigger the instrument for the parameter named in S11 are given as XML attributes. Attributes are included inside the element and use the '=' character for assignment.

```
14      <Index_File_Info>
15          <Index_File_Location>http://www.newportinstruments.com/IF1</Index_File_Location>
16          <Indexing_Parameters_Name>1 to 64 Chars</Indexing_Parameters_Name>
17      </Index_File_Info>
18      <Index_File_Info>
19          <Index_File_Location>http://www.newportinstruments.com/IF2</Index_File_Location>
20          <Indexing_Parameters_Name>1 to 64 Chars</Indexing_Parameters_Name>
21      </Index_File_Info>
22      <!--0 to 100 index files-->
23  </Acquisition_Context>
24</wave:Waveform>
```

The addition of cell subsets to FCS[15] has been an important improvement and can be enhanced by employing the functionality provided by the DICOM Waveform to express subsets as index files. The locations of two of these index files are given in S15 and S19. The names are specified in S16 and S20. The indices are based on DICOM Referenced Sample Positions. These are positions in one or more Channels in the Multiplex Group, which corresponds to flow cytometry list mode parameters. These positions are numbered starting with 1 and are equivalent to the indices of an array. Specific ranges or Segments of the array can be addressed. This capacity to specify a collection of individual events permits the identification of these events as members of a subset.

The use of an index in DICOM, as opposed to the addition of a parameter in the FCS list mode data, both simplifies the software and increases its execution speed. Since the software can index through all of the data that applies to a specific cell subset, the subsets can be analyzed or rendered sequentially rather than simultaneously. These Referenced Sample Positions can also be applied to single channels and employed to gate the list mode data.

The Acquisition_Context and Waveform respectively end on S23 and S24.

The elements of both the FCS Fluorescence Compensation Matrix and the DICOM Frame of Reference Transformation Matrix are listed in row-major order. However, there is no reason to be limited to DICOM for mathematical formulae, since this is the domain of the XML Mathematical Markup Language (MathML) Version 2.0[16]. The description of the compensation matrix can be based on Section 3.5.1 Table or Matrix (mtable) of reference 16, p. 91.

For Image data specific to the Acquisition Context, parts of the Visible Light Slide-coordinates Microscopic Image information object will be employed.

3.2 Multiplex Group.

A selection from the XML document that describes the data in a Multiplex Group is given below.

```
1 <Num_Waveform_Channels>10</Num_Waveform_Channels>
2<Num_Samples>50000</Num_Samples>
```

S1 and S2 state respectively that data from 10 parameters and from 50,000 cells have been acquired.

A slightly abbreviated schema for the Multiplex Group XML document is given below.

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <schema targetNamespace="file://C:\Stds\Cytometry_ML\multiplex_groups.xsd"
2a      xmlns="http://www.w3.org/2001/XMLSchema"
2b      xmlns:fcs="file://C:\Stds\Cytometry_ML\fcs_3_0.xsd"
2c      xmlns:multi="file://C:\Stds\Cytometry_ML\multiplex_groups.xsd"
2d      xmlns:dicom="file://C:\Stds\Cytometry_ML\dicom.xsd"
2e      xmlns:nums="file://C:\Stds\CytometryML\num_types.xsd
2f      elementFormDefault="unqualified" attributeFormDefault="unqualified">
3       <import namespace="file://C:\Stds\Cytometry_ML\dicom.xsd"
3a          schemaLocation="file://C:\Stds\Cytometry_ML\dicom.xsd"/>
4       <annotation> <documentation> The DICOM Number of Waveform Channels Type includes one element, Number, and two constant attributes, Tag and VR. The use of  attributes with a fixed value eliminates the need to enter a value into the Web page; yet, permits the value to be accessed by the application. There is no FCS keyword for this type. It is the 'n' in many FCS keywords. </documentation> </annotation>
```

The targetNamespace attribute for the Multiplex Group schema (S2) presently is a file. S2a-2e provide the namespaces and prefixes of the four other schemas and the multiplexed_groups.xsd schema itself (S2c). All of these schemas can supply elements, data types, and attributes. Since the preponderance of these come from the XMLSchema (S2a), it was chosen as the one schema without a prefix. S2f permits the elements and attributes to exist without prefixes. An import statement (S3 and S3a) is required to make the dicom (DICOM) schema visible. Similar import statements (not shown) exist for the fcs_3_0 (FCS3.0) and num_types schemas.

S4 is an annotation element that encloses documentation. Although this is essentially a typed comment, its use provides information to the XML parser.

```
5  <simpleType name="Num_Samples_Simple_Type">
6   <restriction base="nums:UInt32_Type">
7    <minInclusive value="1"/>
8    <maxInclusive value="2000000000"/>
9   </restriction>
10  </simpleType>
     <!--+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++-->
11  <annotation><documentation> DICOM Tag = (50xx,2006) Number of Samples, Unsigned Long, Value Multiplicity = 1.  The FCS Keyword is $TOT, which "specifies the total number of events in the data set." </documentation></annotation>
12  <complexType name="Num_Samples_Type">
13   <simpleContent>
14    <extension base="multi:Num_Samples_Simple_Type">
15     <attribute name="Tag" type="dicom:Tag_Type" fixed="50xx,2006"/>
16     <attribute name="VR" type="dicom:VR_Type" fixed="UL"/>
17     <attribute name="FCS_Keyword" type="fcs:FCS_Keyword_Type" fixed="$TOT"/>
18    </extension>
19   </simpleContent>
```

```
20  </complexType>
```

S5 through S10 describe Num_Samples_Simple_Type, which is based on an 32 bit unsigned integer (S6) and restricted to a range of values encompassing 1 (S7) to two billion (S8). For most studies, this maximum should be reduced. The annotation, S11, documents the complexType, Num_Samples_Type (S12). The Num_Samples_Type (S12) is based on an extension of the Num_Samples_Simple_Type (S14) to include constant values of the DICOM Tag (S15) and VR (S16) attributes. This provides a 1 to 1 correspondence with DICOM and will facilitate interconversion. Since this type is present in FCS3.0, an attribute (S17) which specifies the FCS keyword, $TOT, has also been included. Because the actual type is based on a simpleType (S14), the content is described as simpleContent (S13, S19). In the element declaration (S14), the multi prefix is required to identify the Multiplex_Groups_Type. The same structure based on the extension of a simple type is used with the Num_Waveform_Channels_Type (not shown).

```
21  <complexType name="Multiplex_Groups_Type">
22   <sequence>
23    <element name="Num_Waveform_Channels" type="multi:Num_Waveform_Channels_Type"/>
24    <element name="Num_Samples" type="multi:Num_Samples_Type"/>
25   </sequence>
26  </complexType>
27  <element name="Multiplex_Group" type="multi:Multiplex_Groups_Type"/>
28</schema>
```

Elements of the two types are combined (S21 to S26) into one global data structure Multiplex_Groups_Type. This combination is a sequence (S22 to S25), which contains two elements (S23 and S24). A sequence is the DICOM and XML format for an aggregate of objects or an element that contains one or more other elements. It is similar to a Pascal or Ada record, or a C struct. The use of a sequence in an XML document permits the number of times each element can be used and their order to be specified. The inclusion of the Multiplex_Group element (S27) permits XMLSpy to automatically produce an XML document.

XML schemas provide a precise means for defining the structure, content and semantics of XML documents. Schemas permit groups or organizations to create classes of documents which include shared vocabularies and a common thesaurus. Schemas include the definitions of: elements, attributes, and types.

3.3 Parameters (Channels)

Each of the parameters (channels) is described by a sequence that includes the equivalent information used in FCS to describe a parameter. In order to facilitate the creation and maintenance of the parameter schema, it has been based on multiple helper schemas: data, detectors, dicom, excitation, fcs_3_0, filters, item, num_types, staining, units, and multiplex_groups. DICOM, fcs_3_0, and multiplex_groups have previously been described.

An XML document that describes a parameter is shown below. In order to increase readability, familiar values for the elements and attributes have been included in this XML document. This document is separated by sections of explanatory text. Most of the individual sections of XML code have a one-to-one correspondence with an individual schema, which has been imported into the main parameters.xsd schema.

```
1<?xml version="1.0" encoding="UTF-8"?>
  <!--Sample XML file generated by XMLSPY v5 rel. 2 U (http://www.xmlspy.com)and
  edited by Robert C. Leif, Ph.D.-->
2<params:Parameter
2a xmlns:params="file://C:\Stds\CytometryML\parameters.xsd"
2b xsi:schemaLocation="file://C:\Stds\CytometryML\parameters.xsd
    C:\Stds\CytometryML\parameters.xsd"
2c xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
3   <Waveform_Channel_Number>1</Waveform_Channel_Number>
4   <Short_Name>FL1</Short_Name>
```

S1 is the same for all of the XML documents presently employed in this work. It describes the XML version and charac-

ter set. It is followed by a comment, which provides the credits for the work. S2 starts with a single encompassing element, Parameter, that was required to automatically generate the entire XML document with XMLSpy. Parameter is the only global element in the Parameters schema and it is of the complexType, Parameter_Type, that includes effectively all of the imported types, which provide the elements and attributes for this XML document. The name-space including the prefix for the schema and its location are given respectively in S2a and 2b. The schema location for XML including its standard prefix is given in S2c, which ends S2. The description of the first Parameter (Channel) sequence starts with the Waveform Channel Number (S3) which is equal to the FCS parameter number, n. The present maximum number of parameters is 100, which is more than ample. Since the sequence construct is used, the value of n needs to be given only once; rather than, as in FCS, being included with each data element. This first parameter has (S4) the common Short_Name (abbreviation), FL1.

```
<Analyte_Info>
  <Binding_Species>IgG</Binding_Species>
  <Binding_Species_Name>Anti5BrdU</Binding_Species_Name>
  <Tag_Name>Fluorescein</Tag_Name>
  <Analyte_Formula_Wt>150000</Analyte_Formula_Wt>
   <Item_General_Info>
    <Manufacturer>Phoenix Flow Systems</Manufacturer>
    <Item_Lot-number>Sean 1</Item_Lot-number>
   </Item_General_Info>
   <Comment>Best There is.</Comment>
</Analyte_Info>
```

The Analyte_Info (S5 to S15) can be based on the LOINC database[17]. For instance, 522 items are listed in the LOINC CellMark Class. The LOINC nomenclature presently describes test results; however, it can be used as a basis for the Binding_Species_Name (S7) in a format suitable for analytical cytology. Following the suggestion of the ISAC Data Standards Committee[18], the lot number (S12) has been included. The complete description of the specimen is separate from that of the Waveform (List Mode) data.

```
<Dectector_Info>
  <Detector>PMT</Detector>
  <Detector_Setting>600</Detector_Setting>
  <Detector_Units Prefix="none" Si_Unit_Name="volt"/>
  <Measurement>Fluorescence</Measurement>
  <Emission_Filter_Info Prefix="nano" Unit="meter">
    <Emission_Filter>Band_Pass</Emission_Filter>
    <Band_Width_Location>unknown</Band_Width_Location>
    <Peak_1>535</Peak_1>
    <Band_Width_1>45</Band_Width_1>
    <Description>535AF45</Description>
    <Item_General_Info>
      <Manufacturer>Omega Optical</Manufacturer>
      <Model_Name>XF3084</Model_Name>
    </Item_General_Info>
  </Emission_Filter_Info>
  <Beam_Splitter_Info Prefix="nano" Unit="meter">
    <Beam_Splitter>Dichroic_Reflect_Low</Beam_Splitter>
    <Low_Cut_Off_1>505</Low_Cut_Off_1>
    <Description>505DRLP</Description>
    <Item_General_Info>
      <Manufacturer>Omega Optical</Manufacturer>
      <Model_Name>XF2010</Model_Name>
```

```
39      </Item_General_Info>
40    </Beam_Splitter_Info>
41  </Dectector_Info>
```

The detector information (S16 to S41) includes the type of detector (S17) and its units. The detector types are coded as an enumerated type, which presently includes: PMT, multi-anode PMT, diode, avalanche diode, diode array, CCD camera, DC impedance, AC impedance, software, and other. Software has been included because a parameter can be based on a calculation, which often can involve more than 1 detector. The inclusion of other is based upon a lack of omniscience. The detector, in this case, is a PMT run at 600 Volts. Both the standard prefixes (S19): milli, micro, etc. and the unit name (S19) come from the Unified Code for Units of Measure (UCUM)[19]. The use of "none" to describe a value that does not have a prefix is new. The measurement (S20) is an enumerated type, which includes: fluorescence, light scatter (low angle, 45 or 90 degree), extinction, dc or rf impedance, and other. Both the beam splitter (S32 to S40) and the emission filter (S21 to S31) can have up to 3 wavelengths.

```
42  <Amplifier_Info>
43    <Mode>Log</Mode>
44    <Gain>100</Gain>
45  </Amplifier_Info>
```

Both the amplifier mode (S43), linear or log and gain (S44) are specified.

```
46   <Data_Info>
47    <Data_16 Num_Bits_Stored="10" Num_Bits_Allocated="16"/>
48  </Data_Info>
```

The Data_Info (S46 to S48) includes a the data class (S47) element. Presently, the data classes include unsigned integers, reals, the time offset since the experiment started, and arrays. The unsigned integers include: 8, 16, 32, and 64 bit d.types. Data_16 is an unsigned 16 bit integer. Both 32 and 64 bit floats have been included. The time in seconds since the experiment started, Time_Offset. Time_Offset is expressed as a decimal with 6 fractional digits. One, two, and three dimensional arrays have also been included. The one dimensional arrays can correspond to a spectrum, such as fluorescence emission. The two dimensional arrays can correspond to images of selected cells from a microscope slide. And the three dimensional images can be selected regions of confocal microscope images. The precision of the measurement is specified by the attribute Num_Bits_Stored and the size of the storage unit by the attribute Num_Bits_Allocated. A character type could be added. However, this does not appear to be necessary. Data_Info only describes the data-type; it does not include any data. Since Cytometry data often is quite large, it is stored in a separate binary list-mode file, which corresponds to a one dimensional array of records. The Data_Info for a given parameter specifies one element of the record or struct.

```
49  <Excitation_Info>
50    <Light_Source_Info>
51      <Light_Source Emitter="Ar" Polarization="None">Gas_LASER</Light_Source>
52      <Power Prefix="milli" Si_Unit_Name="watt">25</Power>
53      <Wavelength>488</Wavelength>
54      <Description>heat exchanger cooled, CW Argon ion</Description>
55      <Item_General_Info>
56        <Manufacturer>Coherent</Manufacturer>
57        <Model_Name>Enteprise II 622</Model_Name>
58        <Item_Serial-number>xyz123</Item_Serial-number>
59      </Item_General_Info>
60    </Light_Source_Info>
61    <Excitation_Filter_Info Prefix="nano" Unit="meter">
62      <Excitation_Filter>LASER</Excitation_Filter>
63      <Band_Width_Location>1/e</Band_Width_Location>
64      <Peak_1>488</Peak_1>
65      <Band_Width_1>0.1</Band_Width_1>
```

```
66      <Description>none</Description>
67      <Item_General_Info>
68        <Manufacturer>none</Manufacturer>
69        <Model_Name>none</Model_Name>
70      </Item_General_Info>
71    </Excitation_Filter_Info>
72  </Excitation_Info>
73</params:Parameter>
```

The excitation information (S49 to S72) includes information on both the light source (S50 to S60) and if there is one, an excitation filter (S61-S71). The formats are respectively similar to the Detector_Info (S16 to S41) and the Emission_Filter_Info (S21 to S31). Presently, the light source emitters include: Ar, GaAlAs, GaAs, HeNe, Hg, Hg-Xe, Xe, YAG, Enzyme-Substrate, and Other"; the polarizations are: Vertical, Horizontal, and None; the types include: Gas-Laser, Solid-State-Laser, Semiconductor-Laser, Arc, Flash-Lamp, Chemo-Luminescence, None, and Other.

4. RESULTS

The necessary DICOM and FCS data types have been represented in XML schema. Besides fulfilling the original requirement for a common standard for flow and microscopic image cytometry[4], CytometryML now also includes Plate_Image and is suitable for its microscopic equivalent, array slides. At present, the settings for flow cell sorting have been omitted. The Waveform Information Object representation of the flow cytometry information was a significant improvement over the original FCS design and fulfills the original requirements for a cytometry standard. The analysis and description of fluorescence compensation can be described in Mathematical Markup Language, MathML[16]. Pilot studies (not published) have demonstrated that the actual binary data can be represented in memory as a single dimensional array of a record type or struct with each element being the value of a parameter's data. This array can then be stored as an index file, where records can be stored and retrieved by their position in a file. This same type of index file can be used to store the values of the positions (indices) of the cells in a subset.

5. FUTURE

The use of XML permits upgrading and extension of CytometryML. When relevant new technologies and modalities arise, a simple consensus system can be employed to extend the enumerated types or, if necessary, the other choice in the enumeration can initially be used. A more formal approach will be necessary when new schema have to be added or new data-types have to be added to existing schema. Investigators and manufacturers can add their own schemas and post them on the Web. Since XML has very good semantics for single inheritance, data-types from one schema can be extended in a new schema.

6. CONCLUSIONS

XML has the advantages over DICOM of: widespread commercial acceptance; the capacity to represent foreign data as schema; excellent formatting and graphics capabilities, suitable for the creation of structured reports; and the ability to exchange data with mainstream software, such as Microsoft Excel and in the near future most of the rest of Microsoft Office and many other programs. The description of the mathematical elements in flow cytometry standard can be greatly simplified by the use of MathML. Since the list-mode data has now been separated from the description of the experiment and is contained in a separate file with a simple structure, analysis of this data with commercial off-the-shelf software should require minimal interfacing. The situation is even simpler for the image cytometry data, which is to be kept in standard file formats, such as TIF and JPEG; this data will be directly accessible by commercial programs, such as PhotoShop.

The collection of XML schemas and pages is programming language independent. DICOM has the advantages of a well tested design of data-types, structures, and methods. The basing of CytometryML on two existing well documented and tested standards, DICOM and XML, should decrease the regulatory documentation burden on commercial manufacturers by permitting them to cite these two existing standards. It would be a worthwhile endeavor to investigate the feasibility of implementing the next version of DICOM in XML syntax with the exception of large binary data types.

ACKNOWLEDGEMENTS

The generous support of Newport Instruments is greatly appreciated.

REFERENCES

1. L. Seamer, B. Bagwell, L. Barden, M. Christofferson, L. E. Magruder, G. Malachowski, R. F. Murphy, D. Redelman, G. C. Salzman, and J. C. S. Wood. “Data File Standards Committee of the International Society for Analytical Cytology (ISAC), “Data File Standard for Flow Cytometry, Version FCS3.0,” http://www.isac-net.org/ 1996.

2. L. C. Seamer, C. B. Bagwell, L. Barden, D. Redelman, G. C. Salzman, J. C. S. Wood, and R.F. Murphy, “Proposed New Data File Standard for Flow Cytometry, Version FCS 3.0.”, Cytometry 28, 118-122, 1997.

3. R. C. Leif and S. B. Leif, "Evolution of Flow Cytometry Standard, FCS3.0, into a DICOM-Compatible Format". *Optical Diagnostics of Biological Fluids and Advanced Techniques in Analytical Cytology*, Ed. A. V. Priezzhev, T. Asakura, and R. C. Leif. A. Katzir Series Editor, Progress Biomedical Optics Series, SPIE Proceedings Series, Vol. 2982, pp. 354-366, 1997.

4. R. C. Leif and S. B. Leif, "A DICOM Compatible Format for Analytical Cytology Data", in *Optical Investigations of Cells In Vitro and In Vivo*, D. L. Farkas, R. C. Leif, B. J. Tromberg, Editors, A. Katzir Biomedical Optics Series Ed. Proc. of SPIE Vol. 3260, ISBN 0-8194-2699-7 pp. 282-289, 1998.

5. R. C. Leif, and S. B. Leif, "A DICOM Compatible Format for Analytical Cytology Data, that can be Expressed in XML", *Optical Diagnostics of Living Cells IV*, D. L. Farkas and R. C. Leif, Editors, SPIE Proceedings Vol. 4260 pp. 238-48, 2001.

6. R. C. Leif, S. B. Leif, and S. H.Leif, “CytometryML, An XML Format based on DICOM and FCS for Analytical Cytology Data”, submitted to Cytometry.

7. H. S. Thompson, D. Beech, M. Maloney, and N. Mendelsohn. *XML Schema Part 1: Structures*, W3C® (2001); www.w3.org/TR/2001/REC-xmlschema-1-20010502/ 2001.

8. P. V. Biron, A. Malhotra. *XML Schema Part 2: Datatypes*, W3C® (2001); www.w3.org/TR/2001/REC-xmlschema-2-20010502/ 2001.

9. *Digital Imaging and Communications in Medicine (DICOM) PS 3.3 - 2001, A.34 Waveform Information Object Definitions*, p. 112 and C.10.8 Waveform Identification Module pp. 489-499, National Electrical Manufacturers Association, VA, USA, 2001. http://medical.nema.org/dicom/2001/01_03PU.PDF. 2001

10. T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, Editors, *Extensible Markup Language (XML) 1.0 (Second Edition)* W3C Recommendation 6 October 2000. http://www.w3.org/TR/REC-xml/ 2000

11. N. G. Leveson, *Safeware, System Safety and Computers,* Addison-Wesley, ISBN 0-201-11972-2 p. 419, 1995.

12. *Digital Imaging and Communications in Medicine (DICOM) PS 3.6 - 2001, Part 6: Data Dictionary, 6 Registry of DICOM data elements,* pp. 5-60, 2001. http://medical.nema.org/dicom/2001/01_06PU.PDF

13. ECMA, Originally, European Computer Manufacturers Association. Now, ECMA International - European association for standardizing information and communication systems. http://www.ecma.ch/

14. *Digital Imaging and Communications in Medicine (DICOM) PS 3.3 - 2001* Table C.10-8, Waveform Identification Module Attributes, Note: “The Acquisition Datetime (0008,002A) is the time of the original waveform data capture. Derived waveforms which are processed (e.g., averaged or filtered) and encoded subsequent to the waveform Acquisition Datetime have a Content Date (0008,0023) and Content Time (0008,0033) representing the time of the processing. In all cases the actual date and time of creation of the SOP (Service-Object Pair) Instance for transmission or storage may be recorded in the Instance Creation Date (0008,0012) and Instance Creation Time (0008,0013)”. 2001

15. Redelman D, Coder DM. “Cell subset (CS) parameter to record the identities of individual cells in flow cytometric data.” Cytometry: **18, pp.** 95-102, 1994.

16. *Mathematical Markup Language (MathML) Version 2.0*, W3C Recommendation 21 February 2001. Carlisle D, Ion P, Miner R, Poppelier N, editors. http://www.w3.org/TR/2001/REC-MathML2-20010221/ 2001.

17. LOINC® and RELMA™. The Regenstrief Institute, http://www.regenstrief.org/loinc/loinc_information.html and *LOINC® Database Version 2.05*, Released: February 8, 2002 http://www.regenstrief.org/loinc/loincdb.pdf, 2002.

18. R. F. Murphy. International Society for Analytical Cytology, Data Standards Committee Meeting Minutes, Date: January 4-5, 2002. http://www.isac-net.org/ 2002.

19. G. Schadow, C. J. McDonald. *The Unified Code for Units of Measure (UCUM), Version 1.4*, April 27, 2000. Regenstrief Institute for Health Care. http://aurora.rg.iupui.edu/UCUM/UCUM.pdf, 2000.