CytometryML and other data formats

Robert C. Leif*

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

ABSTRACT

Cytology automation and research will be enhanced by the creation of a common data format. This data format would provide the pathology and research communities with a uniform way for annotating and exchanging images, flow cytometry, and associated data. This specification and/or standard will include descriptions of the acquisition device, staining, the binary representations of the image and list-mode data, the measurements derived from the image and/or the list-mode data, and descriptors for clinical/pathology and research. An international, vendor-supported, non-proprietary specification will allow pathologists, researchers, and companies to develop and use image capture/analysis software, as well as list-mode analysis software, without worrying about incompatibilities between proprietary vendor formats.

Presently, efforts to create specifications and/or descriptions of these formats include the Laboratory Digital Imaging Project (LDIP) Data Exchange Specification; extensions to the Digital Imaging and Communications in Medicine (DICOM); Open Microscopy Environment (OME); Flowcyt, an extension to the present Flow Cytometry Standard (FCS); and CytometryML.

The feasibility of creating a common data specification for digital microscopy and flow cytometry in a manner consistent with its use for medical devices and interoperability with both hospital information and picture archiving systems has been demonstrated by the creation of the CytometryML schemas. The feasibility of creating a software system for digital microscopy has been demonstrated by the OME. CytometryML consists of schemas that describe instruments and their measurements. These instruments include digital microscopes and flow cytometers. Optical components including the instruments' excitation and emission parts are described. The description of the measurements made by these instruments includes the tagged molecule, data acquisition subsystem, and the format of the list-mode and/or image data. Many of the CytometryML data-types are based on the Digital Imaging and Communications in Medicine (DICOM). Binary files for images and list-mode data have been created and read.

Keywords: CytometryML, DICOM, FCS, LDIP, OME, flow cytometry, pathology informatics, specification, standard, XML schema

1. INTRODUCTION

A cytometry-pathology specification and/or standard for research and clinical use needs to be created in an open manner, which permits peer review. The oversight and development of a specification should be by a committee composed of researchers, service-oriented staff including pathologists and scientists, and vendors. In order to achieve acceptance by multiple societies and other organizations and avoid the development of multiple competing specifications, these groups need to pool their efforts. Thus, a development committee should contain representatives from as many of the societies and groups that are involved in the field, as possible. Two of these societies are the International Society for Analytical Cytology (ISAC), which has developed FCS[1,2] (Flow Cytometry Standard), and the Association for Pathology Informatics, which has started on the Laboratory Digital Imaging Project (LDIP) Data Exchange Specification[3]. There also exists the Digital Imaging and Communications in Medicine (DICOM) Working Group 26, the Open Microscopy Environment (OME)[4], and CytometryML[5].

The specification should consist of a collection of XML schemas that define common data types, elements, and attributes. Wherever possible, it should be based on existing standards. The structures of the data-types must be compatible with their storage in conventional data-bases including those that are part of laboratory information systems. The binary image and list-mode data should be available to the end user in the form of one or more standard file formats. The elements in or produced by these schemas can serve as the subjects and objects for XML languages, such as RDF[6] and OWL[7].

The problem of cytology-pathology standardization has been greatly confounded by the differences in the requirements, needs, and past histories of the societies and groups. At this point, LDIP's main interest is in the relationships between different objects (elements); ISAC and OME share a common interest in the transfer of data; and the DICOM Working Group 26 plans to evaluate and extend the current DICOM standard as it relates to newer microscopic techniques includ-

*rleif@rleif.com; phone 1 619 582-0437

ing whole slide imaging. The heterogeneity of the imaging modalities, flow, two dimensional, three dimensional and temporal renditions of these data, and the differences in data obtained from single cells. as opposed to tissues, provide significant complications to the creation of a common data standard.

The first implementation of LDIP (http://www.ldip.org/) will be in the form of XML Resource Description Framework (RDF) schemas[6], which will express the relationships between data-types. OME and CytometryML employ schemas written in XML Schema language (XSD)[8,9], which defines data-types, and DICOM employs a rich syntax[10], which defines and expresses the relationships between data-types, as well as providing methods for their uses.

The Flowcyt project (http://www.flowcyt.org/) is working on the development of bioinformatics standards for flow cytometry and statistics packages "to allow users to implement analyses in a high throughput fashion, as well as exchange these analyses in more meaningful ways then are currently available." Their first project is the development of a gating standard for flow cytometry and sorting.

A promising solution to this problem of multiple implementations is to develop a common ontology, shared vocabulary, that specifies the data-types and objects derived therefrom. Although these data-types will be expressed employing different syntaxes, their semantics and definitions will be identical. This will facilitate translation between the multiple implementations and interfacing of software written with standard programming languages, such as Ada, C, C#, C++, FORTRAN, Java, etc. The use of common data-types and definitions will also facilitate the creation and use of databases that can import and export data to and from implementations based on syntactically different standards.

DICOM[10] (http://medical.nema.org/dicom/) is a functioning, reliable clinical information system standard, as demonstrated by its extensive use in radiology. The College of American Pathologists (CAP) was involved in DICOM Supplement 15: Visible Light Image for Endoscopy, Microscopy, and Photography[11]. The newly created DICOM Working Group 26 plans to update and extend Supplement 15. In which case, DICOM will serve to connect the data with the pathologist or other individual who makes a clinical decision based on the data. DICOM already includes digital slide microscopy and a Waveform Information Object that is the basis for the design of the list-mode metadata in CytometryML[5,12]. Since DICOM was created prior to the creation of the World Wide Web, it has many of its own facilities. At present, the DICOM interface to the Internet is described in DICOM Part 18: Web Access to DICOM Persistent Objects (WADO)[13]. Presently images, regions of images, and structured reports in HTML format can be retrieved via the Internet. Annotations containing the patient name and technical information can be retrieved by being burned into an image. It appears, at present, that there is no means to retrieve the information present on a DICOM server as XML tagged information or to transmit XML tagged information to a DICOM server. Because DICOM is the largest relevant ontology, its data-types and documentation should be considered for reuse in the RDF and XSD schemas. However DICOM data-types and methods should not be reused where existing XML data-types and tools are available.

The availability of XML artifacts that are based on DICOM information objects will greatly facilitate bidirectional translation between DICOM and XML. This will permit DICOM to use XML tools, such as XForms[14] for input, and XML based commercial off-the-shelf software, such as word processors, spreadsheets, and databases.

The XML Schema definition language, XSD can be used as a design language for data-types of software constructs including a software standard, such as DICOM. The arguments for this use include: 1) Since XSD is an XML language, it was designed to maximize readability; 2) XSD can be used for object-oriented development including the creation of strongly typed data structures; 3) XSD schemas can be checked for being well formed and valid.

2. METHODS

The CytometryML schemas were developed using the XSD language. They primarily consisted of DICOM data-types with XSD documentation elements that included references to the descriptions of the data-types in the DICOM standard[10], specifically Part 3: Information Object Definitions[15]. Data-types that could be part of an extension of the DICOM standard followed that standard, as much as reasonable, and were constructed primarily from previously created CytometryML data-types that had been based on the DICOM standard. The only major exceptions to this were 1) the use of the unambiguous names for the numeric types that were included in the ECMA standard[16], which is used by Microsoft and 2) the use of data-types that are specific to XML, such as AnyURI. As previously[12,17], in order to completely identify the data elements and facilitate traceability, the DICOM Data Element Tag and Value Representation, VR, have been included as fixed (constant) attributes. Similarly, where relevant, the FCS Key Words have been included as fixed

attributes. Two sets of schemas have been created. The CytometryML schemas, which are specific to cytometry and the XML_Utilities that consist of general purpose tools.

The new schemas and the updates to those previously published[12,17] were made with Stylus Studio® XML 6 Professional Edition (http://www.stylusstudio.com). 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 pretty-printers of case insensitive languages. All types end in _Type. When elements and attributes are based on new types, they have the same name as the type except that the type has the suffix _Type.

3. DESCRIPTION OF SCHEMAS

3.1. Description of the XML Utilities

The XML Utilities consist of 4 schemas: configuration, person_name, num_types, and units. Configuration is to be imported by every schema. It provides the meta-information that will permit a group of programmers to develop and maintain the XML Utilities and CytometryML schemas. This information includes: the subject, a short description, a reference to the source of the data-types, version, location on the web, regulatory status, copyright holder and permissions, maintainer's name and E-mail address, release date, and keywords. The person_name schema includes elements that provide its complete description including form of address and degrees. The names of the types in num_types are based on the simple, concise ones provided by a public standard, ECMA[16]. 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 Units schema is primarily based on the Unified Code for Units of Measure (UCUM 1.4)[18]. It includes the abbreviations, name, and prefix (milli, micro, etc.) of scientific units. Each type of scientific unit has its own data-type; and the units are derived by restriction from a parent starting with Units_Type.

3.2. Description of CytometryML Higher Level Schemas

CytometryML presently comprises 19 schemas. The 5 schemas added to those previously described[12.17,19] are optics, instrument, slide, pixel, and image. The references to the specific DICOM data-types given below are all to Part 3: Information Object Definitions. The specific locations referenced in the following text are from DICOM Part 3.

Optics: The optics schema includes simple optical elements, such as lenses, mirrors, prisms, and polarizers, as well as, complex optical elements, such as objectives and condensers. The complex optical elements are further described in terms of their field flatness, immersion, chromatic properties, NA, working distance, magnification, and contrast method, which includes none.

Instrument: The instrument schema includes a generic optical instrument type, which combines the characteristics of both a microscope and a flow cytometer. The common characteristics are: the orientation of the optical system, upright, inverted, or plainer; one or more objectives; none or more condensers; the type of carrier, liquid or a fixed substrate, such as a microscope slide; the type of image, stereo or mono; and whether the instrument sorts. Since the Instrument_Type is a description of the basic hardware of the instrument, it does not include information that is specific to a particular type of measurement. This information is described in the channels schema. The Microscope_Type and Flow_Cytometer_Type are formed by restriction from the Instrument_Type. For instance, a microscope very often has more than one objective; whereas a flow cytometer has only one. The optical assembly that focuses light at 90^{O} to the objective into the flow cell can be considered to be a condenser.

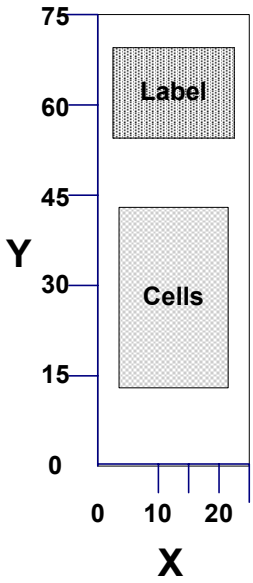

Figure 1. Locations of the DICOM slide coordinates.

Slide: The slide schema includes an Offset_Type, which is a vector consisting of the X, Y, and Z offsets. The X and Y offsets are the distance in millimeters from the origin of the slide. As is shown in Figure 1, the origin (0,0) is on a corner of the end of the side that is opposite from the label. The Y axis is the long direction and the X axis is the short direction of the slide. The zero points of both axes are located at the origins. Therefore increasing the distance in the Y direction between the label of the slide and the objective decreases the value of Y. The Z offset is the distance in microns from the surface of the slide closest to the objective.

Pixel: Since the DICOM pixel type is limited to monochrome and the three primary colors, it is inadequate for cytometry. A simple solution is to incorporate the Channel_Type (parameter) into the Pixel_Type, which is described in the pixel schema. This results in the list-mode data produced by either digital microscopy and flow cytometry being very similar. The major differences are flow often uses the time after the start of a run as a parameter and digital microscopy uses the X, Y, and often Z coordinates. The rest of the parameters for both modalities are overlapping subsets of the class of cytometry parameters, Channel_Type, which has been defined in the channels schema. The XSD code for the complexType Pixel_Type is shown below.

```
1. <complexType name="Pixel_Type">
2.    <sequence>
3.       <element name="Samples_Per_Pixel"
            type="pixel:Samples_Per_Pixel_Type"/>
4.       <element name="Photometric_Interpretation"
            type="pixel:Photometric_Interpretation_Type"/>
5.       <element name="Pixel_Aspect_Ratio"
          type="pixel:Pixel_Aspect_Ratio_Type"/>
6.       <element name="Channels" type="channels:Channel_Type"
            minOccurs="1" maxOccurs="20"/>
7.       <element name="Location" type="slide:Offset_Type"/>
8.    </sequence>
9. </complexType>
```

Since the XML Schema language may be unfamiliar, a brief description of its syntax is necessary. All XML statements begin with the '<' character and end with the '>' character. The end of an XML statement includes the '/' character. Line numbers have been added at the far left. XML is a nested language that, as shown in line 1, employs the less than character '<' to begin a construct. The first element is a complexType, which indicates that it is composed of a structure that includes multiple entities or includes one or more attributes. It has an attribute, name. Attributes are always based upon simple_Types, which are only composed of a single elementary type, such as a string or an integer. The value of an attribute is set in quotation marks and follows the equals sign. In this case, it is Pixel_Type. Since line 1 encapsulates the statements that follow, it ends in the > character. Pixel_Type is a sequence (line 2) (record or struct) that includes the element type pairs specified in Table C.7-11a, Image Pixel Module Attributes, with the addition of Channel_Type. The Samples_Per_Pixel (line 3) is the number of parameters measured. The pixel: following the type attribute indicates that the data-type was imported from the pixel schema. Samples_Per_Pixel_Type was derived from the Num_Waveform_Channels_Type in the multiplex_groups schema, which is used to describe list-mode data. The Photometric_Interpretation (line 4) is an enumeration of types including Monochrome, RGB, Multidimensional, and others specific to DICOM. The Pixel_Aspect_Ratio (Line 5) specifies the ratio of the vertical size (Y) and horizontal (X)

size. The Channels (line 6) describes each of the channels (parameters); it was imported from the channels schema. Each channel is itself a sequence that describes a complex data structure that includes: a description of the analyte binding species, which includes a description of the analyte; the excitation source including its filter; the detector and its optics; the amplifier; and the data-type of the parameter. At least one and up to 20 parameters are included. The Location (line 7) specifies a pixel's X and Y coordinates and, if it is included, the Z coordinate.

Image: The image schema provides the metadata for the binary image files. The XSD code for the complexType Image_Type is shown below.

```
1<complexType name="Image_Type">
2  <sequence>
3     <element name="Columns" type="image:Columns_Type"/>
4     <element name="Rows" type="image:Rows_Type"/>
5     <element name="Planar_Configuration"
         type="image:Planar_Configuration_Type"/>
6     <element name="Pixels" type="pixel:Pixels_Type"/>
7     <element name="General_Image" type="image:General_Image_Type"/>
8     <element name="Acquisition_Date_Time"
         type="time:Acquisition_Date_Time_Type"/>
9     <element name="Compression" type="image:Compression_Type"/>
10    <choice>
11       <element name="Microscope" type="instr:Microscope_Type"/>
12       <element name="Flow_Cytometer"
            type="instr:Flow_Cytometer_Type"/>
13    </choice>
14    <element name="File_Location" type="anyURI"/>
15 </sequence>
16 <attribute name="Endian" type="dicom:Endian_Type"
      default="Little_Endian"/>
</complexType>
```

The Columns and Rows elements (lines 3 & 4) specify respectively the number of pixels in the horizontal and vertical directions. The Planar_Configuration element (line 5) specifies whether the image will be by color-by-plane where the image is a stack of monochrome images or color-by-pixel where there is one image with each pixel being a vector of multiple channels (parameters). The Pixels element (line 6) is from the pixel schema. It is identical to the Pixel_Type described above except that its sequence lacks the Location element, which is only relevant to an individual pixel. The General_Image element (line 7) includes whether the image is original or has been derived by some process and whether the image is the left or right member of a stereo pair or is monocular. The Acquisition_Date_Time element (line 8) is the date and time that the image was acquired or was derived. The Compression element (line 9) can either be lossy or lossless. For both, the image file format type JPEG, JPEG 2000, DICOM, TIFF, etc. is given. For lossy compression, the compression method and ratio are also given. Since the instrument, as described above, can be either a Microscope or a Flow_Cytometer element, the instrument is represented as an XML Schema language Choice element (line 10). Since the File_Location element location (line 14) is a URI, it can either be stored on the web, a workstation disc, or on a local server. The Endian attribute (line 16) is from the dicom schema; In agreement with DICOM, the default is little endian. Unfortunately, DICOM and OME do not agree on the default value for endian. It appears that the OME[4] has big endian as the default in the OME schema.

4. DISCUSSION

The development process for the CytometryML and XML_Utilities schemas adhered to the principle of standards pau-

city. Most of the data-types and their documentation were reused from existing standards, principally DICOM and ECMA. This reuse minimized the design effort and maximized the reliability of the schemas due to the previous successful use of many of the data-types in implementations of DICOM. The creation of new data-types was facilitated by extending and/or enhancing already existing data-types. In the future, the CytometryML schema should be enhanced by reuse of data-types and members of enumerations from the OME schema[4].

The CytometryML schemas could be used as a means to increase the compatibility of DICOM with XML for both data input and output. It would be useful to employ XML tools, such as XForms[14], for data input and XML with XSL-FO[20] or CSS[21] for the output of structured reports. Scalable vector graphics[22] (SVG) could be employed for image overlays. The design of translators between CytometryML and DICOM should be simplified by the use of data-types in the CytometryML schemas that are based on DICOM, which have constant attributes for DICOM tags and value representations. The tags can be used for look-up tables.

Although the ease of use of separate binary files has been demonstrated[23], there is a real possibility that the XML and binary files could be separated. The utility of the information provided by the combination of both files is much greater than the utility of either the XML or binary file. Conversely, when the two types of files are combined their utility is minimal if they can not be separated. One standard approach to combining XML metadata with binary data is the use of Zip files and, which according to Microsoft[24] are "The XPS (XML Paper Specification) physical format understood in Windows Vista™". A second is MHTML[25], which is a MIME Encapsulation of Aggregate Documents, such as HTML. It is the standard for saving web pages as a single file. The extension used by Microsoft Windows for MHTML files is MHT.

In order to facilitate the maintenance and extension of the CytometryML schemas, they were organized as a hierarchy with the two top-level schemas being image and waveform. This hierarchical organization together with the inclusion of the elements in the configuration schema into each schema should permit multiple maintainers to simultaneously maintain or extend the schemas.

5. CONCLUSIONS

The feasibility of a combined specification for the metadata for both digital microscopy and flow cytometry has been demonstrated by the above results and previous publications[12,17,23]. The ability to store and manipulate binary data stored in an open simple formats has previously been demonstrated[23]. The list-mode data were stored as a file that corresponded to a simple array of feature vectors (records) and the images in the Adobe® Photoshop® RAW format, which can be as simple as a two-dimensional array of monochrome or RGB pixels. The CytometryML schemas demonstrate that the principle of standards paucity works. These schemas also demonstrate that in the unfortunate situation where there is not agreement on a single standard or where there is a large amount of legacy software, it is possible to employ common, traceable data-types as a means to mitigate the situation.

6. ACKNOWLEDGMENTS

I wish to thank the following individuals for providing information on and explanations of their standardization efforts: Ulysses Balis, Bruce Beckwith, Jules Berman, Ryan Brinkman, Bruce Friedman, and Ilya Goldberg. Newport Instruments internal development funds have supported this project.

7. REFERENCES

1. L. C. Seamer, C. B. Bagwell, L. Barden, D. Redelman, G. C. Salzman, J. C. Wood, R. F. Murphy, "Proposed new data file standard for flow cytometry", version FCS 3.0. *Cytometry* **28**, pp. 118–122 1997.
2. FCS, Flow Cytometry Standard http://www.isac-net.org/ Then search for FCS.
3. LDIP, Laboratory Digital Imaging Project (LDIP) Data Exchange Specification, http://www.ldip.org
4. OME, Open Microscopy Environment, http://www.openmicroscopy.org
5. R.C. Leif, S.H. Leif, S.B. Leif, "CytometryML, a markup language for analytical cytology", in *Manipulation and Analysis of Biomolecules, Cells and Tissues*, D. V. Nicolau, J. Enderlein, and R. C. Leif, Editors, SPIE Proceedings Vol. 4962, pp. 288-297, 2003.
6. RDF/XML Syntax Specification (Revised), W3C Recommendation 10 February 2004, Latest version: http://www.w3.org/TR/rdf-syntax-grammar/
7. OWL Web Ontology Language Overview, W3C Recommendation 10 February 2004, http://www.w3.org/TR/owl-features/
8. XML Schema Part 1: Structures Second Edition, W3C Recommendation 28 October 2004, Latest version: http://www.w3.org/TR/xmlschema-1/

9. XML Schema Part 2: Datatypes Second Edition, W3C Recommendation 28 October 2004, Latest version: http://www.w3.org/TR/xmlschema-2/
10. DICOM, Digital Imaging and Communications in Medicine http://medical.nema.org/dicom/2004.html
11. DICOM Supplement 15: Visible Light Image for Endoscopy, Microscopy, and Photography, http://www.dclunie.com/dicom-status/status.html#SupplementsByNumber, 1998.
12. R. C. Leif, S. B. Leif, and S. H. Leif, "CytometryML, An XML Format based on DICOM for Analytical Cytology Data", *Cytometry* 54A pp. 56-65, 2003.
13. DICOM Part 18: Web Access to DICOM Persistent Objects (WADO), http://medical.nema.org/dicom/2004.html
14. XForms 1.0, W3C Recommendation 14 October 2003, http://www.w3.org/TR/xforms.
15. DICOM Part 3 from http://medical.nema.org/dicom/2004/04_03PU3.PDF.
16. Standard ECMA-335, Common Language Infrastructure (CLI), 3rd edition (June 2005), http://www.ecma-international.org/publications/standards/Ecma-335.htm
17. R. C. Leif and S. B. Leif, "A DICOM compatible format for analytical cytology data, that can be expressed in XML." In: Farkas DL, Leif RC, editors. Optical diagnostics of living cells IV. Vol. 4260. Bellingham, WA: SPIE Proceedings; 201. pp. 238–248, 2001.
18. 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.
19. CytometryML, Cytometry Markup Language,. http://www.newportinstruments.com/cytometryml/cytometryml.htm.
20. Extensible Stylesheet Language (XSL)Version 1.0, W3C Recommendation 15 October 2001, http://www.w3.org/TR/xsl/.
21. Cascading Style Sheets, level 1, W3C Recommendation 17 Dec 1996, revised 11 Jan 1999, http://www.w3.org/TR/REC-CSS1.
22. Scalable Vector Graphics (SVG) 1.1 Specification, W3C Recommendation 14 January 2003, http://www.w3.org/TR/SVG11/.
23. R. C. Leif, "CytometryML, Binary Data Standards", in *Manipulation and Analysis of Biomolecules, Cells, and Tissues II*, D. V. Nicolau, J. Enderlein, R. C. Leif, and D. Farkas, Editors, SPIE Proceeding Vol. 5699 pp. 325-333 (2005).
24. Integrate Data, Finding and sharing data using Windows Vista™ developer technologies, Microsoft, http://msdn.microsoft.com/windowsvista/integrated/ (last visited Oct. 2005).
25. J. Palme, A. Hopmann, and N. Shelness, Network Working Group, Request for Comments: 2557
Category: Standards Track, March 1999, http://www.ietf.org/rfc/rfc2557.txt (last visited Oct. 2005) or MIME Encapsulation of Aggregate Documents, such as HTML (MHTML), 2557, http://www.rfc-editor.org/rfcxx00.html.

