

## The Creation of Multiple Standards with Common Data-Types

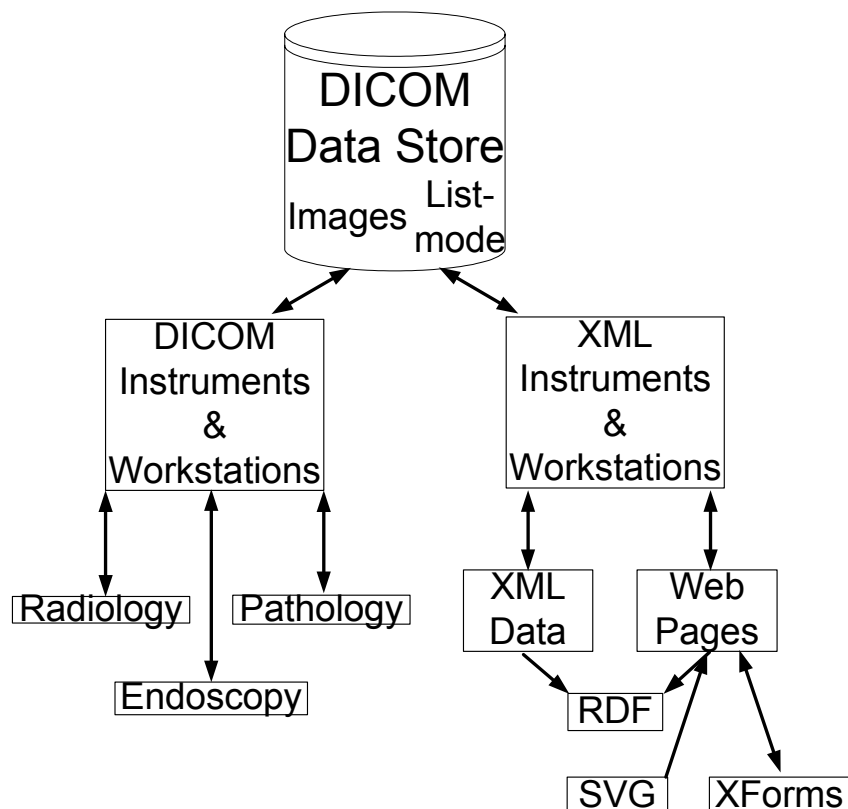
Robert C. Leif (in absentia)

Newport Instruments, San Diego, CA. rleif@rleif.com, www.newportinstruments.com

### Need for a New Standard that is Supported by Multiple Societies

- A clinical cytometry data standard, as opposed to a research data standard, has the extra requirement of exchanging data with hospital information systems.
  - The list-mode and image data from flow cytometry, digital microscopy including pathology images, new analysis techniques, and much of the related data that describes it (meta-data) must be integrated into the patient's clinical information and should be stored together.
- This integration of cytometry data with clinical information systems would be greatly facilitated by the use of a common data standard by the interested societies or groups.
  - However, since the pathologists plan to use DICOM and ISAC does not want to use DICOM, interoperability can still be achieved if all the groups use the same set of data-types and any new standard is based on the eXtensible Markup Language, XML.
- Present advances in digital microscopy and flow imaging will result in their use by the members of the Clinical Cytometry Society.
  - Since the differences between the software models of a digital microscope and a flow cytometer are minimal and both modalities are employed in the same laboratory for similar purposes, it is reasonable to apply a common data standard to these 2 modalities. This has already been demonstrated in the creation of a collection of XML schema (CytometryML) where the descriptions of both a flow cytometer and digital microscope were derived from a generic instrument (1).

### Relationships Between the Parts



- As shown in the figure, DICOM based instruments including digital microscopes will, as at present, create and transmit their results to the data store, which will then be able to exchange this data with the laboratory and hospital information systems.
- Flow cytometry XML and list-mode binary data can also be transferred and retrieved from the store. This provides the following benefits:
  1. The combination of image and list-mode data obtained with either a digital microscope or an imaging flow cytometer will be kept together.
  2. Laboratories that use both modalities will be able to store their data in the same place.

- 3. This will satisfy many of the pathologists and physicians from other specialities.
- The part of the data that is created by a human will be entered using an XML standard form, XForm (2).
- The analyzed data will be presented either using XML in the form of an office suite or as an XHTML based web page.
- Scalable vector graphics, SVG (3) is a portable web standard, which can be used to present cytometry data in the form of graphs.
- The relationships between the data will be codified with the resource description framework, RDF, (4).
  - A flexible tool like RDF is essential for research, which by its nature, must be free to change the characteristics of a test.
  - Since clinical data must be obtained using predefined tests, this use of RDF may be unnecessary. However, as described above, it can be very useful when creating a clinical test.

#### **CytometryML Schemas**

- The purpose of CytometryML schemas is to precisely specify data types that can be used to facilitate the transfer, storage, presentation, and creation of data, while minimizing the probability of mistakes that can occur during these processes. These uses include:
  1. Precisely describing in detail objects, such as: parts of a flow cytometer or microscope, slides, staining, images, and binary data that describes individual cells.
    - This description will provide the datatypes necessary to repeat a cyto- or histochemical measurement.
  2. Assisting in the design and creation of databases.
  3. Providing data in a form suitable for reports and forms.

#### **Partial Solution Based on Reuse**

- Reuse is a well known software engineering practice, which besides being applied to code, has been applied to many other parts of the development environment, such as designs, documentation, and tests.
- Standards paucity, the practice of reusing datatypes and their documentation from other standards, is an extension of reuse methodology to standards. This reuse minimizes the design effort, facilitates interoperability, reduces the paperwork for FDA approval, and maximizes the reliability of the CytometryML schemas due to the previous successful use of many of the data-types in implementations of DICOM and FCS.
- Because of the reuse of existing data-types, interconvertibility between CytometryML and pre-existing standards will be maximized.
- The extension of DICOM to include pathology imaging by Working Group 26 will result in the important benefit of having one set of semantics and terms for medical imaging that is used throughout the medical profession and can be reused by scientists engaged in Cytometry.
- CytometryML is in part an attempt to create a pilot implementation in XML schemas of the designs developed by Working Group 26 and a means to permit laboratories that cannot or do not wish to use DICOM to store their data in XML and enter data with XForms.
- The creation of new data-types was facilitated by extending and/or enhancing already existing DICOM data-types.

#### **METHODS**

1. A requirements document and a hazard analysis were published to acquire appropriate peer review.
2. The CytometryML schemas were developed using XML schema and were validated with both StylusStudio and XMLSpy. These XML schemas are primarily derived from DICOM data-types with documentation elements that included references to the descriptions of the data-types in the FCS & DICOM standards and data-types that have been created by DICOM Working Group 26.
3. A schema that describes flow cytometry data in greater detail than FCS was created. Since this is based on object-oriented design, it employs data-types imported from multiple relatively simple schemas.
4. XMLSpy was used to produce an example XML file (document) from the waveform (list-mode) element in the waveform schema. This waveform.xml page was then filled in with values and validated with XMLSpy.
5. A use case was created by filling the waveform.xml document with data from an existing flow measurement and the XML was validated against the waveform schema.

6. Visual inspection of the XML document was used to detect design defects including the order of the elements. The appropriate schema was corrected and the process was iterated.

### Brief Description of XML Syntax

An XML document that describes a flow experiment has been generated from the Waveform schema. The term Waveform is used instead of list-mode because the DICOM Waveform served as the model for the description of list-mode. The XML code at the beginning of the XML document is shown below and is color coded in the online version of this document (<http://www.newportinstruments.com/cytometryml/cytometryml.htm>). XML and the other XML languages are nested languages. The document below starts with items that describe the flow measurement in general. The **numbers** are not part of the XML document.

1 <Acquisition\_Date\_Time BTIM="\$BTIM" DATE="\$DATE">2006-09-17T09:30:47.0Z  
</Acquisition\_Date\_Time>

Element 1. (E1) **Acquisition\_Date\_Time** is the first element. It starts with a < character and is followed by two **attributes**, the FCS keywords BTIM and DATE. These **attributes** are separated from their **values** by an = character. Since these attribute values have been coded as constants in the schema, these have only been shown in the first **element**. The first part of the element is closed by a >. The value (17 September, 2006, 9:30 AM 30.47 seconds, (Coordinated Universal Time, UTC or GMT)) of the element, the DICOM date and time is shown between the first and second part of the element. The second part of the element starts with a </ and ends with a >.

2 <Modality>Flow</Modality>

E2 states that the data was obtained with a flow cytometer. The other modalities are: Sorter, Slide\_Image, and, Plate\_Image.

3 <Waveform\_Originality>Original</Waveform\_Originality>

E3 provides information whether the data was directly produced by the flow cytometer (Original) or was the result of a computer program (Derived).

4 <List\_Mode\_Sequence Start="1" End="250000">

E4 starts the description of the list-mode data by providing attributes that provide the values of the starting and ending point of the data structure that contains the binary list-mode data. The **List\_Mode\_Sequence** element contains other elements nested within that provide information pertaining to the list-mode binary data.

5 <List\_Mode\_Location>

file:///C:/ABSOLUTE-S/List2006-09-17T09:30.Data</List\_Mode\_Location>

E5 shows the URL of the file containing the list-mode data and thus serves to connect the file with the description of its contents and by its extension, Data, to its file type, which could be FCS 3.0, but preferably will be a new generation of FCS, where only the binary data is stored (5) or with a small amount of text data (6).

6 <Index\_File\_Info>

7 <Index\_File\_Location>

file:///C:/ABSOLUTE-S/Indx\_S2006-09-17T10:30.Indx</Index\_File\_Location>

8 <Indexing\_Parameters\_Name>S phase</Indexing\_Parameters\_Name>

</Index\_File\_Info>

</List\_Mode\_Sequence>

E6 through E8 describe an index file that contains a list of positions of the S phase cells in the list mode file. These S phase cells have previously been found by analysis. These index files can be used to directly locate the S phase or any other subset in the list-mode file. The use of these indices will eliminate the time required to read the subset field of the data structure that describes each cell. Instead, the data from only the cells that are members of a specific subset can be retrieved without looking through the entire list-mode file.

There is a potential problem of losing the relationship between the files when the XML page describing the data in the binary files, the list-mode binary file, and the index files are stored separately. This problem is easily solved by combining the files into a zip file, which is what Microsoft has already done with Office 2007 and other products.

### Other items that are based on the waveform schema

The rest of the items that complete the description of the image follow the order of the complete XML document and are described below. The CytometryML schemas and XML documents are available at <http://www.newportinstruments.com/cytometryml/cytometryml.htm>

## Items describing specific classes (schemas) present in the waveform

### Acquisition\_Context

The schemas created by the flowcyt group (<http://www.flowcyt.org/>) that describe the triggering of a flow cytometer and gating of the data have been reused by importing them into CytometryML.

### Channel\_Sequence (Parameter)

- A list of the channels (parameters or colors);
- The number of channels contained in the record describing the cell, which can be from one to 50 channels.

The description of each channel includes:

- Both a Short and Long Name

--For the monoclonal antibody described below (Reagent Info Example) these are respectively FI-Anti-5BrdU and **Fluorescein-Anti-5BrdU**. The **Short\_Name** is suitable to identify the axis of a graph and the **Long\_Name** provides a more complete description of the reagent.

- data on the Reagents used to perform the measurement. This data is described in the Reagent Info Example below;
- the excitation source (Arc, LED, laser) and its filters and collimator (condenser);
- the measurement (absorbance, fluorescence, polarization, etc.) the detector (CCD, PMT, array, etc.), and its optics (beam splitter (dichroic mirror), filter, spectrograph, etc.);
- the amplifier, transform (linear, log) and the data-type of the channel (parameter).
- The type of the data 8 or 16 bit integer or real.

### Flow cytometer

- Capacity for sorting cells;
- Platform (stand) direction (upright or inverted);
- Objective used
- Condensers used to collimate the lasers' emissions
- Detectors and emission filters

### General Information Schema

- For items that can be purchased, the following information can be included:
  - Manufacturer
  - Model name and number
  - Serial number
  - Web address of the manufacturer.

### Reagent Info Example

In the future, much of the constant information described above will be supplied by the manufacturer as an XML file, such as the one below, which describes the anti-5BrdU employed to label the S phase cells.

```
1 <stains:Reagent_Info Binding_Species="IgG" Binding_Species_Name="PRB1">
2   <stains:Label Label_Abbreviation ="FL" Label_Name="Fluorescein">
3     <stains:Reactive_Functionality Name =isothiocyanate Num= "mono"/>
   </stains:Label>
```

E1 is the beginning of a nested structure that describes the reagent. The beginning of E1 includes two attributes, which describe the monoclonal antibody. E2, which is located within E1, describes the label and includes in its beginning attributes that identify the label. E3 ends the description of the label and is followed by the second (closing) part of the label element.

```
4   <stains:Reagent_Formula_Wt>150000 </stains:Reagent_Formula_Wt>
5   <stains:Item_General_Info>
6     <item:Manufacturer>Phoenix Flow systems</item:Manufacturer>
7     <item:Model_Name>ABSOLUTE-S</item:Model_Name>
8     <item:Model_Number>AS1001 </item:Model_Number>
9     <item:Item_Lot-number>99</item:Item_Lot-number>
10    <item:URI_Var>http://www.phnxflow.com</item:URI_Var>
```

</stains:Item\_General\_Info>

E 5 to 10 provide the information about a reagent that could be used for reordering or inclusion in a paper.

11 <stains:Comment>My favorite antibody.</stains:Comment>

</stains:Reagent\_Info>

#### **A Single Standard for both Digital Imaging and Flow Analysis**

- Both flow cytometers and digital microscopes can produce each other's type of data. The Amnis® ImageStream (<http://www.amnis.com/>), which is a flow cytometer, produces images and the CompuCyte iColor™ Fluoro-Chromatic Imaging Cytometer (<http://www.compucyte.com/>), which is a laser scanning microscope, produces Flow Cytometry Standard list-mode files. Flow cytometry software is often employed to analyze list-mode data obtained with other digital microscopes.
- The Microscope and Flow Cytometer datatypes in CytometryML were both derived from a generic cytometer datatype.
- Since both instruments can be epi-illuminated for fluorescence, the minimum number of condensers is zero
- A flow cytometer has 1 objective. A microscope nosepiece very often has more than 1 objective.
- A microscope can have only one condenser; each of the light sources of a flow cytometer requires its own condenser.
- Objects on a solid support, slide, are imaged by microscopes; cells and particles in a flowing fluid are measured and/or imaged by flow cytometers.

#### **Results**

- A collection of schemas written in CytometryML has demonstrated the feasibility of reusing the semantics of datatypes from ISAC's FCS standard and from DICOM.
- The CytometryML schemas have been interfaced to the new gating and scaling schemas being developed by the Flowcyt ([www.flowcyt.org](http://www.flowcyt.org)).
- The creation of a collection of XML schemas, CytometryML, has demonstrated the feasibility of reusing the semantics of data-types from ISAC's FCS standard and those from DICOM, as well as reusing their documentation.
- CytometryML has been used to rapidly prototype a WG 26 design.
- The inclusion of constant attributes to link CytometryML to the legacy Flow Cytometry Standard (FCS) and the DICOM standard has been done.
- However this inclusion results in a small but acceptable increase in complexity of the code, the use of XML schema complexTypes.
- The essential unity of flow cytometers and digital microscopes has been demonstrated by deriving both data-types from a common cytometer data-type.

#### **Conclusion**

The Clinical Cytometry Society should join in the cytometry-pathology data standardization efforts by suggesting a representative to DICOM Working Group 26 and consider representation in the Flowcyt group and The Laboratory Digital Imaging Project, LDIP, Data Exchange Specification, LDIP.

In situations where multiple standards are being created, interoperability can be facilitated by employing a common set of datatypes.

The CytometryML schemas will be extended to include many of the new data-types that describe patients, specimens, etc. being prepared by DICOM WG 26 and other groups.

#### **ACKNOWLEDGMENTS**

I wish to thank Ryan Brinkman and Josef Spidlen for helpful discussions, the members of DICOM Standards Committee, Working Group 26 (Pathology) for sharing their designs, and LDIP for an education on RDF. Newport Instruments internal development funds have supported this project.

#### **Financial Disclosure**

Newport instruments is owned by Robert C. Leif, Ph.D and his partners. Our plan is to offer royalty free licenses to scientific and medical societies provided that they charge a reasonable amount to sublicense the schemas, take over the responsibility for their maintenance, that any sublicense does not include the GNU poison pill, prohibition on software patents, or other similar commercial limitations.

Copies of the XML files described above and the schemas used to generate it are available at <http://www.newport-instruments.com/cytometryml/cytometryml.htm>

### References

- 1 R.C. Leif, Development of an Intersociety Laboratory Flow & Imaging Data Exchange Standard, ISAC XXIII, Poster 139 (2006). ([http://www.newportinstruments.com/cytometryml/pdf/flow-imaging\\_std\\_isac2006.pdf](http://www.newportinstruments.com/cytometryml/pdf/flow-imaging_std_isac2006.pdf)).
2. XForms 1.0 (Second Edition) W3C Recommendation 14 March 2006 (<http://www.w3.org/TR/2006/REC-xforms-20060314/>).
3. Scalable Vector Graphics (SVG) 1.1 Specification, W3C Recommendation 14 January 2003 (<http://www.w3.org/TR/SVG11/>).
4. RDF Primer, W3C Recommendation 10 February 2004, (<http://www.w3.org/TR/rdf-primer/>).
5. R.C. Leif CytometryML, Binary Data Standards Manipulation and Analysis of Biomolecules, Cells, and Tissues II, SPIE Proc. Vol. 5699, pp. 325-333 (2005).
6. A Proposal for FCS4, <http://www.flowcyt.org/FCS4/>

### Appendix: Present Standards- Creation Efforts

Except for DICOM and the Flow Cytometry Standard (FCS), these efforts are all based on XML languages: XML Schema Design Language, (XSDL) and XML Resource Description Framework (RDF). Present efforts include:

The abbreviation of the XML languages is given in parentheses below

- The Laboratory Digital Imaging Project, LDIP, Data Exchange Specification (RDF) <http://www.ldip.org>
- DICOM Working Group 26, WG 26, <http://medical.nema.org/>
- International Society for Analytical Cytometry, ISAC, Data File Standard for Flow Cytometry, FCS, <http://www.isac-net.org>
- Health Level 7, HL 7, (XSDL), <http://www.hl7.org>
- Open Microscopy Environment, OME, (XSDL) <http://www.openmicroscopy.org>
- Flowcyt (XSDL & RDF), <http://www.flowcyt.org>
- Cytometry Markup language, CytometryML, (XSDL), <http://www.newportinstruments.com/cytometryml/cytometryml.htm>

### Interests

- Pathology images: LDIP and DICOM WG 26
- Measurements on large numbers of individual cells generated from flow and image cytometry: Standards created by members of the International Society for Analytical Society (ISAC).
- LDIP's interest is in "developing a set of definitions to clarify the uses of imaging in pathology and laboratory medicine."
- ISAC and OME share a common interest in the transfer of data.
- DICOM Working Group 26 plans to evaluate and extend the current DICOM standard as it relates to newer microscopic techniques including whole slide imaging. (I am trying to extend this to include flow and list-mode)
- CytometryML is creating a detailed description of the the datatypes and data structures necessary to repeat a flow or digital microscopy measurement.