Cytometry Standards Continuum

Robert C. Leif
XML_Med, a Division of Newport Instruments,
5648 Toyon Road, San Diego, CA, USA 92115-1022
E-mail: rleif@rleif.com; phone 1 619 582-0437
www.newportinstruments.com

1. Introduction:

Note: **This is an informal draft, which hopefully will mature into a publishable manuscript.**

Before a software standard is created, its use cases should be defined. After its presumed uses have been described, both a requirements document (1) and a hazard analysis (2) should be and in this case have been published to acquire appropriate peer review. However both documents still could use greater review and extension by others.

1.1. Uses

1) Data transmission

1). Data storage

1). Data entry

1). Data retrieval

1). Data mining (relationship)

1). Documents

1). Reports

1.1) Publications

1.2) Web sites

In the case of cytometry and many other areas of science and medicine all of the above uses can be handled by XML except for large binary objects, such as images and list-mode data (3). Since both humans and software share the common ability to make mistakes, the XML metadata should be validated. The use of defined datatypes also facilitates and increases the accuracy of data entry and software creation. The more that the software environment knows about an object; the less the programmer needs to describe. Abstraction and data hiding are two major software engineering techniques. XML schemas based on the XML Schema Definition Language, XSDL (4,5), have been employed in the cytometry markup language, CytometryML, to define many of the datatypes used in cytometry (1,6). Of greater significance, XSDL is used by Health Level 7, HL7, to define the datatypes for HL7 Version 3, which includes the infrastructure necessary to build a clinical messaging system (see below).

The International Society for Analytical Cytology, ISAC, has two standards, an established one for flow analyses (7,8) and an infrequently used other for digital microscopy (9). The Flow Informatics and Computational Cytometry Society (http://flowcyt.sourceforge.net/) has developed schemas that describe gating (10), transformation (11) and fluorescence compensation (12) and has proposed their use to ISAC. The Digital Imaging and Communications in Medicine (DICOM) Working Group 26 is developing Supplement 122, which includes pathology specimen identification and revised pathology storage classes. In the present preliminary version of Supplement 122, the specimen is a “discrete physical object that is the subject of pathology examination”. It can be an organ part, block with embedded tissue, or a slide with something (sections, cells, etc.) on it. The

physical characteristics of microscope slides and coverslips have previously been described in DICOM. The harvesting procedures, conservative conditions (fixation, freezing, etc.), organ and location within the organ, procedure step (sectioning, hisopathological or cytological examination, fixation, and staining are also to be included in Supplement 122. Tissue microarrays are presently also included in Supplement 122. Whole slide imaging is an important new technology, which is being introduced into DICOM with this supplement. Accession numbers and containers are also included. DICOM already includes key complex datatypes like patient and physician. Presently in DICOM, the number of colors (measurements) is being extended to be greater than 3. Neither list-mode nor flow cytometry are mentioned in DICOM.

This work is being performed in conjunction with the Anatomic Pathology Special Interest Group of HL7 (http://www.hl7.org/), which includes in its Charter the following statement:

"Formal Relationships with Groups Outside of HL7"

> "This SIG acknowledges and cooperates with the DICOM WG 26 (Pathology Imaging), includes representation from the existing College of American Pathologists and will seek participations from specialty societies in the field of anatomic pathologist, surgeons, cancer researchers, and other users of surgical pathology reports. These liaisons will be with the approval of the HL7 Board in accordance with Policy and Procedures."

Since none of the directly or loosely related analytical cytology groups that are trying to create standards have the capacity or interest and probably do not have the required domain knowledge to create the datatypes that have or are being created by the DICOM-HL7 cooperative, it will greatly increase our chances of success if we limit our works to those areas where we have special expertise and reuse (13) or better yet interoperate with the products of the DICOM-HL7 cooperative.

2. One Standard is better than two.

The overlap between image and flow cytometry is sufficient to warrant the use of a common standard for both. The amnis ImageStream® flow cytometer acquires images and many digital microscopes store data in list-mode. A generic instrument schema, which contained an Instrument_Type has been described (14). Both a Microscope_Type and an Flow_Cytometer_Type were created by restriction from the generic Instrument_Type (Figure 1).

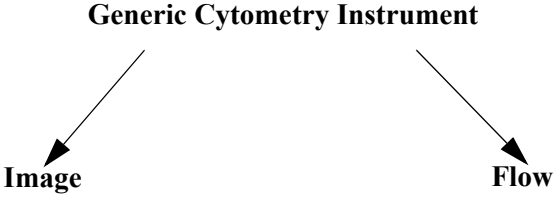

Figure 1. Derivation of image and flow cytometers from a generic cytometry instrument

A combined standard has the following advantages:

1) Since many datatypes can be used for both modalities, the amount of effort it takes to develop, maintain, understand, and extend a combined standard will be considerably less than for two separate standards.

2) The development risk is minimal, since it has already been demonstrated (14) that it can be done.

3 A single standard can be employed for instruments that produce both image and list-mode data.

4 This combination of image and list-mode data will be stored together.

5) Laboratories that use both flow cytometers and digital microscopes will be able to analyze their data using the same tools, which will reduce software development and user education costs.

The following similarities exist between multidimensional list-mode and image data:

1) Multidimensional list-mode data is viewed as an image.

2) Gating of list-mode data can be thought of as placing an overlay plane on top of an image.

3) In principle, clustering algorithms for list-mode data can use morphological operators to isolate populations. Subtraction of control populations from experimentals can be performed in a similar manner to image processing operations.

4) The separation of cells present in an image often employs the same parameters that are used for list-mode data.

5) The analyses and visualization of the individual populations of cells found in an image is often identical to that performed in list-mode.

6) The parameters that are the content of the records (structs) that describe individual cells can be displayed along with their images.

3. HL7

The XML version of HL7 is version 3 -- Reference information model -- Release 1 Standard. It is a joint international standard with ISO, ISO/HL7 21731:2006Health informatics. This means that one needs to pay for the standard document or be a member of HL7. Since the standard is continuously being extended and modified, the necessary information is available at:

http://www.hl7.org/v3ballot/html/welcome/downloads/downloads.htm

The instructions recommend downloading the full site because that will maintain the hyperlinks.

This download of the full 257 Meg ZIP file includes all materials one will find in this Version 3 Ballot site.

3.1 Navigation through the HL7 Version 3 Standard ballot documentation

After you have downloaded and unzipped the ballot, go to the directory that contains the unzipped files and open /html/welcome/environment/index.htm. There is a large Table of Contents in the center, which should, at least temporarily be ignored. At the upper left will appear the menu below:

Ver3_Ballot/html/welcome/environment/index.htm

HL 7 Version 3 Standard

- Introduction
- Foundation
- Specification Infrastructure

- Implementation Technology Specifications
- Services
- Universal Domains
- GB Domains
- Background Documents
- Support Files
- Known Issues

Since HL7 has its own large group of abbreviations, the first step is to go to the Glossary. All path steps in the HL7 menu tree will be separated by slashes "/". Goto Universal Domains.Common Message Elements Types/ Glossary. In the glossary, you will find that CMETS means Common Message Elements Types and you will not find the item directly below, GB Domains; however, GB probably means Great Britain. A good way to navigate the HL7 documentation is visit the Universal Domains, which will then lead you to the schemas. Since the schemas have unintelligible names, going to them first will be very frustrating. The path for CMETS and then schemas is: Universal Domains/Common Message Elements Types/CMET Definitions for all domains. It should be noted that no guarantee can be given that this path will not be changed in a subsequent version of the ballot. For each datatype, there is a thumbnail image of an object diagram below the top center and an icon containing text at the bottom right. Activation of the thumbnail image reveals an object model; and activation of the icon containing text leads to the schema that describes the object model.

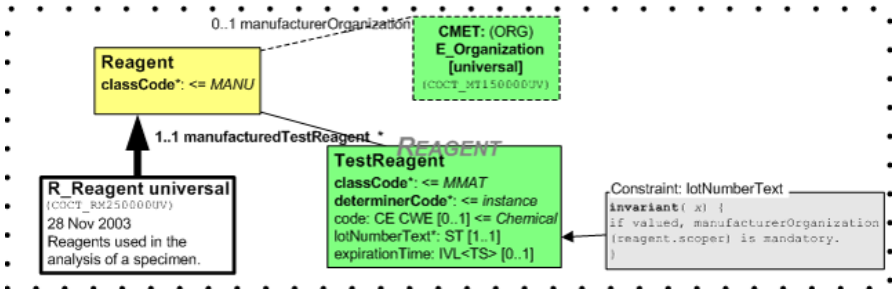

Figure 2. Object Diagram of HL7 R_Reagent universal data-type. This figure was copied from the HL7 Ballot and it and the schema and other software generated from it are the property of HL7.

The schema for R_Reagent universal is named by a code, COCT_RM250000UV03.xsd, which is partially based on the name of the HL7 committee that is responsible for the data-type; instead of the name of the datatype, R_Reagent universal, or preferably something like Reagent_Universal. This schema has the same targetNamespace as apparently all of the other HL7 schema, urn:hl7-org:v3, and it makes directly visible the supporting datatypes provided by other schemas by use of an include element. This eliminates the use of prefixes that point to the source schemas of these supporting datatypes and thus eliminates the validation of the sources of the individual datatypes by the schema parser. It also makes it extremely difficult for a human to determine which schema provided a data-type. The combined use of multiple targetNamespaces and import elements eliminates these problems (3,15), which increases maintainability and reuse. The HL7 schemas are, at present, automatically generated from the object model; the present problems can be changed by reconfiguring the schema generator. A request to do this has already been submitted to HL7.

The present, HL7 basic foundation schema, Voc.xsd, is not object oriented. Voc.xsd contains 19,366 lines of code and simpleTypes are only described. These datatypes range from RealmOfUse (Country) to UnitsOfMeasureCaseInsensitive. There is an extreme overuse of abbreviations, which includes providing different meanings to a two letter abbreviation based on the case of the letters. These abbreviations include letter pairs such

as: bl, BL, bn, ED, st, ST, cs, CD, CE, CV, CS, CO, CR, SC, II, ts, TS, and ADXP. It would be possible to import Voc.xsd into new schemas, which each is devoted to one class and create one or more complexTypes that describe that class. Similarly it will be possible to provide individual abbreviations, when they are included in the appropriate class, with meaningful names. A request to remedy this situation has already been submitted to HL7.

4. Suggested organization for the development of a standard(s)

The only modality that will permit reuse of the HL7 Version 3 is XML schemas, XSDL. For those, who wish to eventually use the Resource Description Framework, RDFa (16,17) apparently supports XSDL simpleTypes that are embedded in XHTML documents and should eventually be able to use XPath statements to work with XSDL complexTypes.

The DICOM-HL7 infrastructure is essential to provide a continuous group of standards, since it includes diverse items, such as billing, specimen description, people, microscope slides, images, and the use of SOAP to send messages. CytometryML provides an infrastructure that describes the instrument and will describe the staining of the specimen. It is planned to have CytometryML include interface schemas that will permit other groups to reuse HL7 datatypes. This way the Data and Image Analysis Special Interest Group (D&IA SIG) of the Society for Biomolecular Sciences could extend the DICOM-HL7 image model and the D&IA SIG could collaborate with the Flow Informatics and Computational Cytometry Society. The Flow Informatics and Computational Cytometry Society would continue to do the specialized flow informatics; however, it would do so in a manner that permits its XML software to interoperate with HL7 and the D&IA SIG.

CytometryML will be augmented by continuing to translate DICOM into XML schemas and working with HL7 to derive the CytometryML utility schemas, such as num_types and units from HL7.

In terms of data exchange, RAW or close to a RAW format for the image and list-mode files should be used. These binary files should be combined with XML based on XML schema for the other information. The combination could then be zipped together in the same manner as Microsoft does with VISTA and Office 2007. This suggestion is not new; it is the modern, logical extension of the original “Proposed Standard for Image Cytometry Data Files” (9).

5. Conclusion

A new combined standard should be based, as much as possible, on existing standards including HL7 and DICOM. The XSDL schemas of CytometryML (http://www.newportinstruments.com/cytometryml/cytometryml.htm), the XSDL schemas of the Flow Informatics and Computational Cytometry Society (http://flowcyt.sourceforge.net/), and the image specific work of Data and Image Analysis Special Interest Group (D&IA SIG) of the Society for Biomolecular Sciences should all be harmonized with each other and HL7.

REFERENCES

1. 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).

2. R. C. Leif and S. B. Leif, "Evolution of Flow Cytometry Standard, FCS3.0, into a DICOM-Compatible For-

mat". in *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).

3. R.C. Leif, "CytometryML, Binary Data Standards", in *Manipulation and Analysis of Biomolecules, Cells, and Tissues II*, D. Farkas, D. V. Nicolau, and R. C. Leif, Editors, SPIE Proc. Vol. 5699, pp. 325-333 (2005).

4. "XML Schema Part 2: Datatypes Second Edition", W3C Recommendation 28 October 2004 (http://www.w3.org/TR/2004/REC-xmlschema-2-20041028/).

5. Pricilla Warmsley, "Definitive XML Schema, Prentice Hall", http://www.phptr.com (2002).

6. 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).

7. 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.

8. "FCS, Flow Cytometry Standard", http://www.isac-net.org/ Then search for FCS.

9. P. Dean, L. Mascio, D. Ow, D. Sudar, J. Mullikin, Proposed Standard for Image Cytometry Data Files, Cytometry, Vol. 11, pp. 561-569 (1990).

10. Data Standards Task Force, Bioinformatics Standards for Flow Cytometry Consortium: "Proposal for International Society for Analytical Cytology (ISAC), Gating-ML: Draft Standard for Gating in Flow Cytometry, version 1.1"; http://flowcyt.sourceforge.net/gating/ (2006)

11. Data Standards Task Force, Bioinformatics Standards for Flow Cytometry Consortium: "Proposal for International Society for Analytical Cytology (ISAC), Transformation-ML: Draft Standard for Transformation Description in Flow Cytometry, version 1.0"; http://flowcyt.sourceforge.net/transformation/ (2006).

12. Data Standards Task Force, Bioinformatics Standards for Flow Cytometry Consortium:" Proposal for International Society for Analytical Cytology (ISAC), Compensation-ML: Draft Standard for Compensation Description in Flow Cytometry, version 1.0"; http://flowcyt.sourceforge.net/compensation/ (2006).

13. B. Boehm, K. Sullivan, "Software Economics: A Roadmap", in *International Conference on Software Engineering, Proceedings of the Conference on The Future of Software Engineering*, ACM Press New York, NY, USA. (2000).

14. R. C. Leif, CytometryML: a data standard which has been designed to interface with other standards, in Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues V, D. L. Farkas, R. C. Leif, D. V. Nicolau, Editors, SPIE Proceedings Vol. 6441, pp. 64410P 1-11 (2007).

15. R. C. Leif, "CytometryML and Other Data Formats", in *Manipulation and Analysis of Biomolecules, Cells, and Tissues III*, D. Farkas, D. V. Nicolau, and R. C. Leif, Editors, SPIE Proceeding Vol. 6088-0L pp. 1-7 (2006).

16. B. Adida and M. Birbeck, RDFa Primer 1.0, Embedding RDF in XHTML, W3C Working Draft 12 March 2007 (http://www.w3.org/TR/xhtml-rdfa-primer/).

17. M. Birbeck, S. McCarron, XHTML RDFa Modules, Modules to support RDF annotation of elements, W3C Editor's Draft 2 April 2007 (http://www.w3.org/MarkUp/2007/ED-xhtml-rdfa-20070402/)