Cytometry metadata in XML

Robert C. Leif[a*], Stephanie H. Leif[a] [a]XML_Med, a Division of Newport Instruments,

3345 Hopi Place, San Diego, CA, USA 92117-3516

ABSTRACT

Introduction: The International Society for Advancement of Cytometry (ISAC) has created a standard for the Minimum Information about a Flow Cytometry Experiment (MIFlowCyt 1.0). CytometryML will serve as a common metadata standard for flow and image cytometry (digital microscopy). Methods: The MIFlowCyt data-types were created, as is the rest of CytometryML, in the XML Schema Definition Language (XSD1.1). The datatypes are primarily based on the Flow Cytometry and the Digital Imaging and Communication (DICOM) standards. A small section of the code was formatted with standard HTML formatting elements (p, h1, h2, etc.). Results:1) The part of MIFlowCyt that describes the Experimental Overview including the specimen and substantial parts of several other major elements has been implemented as CytometryML XML schemas (www.cytometryml.org). 2) The feasibility of using MIFlowCyt to provide the combination of an overview, table of contents, and/or an index of a scientific paper or a report has been demonstrated. Previously, a sample electronic publication, EPUB, was created that could contain both MIFlowCyt metadata as well as the binary data. Conclusions: The use of CytometryML technology together with XHTML5 and CSS permits the metadata to be directly formatted and together with the binary data to be stored in an EPUB container. This will facilitate: formatting, data- mining, presentation, data verification, and inclusion in structured research, clinical, and regulatory documents, as well as demonstrate a publication's adherence to the MIFlowCyt standard, promote interoperability and should also result in the textual and numeric data being published using web technology without any change in composition.

Keywords:

1. Introduction

1.1. What is MIFlowCyt?

Presently, it is a file or a short appendix to a scientific article, such as one published in Cytometry. MIFlowCyt[1],[2] provides a concise description of the technology and findings present in the article. MIFlowCyt does not require sufficient information to repeat the experiment, such as would be provided by a well written Materials and Method section of a paper and/or supplementary materials. The information contained in a MIFlowCyt document can be textural and/or structured.

Problem: How does one verify that a MIFlowCyt section of a manuscript is sufficiently complete for publication? The authors believe that this presently is a highly subjective judgement. In the case of most of the CytometryML major schemas, including those of the CytometryML version of MIFlowCyt, an XML page was generated from elements, particularly, the major element of the individual schemas. The element and attribute values for this page were then manually entered. As will be demonstrated below, since these pages were not formatted, they were difficult to read and check.

Since the reference example is Blimkie et al.[3], a good test of the CytometryML implementation of MIFlowCyt is, do the X ML pages generated from the CytometryML schemas validate with the values present in Blimkie et al? Since the text of Blimkie et al. is 9 pages of 12 point type, it is a considerable duplicative part of a Cytometry paper. An important question is, if there is a discrepancy, is the text of the paper or is the MIFlowCyt section nominative?

The present progress in producing an XML implementation of MIFlowCyt is best described in terms of the Miflowcyt_Info_Type, which is the major datatype declaration (Code Fragment 1) of the miflowcyt schema. Miflowcyt_Info_Type consists of the major elements of MIFlowCyt, which are contained in the MIFlowCyt_Info element.

Code Fragment 1, Description of the CytometryML MIFlowCyt Element

```
1<complexType name="Miflowcyt_Info_Type" mixed="true">
2    <sequence>
```

```
3       <element name="Experiment_Overview"
           type="exper_overview:Experiment_Overview_Type"/>
4       <element name="Sample_Info" type="sample:Sample_Info_Type"/>
5       <element name="Sample_Treatment_Description"
6          type="protocol:Sample_Treatment_Description_Type"/>
7       <choice>
8          <element name="Flow_Series_and_Instance_Info"
              type="flow:Flow_Series_and_Instance_Info_Type"/>
9          <element name="Microscope_Series_and_Instance_Info"
              type="micro:Microscope_Series_and_Instance_Info_Type"/>
        </choice>
    </sequence>
</complexType>
```

Element 1 is a complexType because it is composed of multiple element-datatype pairs (lines 2-9). It is mixed because XML pages derived from it permit the interspersion of text. Although, the Miflowcyt_Info_Type is part of a schema that defines one data type, the Miflowcyt schema imports datatypes from 7 other schemas, which each have schemas to import. Thus, the Miflowcyt schema is the apex of a tree of schemas. In keeping with object-oriented design, each of the schemas referenced by the elements above describes a single major class. The schema references (prefixes) are words that are separated by a colon in front of the type name. The choice element is included because to distinguish whether the movement of the cells in the sample was by fluid transport or movement of a slide.

2. MIFlowCyt with XHTML5 Elements

2.1 Software Development Approach

Since an element that describes the properties of a fluorescent dye is currently being developed by the International Society for Advancement of Cytometry, ISAC, Data Standards Task Force, DSTF, it was decided to follow the example of the ISAC MIFlowCyt Standard document[1] and to show the data in the form of a table. The first example is a visualization of the components of a fluorochrome_Info element.

The most familiar language for formatting web documents is html-xhtml. The latest version of html is html5.1[4]. Fortunately XHTML5 has an XML schema[5]. This approach has a serious problem in that the part of the html community that is creating the new html5 standard has only recently shown significant interest for xhtml5[4]. The following line of code because of the presence of the xhtml <td> element being applied to a CytometryML data element to place it in a cell within a table, will not validate with a standard XML parser.

```
<td><fluor:Common_Name>EuQuantum Dye</fluor:Common_Name></td>
```

The strong typing of XML schemas precludes an element of one schema from containing an element from another schema except when one schema is part of the data structure contained in the other schema. This requirement for importing the element of one schema into an element that is within another contradicts and interferes the separation of the markup language from the data language.

3. Results

3.1 Simple Solution

The following is a work in progress! One solution to the combination and storage of a heterogenous collection of files is to create the text pages as formatted XHTML 5 files and to organize and store them together with others in an EPUB[6],[7]. There is a major problem with this solution: XHTML 5 was not written in an XML schema language[4]. A reasonable solution was achieved by finding a XML schema for XHTML 5, fortunately one was created by Olivier Ishacian[5]. The first solution

tried was the use of an addition to the XML Schema Definition Language (XSD), which occurred as part of the creation of XSD1.1. This addition, the openContent element, appeared to be the solution.

Unfortunately, it presently does not work. It permits elements from different schemas to be sequentially interleaved; however, they cannot be interspersed.

```
<openContent mode="interleave">
  <  any namespace="##other" processContents="lax"/>
</openContent>
```

and `processContents`="`skip`" do not work

The parser sees an unknown element between the two parts of an XHTML or XML element and rejects the construct. An "ignore" value should be added to the processContents attribute. The ignore value would permit the interspersion of elements from other schemas by making the parser blind to the contents of the enclosed element(s).

3.2 Alternative Solution

Radu Coravu of Syncro Soft SRL, the manufacturer of oXygen (www.oxygenxml.com), which is the editor tool used to create the CytometryML schemas, suggested the use of a meta-schema language control language, NVDL, which is formally referred to as: ISO/IEC 19757-4 NVDL (Namespace-based Validation Dispatching Language)[8]. · According to the website[9], NVDL is Part 4 of ISO/IEC 19757 DSDL (Document Schema Definition Languages and has an XML syntax.The NVDL independently validates elements in an XML or XHTML page against their parent schema. This permits an XHTML5 phrasingContentElement such as: <p>, <h1><h2>or<td> to surround elements that have been derived from other schemas. Thus, the parsing of the XHTML5 elements ignores the data containing elements from CytometryML and the parsing of the CytometryML data containing elements data ignores XHTML5 phrasingContentElements. This design requires that the web page be validated by including a separate NVDL page. Standard direct validation of the XML or XHTML pages shows errors at the XHTML phrasingContentElements.

The first element to be marked up with phrasingContentElements is Fluorochrome_Info, which is based upon a tentative design of one of the authors (RCL). The data in the Fluorochrome_Info element is described by the elements of the Fluorochrome_Info_Type, The goal of this software study is to present the data in the Fluorochrome_Info element as a table.

The first step, as described previously, is to prepare an XSD1.1 schema. Code Fragment 2 shows the Fluorochrome_Info_Type

3.3 Procedure

An xsd1.1 schema, fluorochrome.xsd, which contains the Fluorochrome_Info element (Code Fragment 2), was created. The Fluorochrome_Info element (line1) is based upon the Fluorochrome_Info_Type (line 2).

Code Fragment 2, Fluorochrome_Info Element

```
1)<element name="Fluorochrome_Info" type="fluor:Fluorochrome_Info_Type"/>
2)<complexType name="Fluorochrome_Info_Type" mixed="true">
3)  <sequence>
4)    <element name="Caption" type="strings:Bounded_256_Type"/>
5)    <element name="Common_Name" type="fluor:Name_Type"
        minOccurs="0" maxOccurs="1"/>
6)    <element name="Abbreviation"
```

```
        type="fluor:Abbreviation_Type"/>
7)    <element name="Excitation_Max"
        type="fluor:Wavelength_Type"/>
10)     <element name="Emission_Max"
        type="fluor:Wavelength_Type"/>
8)    <element name="Vendor_or_Source_URI" type="anyURI"
        minOccurs="0"/>
9)    <element name="CAS_Num_or_Other_Value"
        type="fluor:CAS_Num_or_Other_Value_Type"/>
   </sequence>
</complexType>
```

Fluorochrome_Info_Type is to serve as a description of the minimum information about a fluorochrome. The sequence, element 3, is a record that starts with a caption (element 4) of up to 256 characters. This is followed by the common name (element 5)[10], which is followed by an abbreviation (element 6). This is followed by the Excitation_Max (element 7) and the Emission_Max (element 8). Both of which are in wavelengths. The URI (element 9) comes next. The URI was used because most scientific products are purchased via the Internet and the physical location of the source is of minimal interest to investigators living in another part of the world. The MinOccurs="0" is for the odd case where the vendor does not sell through the Internet. Although MIFlowCyt is supposed to provide the minimum information about a flow cytometry experiment, the schema that controls the available datatypes can provide more than the minimum by the simple expedient of providing a minOccurs attribute with a value of zero. This permits datatypes of elements that are thought to be relevant but not mandatory to be in the standard without being required. The CAS_Num_or_Other_Value (element 10) is of multiple types because this number should provide a reliable key to a database or similar data storage element. Unfortunately not all fluorochromes have CAS (Chemical Abstracts Service) numbers. Other possible temporary types of values possibly include: an ISAC_Num_Type, Vendor_Num, Other_Value, NA, More_Info_Needed, and No_Name_Found.

The Analyte_Reporter is the labeled species, which consists of a macromolecule, label and a coupling species that covalently joins the other two. The labels include: fluorochromes, phosphors, or isotopes. The software technology presented here could be used to describe analyte reporters. However in most cases, the listing of the three components including the information that describes each would result in an excessively wide table.

Code Fragment 3 shows part of the content of that XML page, which was generated by oXygen from element 1 of Code Fragment 2, Fluorochrome_Info. Each of the elements shown in the schema shown in Code Fragment 1 maps to an element in the XML page of which two cells (elements 5 and 6) are shown in Code Fragment 3.

Code Fragment 3, two Table Cells from an XML page

```
<tr>
   <td><fluor:Common_Name>EuQuantum Dye</fluor:Common_Name></td>
   <td><fluor:Abbreviation>EuMac</fluor:Abbreviation></td>
```

Code Fragment 3 shows the first 2 data elements of the XML page that was generated from Code Fragment 2. Subsequently, the XHTML phrasingContentElements <tr>, which signifies the beginning of a row and <td>, which indicates the presence of a data cell, were added. The information is located between the beginning and end parts of the elements. One can read the information containing text: EuQuantum Dye and EuMac; however, it is a bit of a strain and there is a low concentration of relevant information. A requirement of this work is that the XML page generated from Fluorochrome_Info be suitable for inclusion in the MIFlowCyt section of a publication.

TABLE 1

Tabular Output of a Page based on the Fluor Schema

Common Name	Abbreviation	Excitation Max	Emission Max	Source URI	ACS CAS# or Other#
EuQuantum Dye	EuMac	365	619	http://www.newportinstruments.com/	Newp00-1 /

Table 1 is much easier to read than an XML page. Since there was a possibility of a problem with copyright(s), Quantum Dye®, which is owned by Newport Instruments, was used for the example. This table presently is approaching being suitable for professional publication. According to W3C, "CSS, the cascading style sheet language, is applicable to XML as it is to HTML." (http://www.w3.org/XML/1999/XML-in-10-points). In principle, XHTML5 documents should work with today's browsers. Presently the World Wide Web Consortium (W3C) states that Formatting XML with CSS is not recommended. Use JavaScript or XSLT (XSL Transformations (XSLT)[11],[12] instead. (http://www.w3schools.com/xml/xml_display.asp). However, CSS has the great advantage of being WYSIWYG (What You See Is What You Get), which is the way commercial word processors work. Changes to the text are made virtually instantaneously.

This html tabular formatting approach can achieve a similar presentation to that provided by the MAGE-TAB standard, which is used with micro titer plates. The creators of MAGE-TAB have "propose(d) a simple spreadsheet-based[13] format for representing primary data and experimental details (metadata) from microarray investigations. We refer to this format as MAGE-TAB (MicroArray Gene Expression Tabular)" They use a commercially available spreadsheet (Microsoft® Excel) as their mode of implementation. The authors explained their abandonment of XML technology. "However, the complexity of the MAGE-ML[14] format has made its use impractical for laboratories lacking dedicated bioinformatics support." As shown in TABLE 1, it is possible to create equivalent tables with XML technology and with use of xhtml5 technology to place XML defined elements in the cells of a table.

3.4.0 Present Status of the CytometyML Implementation of MIFlowCytProject

1. The software has been tested to work with Chemical Abstracts Service CAS number of one dye and with strings of numbers that are out of scope for the CAS specification. The standard XML pages validated as expected; however, when the html table formatting elements were added The page would not validate. When these pages were tested with the NVDL file present, they validated. They did not validate when the value of the CAS number was out of the range of the CAS datatype. More extensive testing will be required to determine if this approach is adequate for the creation of medical grade software.
2. A significant part of MIFlowCyt, Code Fragment 1, has been coded in XSD1.1. MIFlowCyt.xml is relatively large with all coded elements present, validates and contains 882 lines of code. This would be appropriate for a structured document, where the materials and methods was primarily composed of the MIFlowCyt elements. When only the required elements were included, it validates with 432 lines of code, which would be suitable for the present synopsis.
3. MIFlowCyt.xml is based upon a tree of approximately 77 schemas. The next step in this project is to finish the Sample_Treatment_Description_Type.
4. The feasibility of using MIFlowCyt to provide the combination of an overview and index to a scientific paper or a report has been strongly suggested.
5. The feasibility of a combination of XHTML5 with XML schema elements has been demonstrated. The use of CSS with these XML web pages is sufficient to be able to monitor the design of a schema from the XML produced from the schema. Unfortunately, these XML web pages are not yet of publication quality.

4. CONCLUSIONS

Maintaining MIFlowCyt as a single schema and associated web page would be very difficult. The CytometryML schemas are highly modularized and each one describes an object or class. This permits reuse and reorganization. Since multiple XHTML5 pages can be combined into one EPUB[6], a potential open, efficient, searchable scientific format, which can be developed without any need for ISAC to support its own standard. A complete publication framework is now possible, where the web

pages generated from the CytometryML MIFlowCyt schemas and the XHTML5 schema can be packaged in an EPUB[6] as part of a patient record, journal article, or medical report. Previously, it has been established[6] that an EPUB can contain and organize multiple maintainable components including images. The web pages in the EPUB can be formatted employing web technology, CSS or XSL (EXtensible Stylesheet Language[11],[12]).

MIFlowCyt can serve as a table of contents, ToC, with hypertext linkages to the detailed descriptions, which will be based on web pages generated from the CytometryML schemas. This material can be contained in the paper or supplementarymaterial.

Interested Parties: reviewers, editors, government agencies (FDA) can automatically check for completeness in answering their requirements by validating the ToC against its schema(s).

Changing an external cascading style sheet (CSS[15]) file permits reuse including: archival, electronic publishing, print publishing, government reports and grant applications.

Style consistency, where the editor and publisher control the style of the document, permits verified publishable manuscripts to be submitted by the authors, which will reduce publishing costs!

4.1 Software Implications

The combination of XSD schemas, an XHTML5 XSD schema and CSS or the Extensible Stylesheet Language Family (XSL) should act as a graphical programming interface, which could permit the development of a What you see is what you get (WISIWIG) products. these products could compete with present commercial products and be portable between operating systems. XHTML5 includes:[16] forms, SVG, MathML, and Paged Media[17].This portable technology should facilitate medical and other high-technology informatics

The XHTML5 schema could be augmented with DITA and/or DocBook elements, which should facilitate the preparation of structured documents including forms. These structured documents could also be stored together as an EPUB, which is effectively an open, portable standard and thus is competitive with PDF.

5. Acknowledgment

This work was sponsored by Newport Instruments internal research funds. One of us, R.C.L. wishes to thank Ryan R. Brinkman, Josef Spidlen and the other members of the ISAC DSTF for many enlightening and pleasant discussions and particularly Kim Blenman who performed the work on obtaining and preparing CAS numbers. Radu Coravu of Syncro Soft SRL, the manufacturer of oXygen (www.oxygenxml.com), suggested and facilitated the use of a meta-schema language control language, NVDL. Olivier Ishacian and his company, Pixware SARL, for making the XHTML5 XSD schema available to the authors and the rest of the public.

6. References

[1] Lee, J.A., Spidlen, J., Scheuermann, R., Brinkman, R., Editors,.MIFlowCyt Standard - ISAC Recommendation MIFlowCyt 1.0 - a standard for outlining the minimum information required to report the experimental details of flow cytometry experiments, http://isac-net.org/Resources-for-Cytometrists/Data-Standards/MIFlowCyt.aspx (2008)

[2] Lee, J.A., Spidlen, J., Boyce, K., Cai, J., Crosbie, N., Dalphin, M., Furlong, J., Gasparetto, M., Goldberg, M., Goralczyk, E.M., Hyun, B., Jansen, K., Kollmann, T., Kong, M., Leif, R.C., McWeeney, S., Moloshok, T.D., Moore, W., Nolan, G., Nolan, J., Nikolich-Zugich, J., Parrish, D., Purcell, B., Qian, Y., Selvaraj, B., Smith, C., Tchuvatkina, O., Wertheimer, A., Wilkinson, P., Wilson. C., Wood, J., Zigon, R.; International Society for Advancement of Cytometry Data Standards Task Force, Scheuermann RH, Brinkman RR, "MIFlowCyt: the minimum information about a Flow Cytometry Experiment," Cytometry Part A. 73A, pp. 926-930 (2008).

[3] Blimkie, D., Fortuno III, E.S., Thommai, F., Xu, L., Fernandes, E., Crabtree, J., Rein-Weston, A.,Jansen, K., Wilson, C.B., Brinkman, R.R., Kollmann, T.R., "Identification of B cells through negative gating—an example of the MIFlowCyt standard applied", Cytometry Part A; (2010): 77A:546–551 Supplement

[4] “HTML 5.1, W3C Working Draft 08 October (2015),” (accessed 10 December 2015) [http://www.w3.org/TR/2015/WD-html51-20151008/]

[5] Ishacian, O., “XHTML5, Pixware SARL. See . HTML5 schema , Copyright (c) (2012) Pixware SARL,” (accessed 10 December 2015) [http://grepcode.com/file/repo1.maven.org/maven2/com.googlecode.l10n-maven-plugin/l10n-maven-plugin/1.5/xhtml5.xsd]

[6] Leif R.C., and Leif S.H., "A shared standard for cytometry and pathology," Paper No.: 8587-50, Tracking No.: PW13B-BO400-86, Symposium: PW13B SPIE BiOS, Conference: Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XI 2013 (2013).

[7] International Digital Publishing Forum (IDPF) “EPUB Open Container Format (OCF) 3.0.1, Recommended Specification (26 June 2014) (accessed 10 December 2015) [http://www.idpf.org/epub/301/spec/epub-ocf.html]

[8] “NVDL, ISO/IEC 19757-4 NVDL (Namespace-based Validation Dispatching Language) information is available at”, (accessed 10 December 2015) [http://www.nvdl.org/]

[9] ISO/IEC JTC 1/SC 34, “ISO/IEC 19757 - (DSDL) Document Schema Definition Languages,” (accessed 10 December 2015) [http://www.dsdl.org

[10] “IUPAC - International Union of Pure and Applied Chemistry: 4. Nomenclature” (accessed 10 December 2015) (accessed 10 December 2015) [http://www.iupac.org/nc/home/publications/technical-reports/guidelines-for-drafting-reports/4-nomencl.html?sword_list%5B%5D=name]

[11] W3C.org, “The Extensible Stylesheet Language Family (XSL)” (accessed 10 December 2015) [http://www.w3.org/Style/XSL/)]

[12] W3C.org XSL Languages, (accessed 10 December 2015) [http://www.w3schools.com/xsl/xsl_languages.asp]

[13] Rayner, T.F., Rocca-Serra, P., Spellman, P.T., Causton, H.C., Farne, A., Holloway, E., Irizarry, R.A.S., Liu, J., Maier, D.S., Miller, M., Petersen, K., Quackenbush, J., Sherlock, G., Stoeckert Jr, C.J., White, J., Whetzel, P.L., Wymore, F., Parkinson, H., Sarkans, U., Ball, C.A., Brazma, A., “A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB” BMC Bioinformatics 2006, 7:489 doi:10.1186/1471-2105-7-489. This article is available from: [http://www.biomedcentral.com/1471-2105/7/489] (2006)

[14] MAGE-ML: MicroArray Gene Expression Markup Language, (accessed 12 December 2015) [http://xml.coverpages.org/MAGEdescription2.pdf]

[15] W3C.org, “CSS Current Status”, (accessed 12 December 2015) [http://www.w3.org/standards/techs/css#w3c_all]

[16]W3C.org, “HTML 5.1, W3C Working Draft 08 October 2015”, (accessed 13 December 2015) [http://www.w3.org/TR/html51/]

[17] W3C.org, CSS Paged Media Module Level 3, W3C Working Draft 14 March 2013 (accessed 13 December 2015) [http://www.w3.org/TR/css3-page/#content-outside-box]