A shared standard for cytometry and pathology

Robert C. Leif[1], Stephanie H. Leif[1]

[1]XML_Med, a Division of Newport Instruments, San Diego, USA

Robert C. Leif*[a], Stephanie H. Leif[a]

[a]Newport Instruments, 3345 Hopi Place, San Diego, CA, USA 92117-3516

ABSTRACT

Introduction: The development of cytometry standards is complicated by their being relevant to pathology and biological science, which already have standards. CytometryML, the cytometry markup language, is an XML standard for flow and image cytometry, which includes both objects and their relationships, and is based upon existing standards: the International Society for Advancement of Cytometry (ISAC) FCS, Digital Imaging and Communication in Medicine (DICOM), and International Digital Publishing Forum (EPUB).

Methods: The CytometryML schemas are written in XML Schema Definition (XSD1.1). Object-oriented methodology was employed to create the CytometryML schemas, which were tested by translating specific XSD elements into XML and filling in the values. The attribute based syntax description of relationships in the Resource Description Framework (RDF) has been replaced by an XSD element based implementation. The ISAC Archival Cytometry Standard (ACS) concept of a zipped data container file was further refined to be a EPUB file. Since Table of Contents information is present in an EPUB container, it was minimized in the Relations schema, which replaced the ToC schema of the ACS and includes a modified and extended version of the ToC RDF capabilities.

Results: An XML based system that includes the DICOM specified separation of series and instances and includes relationships has been created.

Conclusions: CytometryML and EPUB could be used for the transmission of research and medical data and be extension some of the pathology part of DICOM. The CytometryML version of RDF in XSD could be extended to provide XSD with full RDF capabilities.

Keywords: CytometryML, DICOM, EPUB, FCS, Series, Standard, XML Schema, RDF

1 INTRODUCTION

In a recent review article on Telecytology[1], it was pointed out that one of the disadvantages of whole-slide imaging was a lack of standardization including software. Similarly, Lyttleton et al.[2] have stated: “The use of multiple encoding formats, instead of a single encoding format, adds to the burden of developing applications that read and write TMA DES (Tissue Microarray Data Exchange Specification) data”. This statement can be extended to cytometry, if DICOM and ISAC develop incompatible standards. The International Society for Advancement of Cytometry (ISAC) Data Standards Task Force (DSTF) is now finishing work on the “Archival Cytometry Standard (ACS), which has been developed to bundle data with different components that describe cytometry experiments”[3]. DICOM Working Group 26 is working on a standard that overlaps this area. The ISAC ACS captures relations among data and other components and includes support for audit trails, versioning, and digital signatures. The ACS already includes Gating-ML[4], which describes gating and color compensation. The ISAC ACS container is a ZIP file[5] that archives the binary data and related metadata files obtained from one or more cytometry data acquisitions of list-mode or similar data. The binary data can include the original binary data produced by the instrument and the binary files that are part of the sequence that results in data suitable for the display of cellular populations. The control data, if used in the processing of the data, can also be included. DICOM Working

*rleif@rleif.com; phone | 619 582-0437; | cytomertryml.org; newportinstruments.com

Group 27 is developing standards for retrieving, storing, and quarrying DICOM and XML metadata, as well as binary data containing standard formats, such as JPEG and FCS. Another ISAC activity is the present proposed ISAC Image Cytometry Experiment Format ICE format[6], which unfortunately is not related to the ACS, DICOM, or Open Microscopy Environment (OME)[7,8] standards.

CytometryML is an XML schema based translation, extension, and amalgamation of the DICOM and ISAC standards[9]. CytometryML is based upon an object-oriented design, which resulted in several major XML schemas: Relations, Series, Instance, Instrument, and Specimen; it also includes Image, and List-Mode schemas and multiple helper schemas. Series metadata, which is specific for an entire collection of images and/or list-mode files produced by a single instrument and derived from a single specimen, is stored together with related metadata files and instance files in an EPUB[10,11] container (ZIP) file. Each Instance metadata file includes relations that include hypertext references to binary image and/or list-mode files together with external related metadata files that are specific for a single or closely related group of instrument runs from a single specimen.

There has been considerable interest and work on the creation of Semantic Webs of data[12]. According to World Wide Web Consortium (W3C)[13], "The Semantic Web is a Web of data — of dates and titles and part numbers and chemical properties and any other data one might conceive of. The Resource Description Framework (RDF) (http://www.w3.org/RDF/) provides the foundation for publishing and linking your data." RDF is a way of documenting the relations between objects. These objects can be represented by URIs. Most embodiments of RDF technology employ schemas that are different and often incompatible with the XML Schema Definition Language (XSD), which is the basis of CytometryML and the ACS schemas. This greatly complicates the creation of schemas that need to both describe objects and their relationships because they would have to employ two different incompatible schema languages. XSD was designed to describe objects and RDF to describe their relationships. A very useful way to employ the equivalent of RDF in XHTML web pages is to use a special set of attributes[14]. However, it would be useful to describe the equivalent of relations employing RDF in standard XML schemas. The ToC schema of CytometryML[15] included a construct to describe relations between file references. The description of and the refinement of the ToC schema to a relation schema, which contains a Relation type is a major part of this paper. The description of Relation_Type is changed from including the standard attributes of RDFa to being a generic (template) that includes multiple elements that still include the necessary hypertext linkages between files and elements. Since the use of an attribute in a schema can only generate or validate one attribute value pair in an XML page, the change to an element, which can generate or validate unlimited element value pairs now simplifies describing cases where there is more than one relation between entities. The use of elements also permits the inclusion of choice elements that permit the user to select either an element that is part of the standard or Other_Type. The availability an Other_Type permits users to include possibilities that were omitted by the developers or are proprietary. A simple example where this could be useful is enumerated types. It is often improbable that the developers should foresee all possibilities of an enumeration.

2 DESIGN

2.1 Top-Level design: Because of adherence in the CytometryML schemas to the DICOM model of information objects[16] and their definitions (IODs)[16,17] has the benefits of increasing acceptability by pathologists and to correctly model specimen preparation. The CytometryML standard does not follow the ACS container file model, in that CytometryML splits the data into Series and Instance(s) files (Figure 1). The group of measurements in the ACS container file is analogous to the sum of the contents of a DICOM series together with its instance(s) files.

Separating a series object from its instance(s) results in the creation of one series and multiple instance XML metadata files. The use of corresponding XHTM web pages is a suitable means to present this data The first instance file is the one that describes the Most Relevant Image or List-Mode binary (Bulkdata) and metadata to the physician or scientist who uses the data. The degree of relevancy is to be determined by the creator of the instance file. Files that are the object of a relationship from another file shall be referred to as Related_Files. These could be instance XML files, or other metadata files, such as those that describe the processing of binary data containing files. In principle, each one of these Related_Files could have its own Related_Files. This can result in a complex tree of trees structure that contains multiple relations between file descriptions located above or below a file description. In the ISAC model, this connection is unidirectional between the original file and its related file. In CytometryML, it is possible to have the converse where the connection is between the related file and the original file; however, it can lead to redundancy. This capability is presently available in CytometryML, but the only recommended use is to create a doubly linked list, Figure 2.

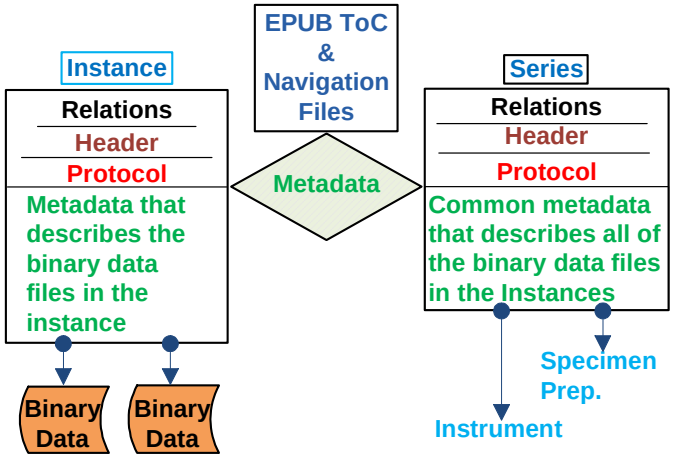

Figure 1 Diagram showing the division of measurement metadata into the Instance and Series XML files together. with the EPUB ToC and Navigation files. The Instance_Data_Type (left) and Series_Data_Type (right) and their corresponding elements each contain Header Information, a list of relations, Relation_List, and a description of the Protocol that contains the metadata necessary to analyse the data and eventually to repeat the measurement. There can be only one Series and will often be multiple Instances.

URIs to the binary data are included in the metadata. These URIs permit the selected binary data to be subsequently retrieved. For DICOM the binary files are external to the EPUB container; whereas, in the ACS, the binary files are inside of the EPUB container.

The Instance and Series schemas now import elements from a new Relations schema, which is an evolution of the ACS ToC schema[3]. The new Relations schema contains an Instance_Relation_List element, which is contained within the Instance schema. It will be described in the form of code fragments from an XML page. The Instance_Relation_List of an Instance schema can either be, as shown in the code fragments below, in a separate XML page or be included in a higher-level element in the Instance XML page, as was previously done[15]. Specific elements of the XML pages generated from the Instance and/or the Series schemas can provide data for an XML equivalent of a DICOM structured report[17,18,19] that includes a description of pathology and/or cytometry data. A preliminary Instance_List element that is appropriate for a DICOM Series is being implemented, but will not be described in detail in this paper.

2.2 DICOM Container Structure: The use of separate DICOM Series and Instance objects significantly simplifies the design by limiting the amount of information and resultant complexity in an Instance XML file to that contained in one measurement or a closely related group of measurements. The elements located within a Series file describe information that is relevant to all the Instances; whereas, an element located within an Instance XML page is specific for that Instance. Figure 1 is a modification of a previously described model[9]. This separation into Series and Instances limits the complexity of XML files in any container. In a client-server environment, such as DICOM, it minimizes the load on the server to transport the entire metadata including the instances for a series in a ZIP file. The size of metadata is small compared to that of binary data and the off-loading of the selection of the relevant parts of the metadata can be performed by the client. The extra step of selecting only the relevant, often quite large, binary files minimizes traffic on the network. Since an E-Pub file is a ZIP file that can contain XML, XHTML, and other files and is based upon an existing, well supported standard; it is suitable for to serve as a container file for cytometry and other XML based informatics standards. In the case of client server environments, once the relevant binary data containing files have been selected, copies of them are subsequently requested from the server and transferred to the client.

The ISAC ACS standard differs from DICOM in that it is for data transfer between computers or retrieval of files. and not specified for a client-server (cloud) system. The ACS specifies the complete transfer of the metadata and binary data in a single ZIP file. The structure shown in Figure 1 would represent the ACS design, if the binary files were shown inside the Instance.

A EPUB or ACS ZIP container that is used for file transfer can include: 1) Binary files: FCS and images, such as TIFF, JPEG, or DICOM. 2) metadata files that describe the content of the binary files and include hyperlinks to the descriptions of other binary files and 3) other metadata files, such as the descriptions of compensation and thresholding (scene segmentation), that describe processes and conditions that are relevant to the binary data files. The first of the instance metadata files is deemed to be of greatest interest to the end user[15]. For example, in the case of clinical data, the first metadata file would describe the binary Data_File that was or will be used for performing the diagnosis and a related data file that was the object of the relationship could be the one that was originally produced by the instrument. The relations between and roles are described in the Relations schema.

As is shown in Figure 1, the Series container includes or points to protocol information files that contain information that is relevant to all Instance files. Series files include the description of the fixed components of the Instrument and all or parts of the Specimen Preparation except those that are specific for an instance. The descriptions of the fixed components of the Instrument and all or parts of the Specimen Preparation can be resident in files that are not located in the Series XML file. These separate files can be used by multiple series of measurements. In fact, the nominative versions of the Instrument file, which describes the fixed components of the instrument, and the Specimen preparation files could be maintained by the manufacturer and reside on the manufacturer's web site. No matter where these files are located, version information for them must be accessible.

3 MATERIALS AND METHODS

Since the design including the element content and documentation was reused from the Digital Imaging and Communications in Medicine (DICOM)[16,17,20,21,22] standard or Flow Cytometry Standard, FCS3.1[23], much of the information and data-types present in the XML schemas and subsequently XML pages were prepared by domain experts. New data-types were created and data-types from other CytometryML schemas[24] were reused.

Because code in the Digital Imaging and Communications in Medicine (DICOM)[16] standard could be included in a FDA Class II device[25], the safety of the software developed as part of a standard should be maximized. A strong effort was made to: maximize readability, facilitate finding the sources of elements by adherence to naming conventions, modularize the code, minimize coupling between major schemas, maximize cohesion of individual schemas, and reuse of existing CytometryML[5,9,26] XML schemas. The schemas and XML pages including those for the EPUB were prepared and validated with oXygen using the Xerxes validation engine for XSD1.1[27,28,29].

4 RESULTS

4.1 EPUB container: The ACS container file design[3] has been modified and extended to be an EPUB file[30]. The use of a format with widespread and growing use, such as EPUB eliminates the need for special software to create and display the ACS container files. The EPUB format was initially described by Open e-book Publication Structure (OEBPS), and now is specified[10] to be the EPUB Open Container Format (OCF) 3.0. The abbreviation OEBPS is the MIME media type application/oebps-package+xmlname. The modification to the ACS container consists of the addition of a total of 4 small XHTML and XML files[11], which in order to facilitate use provides a means to include a human- and machine-readable global navigation layer in the document or publication. An EPUB file has the capacity to include XHTML documents that contain the data in an easy to read format. Many of these files include an XML Signature, which is a W3C Recommendation[31],which has been adopted by the ISAC DSTF to allow for digital signatures of data and other components within the ACS container[3].

The only significant difference between EPUB and ISAC container files is in their ToC pages The EPUB zip container can include two Table of Contents, toc.xhtml and toc.ncx (Code Fragment 1), as well as a manifest[30] (Code Fragment 2) that contains references to “All Publication Resources” “regardless of whether they are included in the EPUB Container or made available remotely”, the ACS design includes a ToC schema, which has been replaced a combination of the 2 EPUB toc elements and a list of relations (Figure 1). The toc.ncx Navigation Control XML page describes the pages that will be visible in the EPUB Container and their order of appearance. The ncx root element contains a navMap, such as the one shown below.

4.2 Special EPUB elements

Code Fragment 1, Fragment of a navMap Element from an EPUB Table of Contents (toc.ncx) file

```
<navMap>
   <navPoint id="front_cover" playOrder="1">
      <navLabel>
         <text>CytometryML Metadata</text>
      </navLabel>
      <content src="front-cover.xhtml"/>
      </navPoint>
   <navPoint id="series" playOrder="2">
      <navLabel>
         <text>Series Info</text>
      </navLabel>
      <content src="series.xhtml"/>
    </navPoint>
   <navPoint id="instance3" playOrder="3">
      <navLabel>
         <text>Data_Of_Greatest_Interest</text>
      </navLabel>
      <content src="instance3.xhtml"/>
    </navPoint>
```

Code Fragment 1, which is a fragment of a navMap element, describes an order of front_cover, series, instance3 (Data_Of_Greatest_Interest) and other instance files (not shown). The navMap element encloses all of the visible files in the EPUB and states their order in the compound document. In this case, only XHTML files have been included. Other files that are directly viewable and understandable by humans after formatting could be included.

Code Fragment 2, fragment of a manifest element from an EPUB (content.opf) file

```
<manifest>
   <item id="ncx" href="toc.ncx" media-type="application/x-dtbncx+xml"/
>
   <item id="htmltoc" properties="nav" media-type="application/
xhtml+xml"
      href="toc.xhtml"/>
   <item id="front_cover" href="front-cover.xhtml"
      media-type="application/xhtml+xml"/>
   <item id="series" href="series.xhtml"
      media-type="application/xhtml+xml"/>
```

```
6     <item id="instance3" href="instance3.xhtml"
         media-type="application/xhtml+xml"/>
7     <item id="instance1" href="instance1.xhtml"
         media-type="application/xhtml+xml"/>
8     <item id="instance2" href="instance2.xhtml"
         media-type="application/xhtml+xml"/>
9     <item id="style" href="stylesheet.css" media-type="text/css"/>
10    <item id="series_data" href="series.xml"
         media-type="application/xml"/>
11    <item id="instance1_data" href="instance1.xml"
         media-type="application/xml"/>
12    <item id="classifier" href="classifier1.xml"
         media-type="application/xml"/>
13    <item id="compensation" href="compensation.xml"
         media-type="application/xml"/>
14    <item id="fig1" href="fig1.svg" media-type="application/svg"/>
15    <item id="img1" href="Cells of Greatest Interest/image001.jpg"
         media-type="image/jpeg"/>
16 </manifest>
```

A content.opf file starts with a metadata element (not shown) that includes Dublin Core elements: title, language, an identifier, and a modification date. This is followed by the manifest element, which differs from the navMap element by including all of the files in the EPUP container. Thus, for each XHTML file (Elements 4 to 8), there exists an XML file (Elements 10 to 13) that is the source of the data that is displayed by the XHTML file. One or more formatting files such as, element 9, stylesheet.css and graphics files such as, elements 14 and 15. The toc.ncx file (Element 2) is also referenced in the manifest.

The inclusion of the XHTM files is based on the requirement to create and read structured reports. Since EPUBs are used for the creation of magazines and books and when they contain XHTML5, capable of rendering media, such as sound and motion pictures, EPUBs have the capability to contain and permit the rendering of multimedia containing structured reports. The XHTML5 form elements are suitable for data entry by clinicians and scientists.

There already are two ToCs (toc.ncx and toc.xhtml) in an EPUB. toc.xhtml (not shown), which basically is a list that includes hyperlinks (href attributes) to the cover and XHTML pages included in the EPUB. Since the addition of a third toc document could cause confusion, the original CytometryML ToC schema, which is based in part from the ACS Table of Contents (ToC) schema[3], has been renamed and reorganized to be a Relations schema. Both the old ToC and new Relations schemas include some of the functionality of the Resource Description Framework (RDF)[12,14], which is used to document relations.

The syntax of the XML Schema Definition Language (XSD1.1) structures[27] and data-types[28] was helpful in this attempt to produce safe code. Since XSD version 1.1 also includes assertions, it can provide extra checks on the correctness of the code and XSD1.1 also includes the generalized capacity to create generic (template) elements. The combination of the new capacity to produce generic elements and the existing capacity to extend an element permits creation of schemas that each describe an object (objected oriented design). The CytometryML schemas have been coded to be similar to object and data-type declarations in classes. In order to satisfy the requirement of the ISAC Archival Cytometry Standard (ACS) for a data file standard[32] that a “Detailed semantic shall be provided to prevent potential misinterpretations and misusages of the standard”, readability has been maximized. Many of the other requirements for the ACS data file standard[33] were met or facilitated by the use of XSD1.1. The absence of methods (functions and procedures) greatly simplified the process; however, it does not permit the CytometryML schemas to be the complete basis of any web page or other entity that is more than a passive document. References (pointers) to the Instance files rather than the binary image files were

included in the relationships because 1) in some cases (FCS) the browsers would be unable to display the binary data; 2) DICOM works at the instance level; and 3) the metadata of images are of interest and may be required by the human reader and/or the computer program to make an image intelligible.

4.3 Relations

Because EPUB files already include a ToC named toc.ncx[30], a Relation_List element has replaced the ACS and CytometryML ToC. This Relation_List is part of the Relations schema and has been included in Series and Instance schemas and prototyped for an EPUB file that is essentially an extension of the ISAC standard ACS Container file[15]. Effectively, the EPUB Container file design has been simplified from the ACS format by following the DICOM design, which divides the content of the ACS container into a Series and Instances hierarchy. The XML code of an individual Instance can contain indirect references to binary data other than the binary data it describes by including direct references to the XML code that describes the binary data of an other instance and this second instance XML code will contain a hyperlink to the second binary data file. Since multiple binary data objects are produced during the transformations of the raw data into interpretable data, multiple metadata files, each of which is relevant to an individual binary data file, will often be included in the EPUB. These will be accompanied by the method metadata files, such as those that describe compensation and gating.

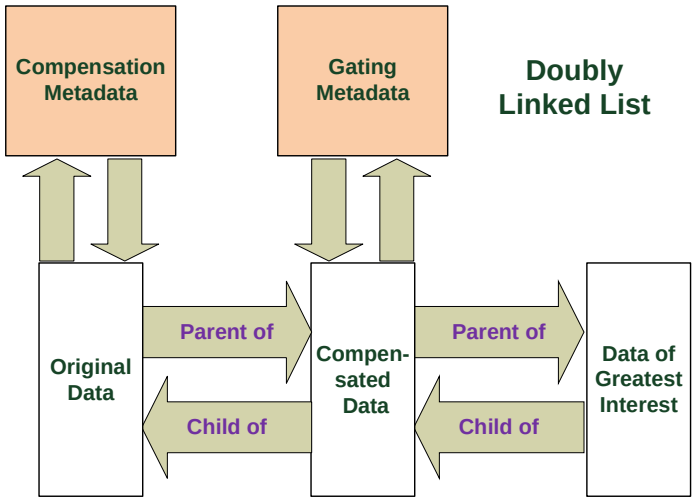

Figure 2 shows relations between three elements that each describe and are linked to a member of a group of related binary files. These are in ascending order 1) the original data, 2) the compensated data, and 3) the data of greatest interest, which in this case is the data that describes the various cell populations present. For each element that describes a binary file except for the Data of Greatest Interest and the Original Data, links (fat arrows) to both a parent and a child metadata file are shown. Since the Data of Greatest Interest is the last file, it does not have any children and the Original Data file does not have any parents.

The original list-mode data (Original Data file) is acted upon as described in the Compensation Metadata. Both the Compensation Metadata and the element that references the Original Data file can have references to each other. Similarly the Compensated Data reference and the Gating Metadata could have references to each other.

Figure 2 is a diagrammatic representation of the relations between 3 XML elements that each directly describes a binary data file and the relations between two pairs of a metadata elements and a binary data describing element. The metadata that describes a transformation is bidirectionally joined with the an element that describes the binary data file that it transforms. The presence of multiple relations in either direction permits the data to be organized as a doubly linked list. The use of a doubly linked list permits the human user to sequentially navigate in both directions between related files. Since hyperlink navigation is essentially how web pages are traversed, the Relations schema is based on an efficient, reliable, well developed technology.

Code Fragment 3, examples of the use of a prefix

```
 <?xml version="1.0" encoding="UTF-8"?>
1<instance:Instance_Metadata
 xsi:schemaLocation="http://www.cytometryml.org/ACS/instance
 instance.xsd">
 xmlns:comp="file:///cytometryML/compensation/"
 xmlns:test="file:///ACS_Image_Data/20Aug2012/"
```

```
 xmlns:instance="http://www.cytometryml.org/ACS/instance"
 xmlns:relations="http://www.cytometryml.org/ACS/relations"
 xmlns:relations_image="http://www.cytometryml.org/ACS/relations_image"
 xmlns:relations_meta="http://www.cytometryml.org/ACS/relations_meta"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
2  <instance:Relation_Image_List>
```

Since standard RDF includes many URIs, which are often strings of considerable length, the structure of the code can be obscured by the presence of these URIs. This problem was initially solved by the creation of the Compact URI or CURIE, which is the abbreviation of Compact URI.[34] Subsequently in RDFa[14], it was changed to prefix, such as the attributes comp and test, which, are shown above in Code Fragment 3. All prefix declarations start xmlns followed by a colon and an equals sign and a quoted reference. This quoted reference replaces the prefix when it is added to the rest of the URI. The prefix, test, is used below in elements 4 and 19 of Code Fragment 4 , Relationships between binary files and a meta-data file.. “The CURIE is replaced with a concatenation of the value represented by the prefix and the part after the colon,[34]. As is shown in Element 4 of Code Fragment 4 , Relationships between binary files and a metadata file., the reference is File3.html. Thus, the value of the URI is file:///ACS_Image_Data/7May2012/File3.html.

Code Fragment 4, Relationships between binary files and a metadata file.

```
1<rel_img:Instance_Image_File_Ref id="File3_Bin">
2     <rel_img:File_Location>
3        <links:Curie_Link>
4           <links:a href="test:File3.jpeg">Image3</links:a>
        </links:Curie_Link>
      </rel_img:File_Location>
5     <rel_img:Significance>Diagnostic</rel_img:Significance>
6     <rel_img:Reference_Std>Most_Relevant_Image_Reference
      </rel_img:Reference_Std>
7     <rel_img:Role_Std>is the Data Of Greatest Interest
      </rel_img:Role_Std>
8     <rel_img:Role_Std>is an Instance</rel_img:Role_Std>
9     <rel_img:Role_Std>is Processed_Data</rel_img:Role_Std>
10    <rel_img:Subject>
11       <relations:Self>Image3</relations:Self>
      </rel_img:Subject>
12    <rel_img:Predicate_Phrase/>
<!--Each Predicate Phrase consists of a verb and one or more objects-->
13    <rel_img:Verb_Phrase_Std>is child of</rel_img:Verb_Phrase_Std>
14    rel_img:Object>
15       <links:IDREF_Link>
16          <links:a href="File2_Meta">Metadata for File 2</links:a>
17       </links:IDREF_Link>
18       <links:Curie_Link>
19          <links:a href="test:File2.jpeg">Image2</links:a>
         </links:Curie_Link>
```

```
      </rel_img:Object>
20    <rel_img:End_Predicate_Phrase/>
21    <rel_img:Predicate_Phrase/>
22    <rel_img:Verb_Phrase_Std>is classification-results of
      </rel_img:Verb_Phrase_Std>
23    <rel_img:Object>
24       <links:Curie_Link>
25          <links:a href="meta:classifier14Aug12.xml">
               gate_and_linear_discriminant metadata</links:a>
         </links:Curie_Link>
      </rel_img:Object>
26    <rel_img:End_Predicate_Phrase/>
27    <rel_img:Additional_Info>This is an example of a relation with one
         object. There could have been multiple objects.
      </rel_img:Additional_Info>
28    <rel_img:figure>
29       <components:Img src="figs:dividing_cell_population.jpeg"
            alt="Fig. 1 shows a dividing cell population"/>
30       <components:figcaption>Fig 1. which is based upon the Image3
            file, shows a dividing cell population stained with Eu labeled
            Quantum Dye(R)</components:figcaption>
      </rel_img:figure>
   </rel_img:Instance_Image_File_Ref>
```

Code Fragment 4 is the metadata that describes to the investigator or physician the relations of the element that describes the binary data file that is the Most_Relevant_Image as stated in Code Fragment 4 Element 6 includes a hypertext reference (href) (Element 4) to the final binary data produced by the processing steps. It is part of a simplified example of an XML page based on the relations_image schema that includes a Relation_Image_List of which one member of the list is the Instance_Image_File_Ref (Element 1). The code shown in Code Fragment 4 would be included in an Instance XML page, which would be included in the EPUB container. Code Fragment 4 includes a description of only one of the three binary containing files. The other two are File 1, which contains the original binary data and File 2, which contains the compensated data. The Instance XML page would also include or refer to two XML based metadata files, one of which describes the compensation algorithm including the compensation matrix and a second the gating algorithm (scene segmentation processing). Many of the values used to test these pages were those used in Spidlen et al[3].

The URI in Element 4 starts with the prefix test (Code Fragment 3) and ends with the file's name. The number that ends the file name is sequentially incremented as the data in the file is processed. It starts with 1 for the original (raw data) and is incremented by 1 with each processing step.Thus, the result of a process has a greater valued number (Element 4) than that of the original raw data, File1.jpeg (not shown) and File2.jpeg (Element 19) The data structure for this element is based on an element in the XLink specification[35]. An optional HREF to an id attribute is also available. Activation of this hypertext reference opens the data file or code fragment.

Instance_Image_File_Ref (Element 1) is based on a datatype that was derived from the Relation_Type complexType. This complexType is based on the unambiguous structure of a simple sentence with three classes of elements: Subject, Predicate, and Object elements, which each can contain other elements. This type of data structure that includes, subject. predicate and object is referred to in the draft RDF Vocabulary Description Language 1.1: RDF Schema[36] as reification. Our design differs from RDFa, which describes its values in the form of attributes including URIs. The Subject, Object and Predicate have an order, since they are contained inside a sequence element (Elements 10 to 20); whereas, no specific grouping of RDF elements is required. The relation element, Instance_Image_File_Ref, contains single Subject (Element

10) and begins with elements that describe this Subject. These elements include: the File_Location (Element 2), Significance (Element 5), Reference (Element 6), and Role, (Elements 7, 8, and 9). A Subject can have Multiple roles. For example, the Role, Data_Of_Greatest_Interest (Element 7), means that the Data_File referenced by this element is of greater interest than all of the other Data_Files. Two other roles have been specified (Elements 8 and 9). These are respectively is an Instance and is Processed_Data. The composition of the Role element is flexible, since it can be specified by restriction of the Relation_Type in the schema were it is used and can consist of a union between two or more enumerations. In order to permit Roles other than those enumerated in the standard, a choice between Role_Std_Type, which is an enumerated type and Role_Other, which can be any user supplied simpleType, such as a string. Similar choices including Other are permitted for many of the enumerated types. This permits a means for users and vendors to create their own types; however, the fact that a type is nonstandard is readily detectable from the code. The creation of a datatype which includes a choice and can have a variable number of occurrences other than one or zero requires the use of an element, which differs from the use of an attribute by RDFa. As shown in Code Fragment 4, the descriptive elements appear before the elements that are involved in relations.

The relations described by CytometryML are based on elements contained in a complexType, which is organized around a single Subject, such as Element 10 and multiple Predicate_Phrase elements (Elements 12 and 21), each of which is terminated by End_Predicate_Phrase element (Elements 20 and 26. These pairs of elements serve as delimiters of Verb_Phrase and Object(s) combinations. A Verb_Phrase can have one or more object elements. Code Fragment 4 contains two Verb_Phrase elements and Two Object elements.

The creation of a list of relations is simplified by limiting each relation element to a single unique Subject, such as Element 10. Since a hyperlink (element 4) to the binary data file was previously included, a Self element was introduced into the design (Element 11) in order to decrease the size of the file and increase readability. In this case the value of "Self" is the previously named File3 (Element 4).

The first Verb_Phrase and Object combination consists of the Verb_Phrase (Element 13) is "is child of" and the Object (Elements 14to 19) is "File2". This unambiguously states that Self (File3) is child of File2. In the second Verb_Phrase and Object pair, the Subject is still Self ("file3"; the Verb_Phrase (Element 22) is "is classification-results of" and the Object (Element 25) is "classifier14Aug12.xml". The availability of both the "is child of" and is parent of Verb_Phrases permits an Instance to contain a doubly linked list. As is the case for the Role element there is a choice of the enumeration in Predicate_Std_Type or user defined values in Predicate_Other_Type. The contents of Verb_Phrase enumerations are flexible in a similar manner to that of the Role elements.

A free text input Additional_Info element is permitted (Element 27). An optional figure (Element 29) and caption (Element 30) have been included. An optional Signature element is also available but because of its size has not been included in Code Fragment 4.

Other relations that are included (but not shown) are those for image metadata elements describing Files 2 and 3 as well as the metadata elements for those relations where the subject is the metadata used for the transformations of image1 and image2.

5 CONCLUSIONS

A design that is based upon conformance to both DICOM and FCS has been created and implemented in the XML Schema Definition (XSD) language. The ACS design for a Table of Contents (ToC) has been extended and modified in CytometryML to be an XML schema datatype that is based upon elements and is appropriate for the DICOM Instance data structure. This extension includes Role and Relation elements based on upon a Relation_Type, which includes the grammatical contents of a simple sentence. This simple sentence format generalizes, organizes, demystifies, disambiguates, and simplifies the RDF expression syntax by including simple directionality and providing a simple syntax for multiple relations with different directions between two XML elements or files. This combination of Role and enhanced Relation types should provide a much richer vocabulary to describe objects and their relations than present RDFa, which is based on attributes and where no specific organization or grouping of RDF elements is required. The use of hypertext links in a structured report will provide a simpler more user friendly means of traversing the data than the standard database query approach. The CytometryML design and that of Spidlen et al.[3] have the advantage that since they have been

created in XSD, they can be easily imported into other XSD schemas.

The CytometryML approach differs from potentially complex structure proposed Spidlen et al[3]. by being a description (metadata) that is limited to a single Instance or a small closely related group of instances. This permits the Relations_List element to be either a single or a small number of doubly linked lists. The order of these lists is based on starting with the element that contains the description of the binary data of greatest interest to the target user. In many cases, this element will be the only one of interest to the user, which will limit the amount of binary data requested from the server or other data storage.

The CytometryML design and that of Spidlen et al.[3] have the advantage over both DICOM and FCS that since they have been created in XSD, they can be easily imported into other XSD schemas. The separation of the descriptions of the metadata and binary data from cytometry measurements into a Series, which primarily contains metadata that is applicable to multiple Instances, and Instances, each of which contains a limited set of binary data, will decrease the amount of extraneous data transmitted. This decrease will, as experienced with DICOM, provide improved response times and lower costs.

In this work, the reuse of an ACS construct in a DICOM based data structure has been demonstrated, which is an extension of previous reuse of DICOM data structures in CytometryML[39]. This leads to the conclusion that the ACS and CytometryML should both reuse DICOM and DICOM should reuse ACS and CytometryML. These standards should be harmonized at the datatype semantics level. The harmonization of these standards and similar work on harmonizing other medical standards and translating them into XML[24,37,38,39] will significantly assist in meeting Executive Office of the President's Council of Advisors on Science and Technology's goal[40] of "capability for universal data exchange." Since DICOM usually is tested by being implemented in a compiled language, the creation of XML schemas and pages presents an opportunity for early, but partial testing of a DICOM design.

The use of EPUB for ISAC standard(s) has the added benefit that if a journal, such as Cytometry Part A were published as an EPUB document, the transfer of information from data files to Cytometry Part A articles and the checking of the MIFlowCyt information[42,43] could be semi-automated.

6 ACKNOWLEDGMENTS

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, the members of DICOM Working Groups 26 and 27 for the knowledge that they have provided, and Jules J. Berman for calling attention to and demonstrating the importance of RDF technology.

7 REFERENCES

1] Thrall M, Pantanowitz L, Khalbuss W. "Telecytology: Clinical applications, current challenges, and future benefits, Review Article" J Pathol. Inform. 2(51) (2011).

2] Lyttleton O, Wright A, Treanor D,1 and Lewis P. "Using XML to encode TMA DES metadata" J Pathol Inform. 2(40) (2011).

3] Spidlen, J. and The ISAC Data Standards Task Force (STF), "Archival Cytometry Standard ACS, International Society for Advancement of Cytometry, Draft Candidate Recommendation" http://flowcyt.sf.net/acs/latest.pdf (2010).

4] Spidlen, J., Leif, R.,C., Moore, W., Roederer, M., International Society for the Advancement of Cytometry Data Standards Task Force, Brinkman, R. R., "Gating-ML: XML-based gating descriptions in flow cytometry," Cytometry A 73A, 1151-1157 (2008).

5] Leif R.C., Spidlen J., Brinkman R.R., "A Container for the Advanced Cytometry Standard (ACS)," Proc. SPIE 7182, 71821Q (2009).

6] Spidlen, J. and Novo, D. (2012), "ICEFormat—the image cytometry experiment format" Cytometry, 81A: 1015–1018. doi: 10.1002/cyto.a.22212

7]Goldberg IG, Allan C, Burel JM, Creager D, Falconi A, Hochheiser H, Johnston J, Mellen J, Sorger PK, Swedlow JR. "The open microscopy environment (OME) data model and XML file: Open tools for informatics and quantitative analysis in biological imaging. Genome Biol 6(5), R47 (2005).

8] “OME, The Open Microscopy Environment” http://www.openmicroscopy.org (2012).

9] Leif, R.,C.,“Toward the integration of cytomics and medicine,” J. Biophoton. 2, 482–493 (2009).

10] International Digital Publishing Forum (IDPF) “New Digital Book Standard Released” http://idpf.org/news/new-digital-book-standard-released (2006).

11] “EPUB Publications 3.0, Recommended Specification 11 October 2011” http://www.idpf.org/epub/30/spec/epub30-publications.html (2011).

12] Beckett D. (Editor), “RDF/XML Syntax Specification (Revised), W3C Recommendation 10 February 2004” http://www.w3.org/TR/rdf-syntax-grammar/ (2004).

13] “W3C, SEMANTIC WEB” http://www.w3.org/standards/semanticweb/ last visited (2012).

14] Adida, B., Birbeck, M., McCarron, S., and Pemberton, S. (Editors), “RDFa in XHTML: Syntax and Processing, A collection of attributes and processing rules for extending XHTML to support RDF, W3C Recommendation 14 October 2008” (http://www.w3.org/TR/rdfa-syntax/) (2008).

15] Leif R C., and Leif S.H., “A CytometryML Table of Contents that Describes Relationships between Elements based upon DICOM and Flow Cytometry Standard,” Proc. SPIE 7902, (790217) 1-9 (2011).

16] Pianykh O.S., [Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival Guide], Springer Publishers, Berlin & Heidelberg, (2008).

17] “DICOM) Part 3: Information Object Definitions, Annex A.35 Structured Report Document Information Object Definitions,” ftp://medical.nema.org/medical/dicom/2009/ (2009).

18] D. A. Clunie, [DICOM Structured Reporting] PixelMed Publishing ISBN 0-9701369-0-0 Available at: http://www.pixelmed.com/srbook.html (2005).

19] Lee, K.,P., Hu, J., “XML Schema Representation of DICOM Structured Reporting” Journal of the American Medical Informatics Association 10, 213-223, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC150374/pdf/0100213.pdf (2003).

20] “Digital Imaging and Communications in Medicine, DICOM” http://medical.nema.org/ (2013).

21] DICOM Working Group 26, “DICOM Supplement 145: Whole Slide Imaging in Pathology” which applies to Parts 2,3,4,6,16,17 of the DICOM Standard (http://www.dclunie.com/dicom-status/status.html#BaseStandard2001) (2009)

22] DICOM Working Group 26, “DICOM Supplement 122: Specimen Module and Revised Pathology SOP Classes,” applies to Parts 2,3,4,6,16 of the DICOM Standard http://www.dclunie.com/dicom-status/status.html#BaseStandard2001 (2008).

23] International Society for Advancement of Cytometry (ISAC) Data Standards Task Force, (DSTF) “Data File Standard for Flow Cytometry Version FCS 3.1, Normative Reference” http://www.isac-net.org/images/stories/documents/Standards/fcs3.1_normativespecification_20090813.pdf (2009).

24] Leif R.C., Leif S.B., and Leif S.H., “CytometryML, an XML Format Based on DICOM and FCS for Analytical Cytology Data” Cytometry 54A, 56-65 (2003).

25] Office of Device Evaluation, “Guidance for the Submission Of Premarket Notifications for Medical Image Management Devices” US FDA, http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm073721.pdf (2000).

26] Leif R.C., “CytometryML, Binary Data Standards,” Proc. SPIE 5699, 325-333 (2005).

27] “W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures, W3C Recommendation 5 April 2012”http://www.w3.org/TR/xmlschema11-1/ (2012).

28] “W3C XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes, W3C Recommendation 5 April 2012” http://www.w3.org/TR/xmlschema11-2/ (2012).

29] Walmsley P., [Definitive XML Schema, Second Edition], ISBN: 0132886723 Prentice Hall (2012).

30] International Digital Publishing Forum (IDPF), “EPUB Open Container Format (OCF) 3.0, Recommended Specification 11 October 2011” http://idpf.org/epub/30/spec/epub30-ocf.html (2011).

31] W3C, “W3C Recommendation - XML Signature Syntax and Processing (Second Edition)” http://www.w3.org/TR/xmldsig-core/ 2010 (2010).

32] Spidlen J., Brinkman R., Leif R. C., and other members of the ISAC Data Standards Task Force. “Advanced Cytometry Standard (ACS) Requirements for a data file standard format to describe cytometry and related analytical cytology data, Version 0.070920,” http://cdnetworks-us-1.dl.sourceforge.net/project/flowcyt/Analytical Cytology Standard/Analytical Cytology Standard/Requirements-v070920.pdf (2007).

33] Spidlen J., Ryan Brinkman R., Robert Leif R. C., and other members of the ISAC Data Standards Task Force, “Requirements for a data file standard format to describe cytometry and related analytical cytology data, Version 0.070920, September 20, 2007” http://iweb.dl.sourceforge.net/project/flowcyt/Archival%20Cytometry%20Standard/Obsolete/Requirements-v070920.pdf “(2007).

34] Birbeck, M., and McCarron, S. (Editors) “CURIE Syntax 1.0, A syntax for expressing Compact URIs”, W3C Candidate Recommendation 16 January 2009,” http://www.w3.org/TR/2009/CR-curie-20090116 (2009).

35] DeRose S., Maler E., Orchard D., Walsh N., “XML Linking Language (XLink) Version 1.1, W3C Recommendation 06 May 2010” http://www.w3.org/TR/2010/REC-xlink11-20100506 (2010).

36] “RDF Vocabulary Description Language 1.1: RDF Schema, W3C Editor's Draft 06 January 2013” https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-schema/index.html# (2013)

37] Catley, C., Frize, M., “Design of a health care architecture for medical data interoperability and application integration” Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint 3, 1952- 1953 (2002).

38] Hongli, L., Chen, Z., Wang, W., “XML Schemas Representation of DICOM Data Model,” Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference,” (2010).

39]Leif, RC.,“An XML Cytometry Standard Based on DICOM,” SPIE BIOS Proceeding, 7264, 72640H (2009).

40] Executive Office of the President, President’s Council of Advisors on Science and Technology, “Report To The President, Realizing The Full Potential Of Health Information Technology To Improve Healthcare For Americans: The Path Forward” http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-health-it-report.pdf (2010).

41] McCarron, S., (Editor) “XHTML+RDFa 1.1, Support for RDFa via XHTML Modularization”, W3C Recommendation 07 June 2012” http://www.w3.org/TR/2012/REC-xhtml-rdfa-20120607/ (2012).

42]. Lee JA, Spidlen J, Boyce K, Cai J, Crosbie N, Dalphin M, Furlong J, Gasparetto M, Goldberg M, Goralczyk EM, Hyun B, Jansen K, Kollmann T, Kong M, Leif R, McWeeney S, Moloshok TD, 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 A. 73A, 926-930 (2008).

43] Spidlen, J., Breuer, K., Brinkman, R., "Preparing a Minimum Information about a Flow Cytometry Experiment (MIFlowCyt) Compliant Manuscript Using the International Society for Advancement of Cytometry (ISAC) FCS File Repository” (FlowRepository.org)” Current Protocols in Cytometry, http://dx.doi.org/10.1002/0471142956.cy1018s61, DO - 10.1002/0471142956.cy1018s61 John Wiley & Sons, Ltd, Published Online: 1 JUL (2012)