

DICOM in the XML Schema Design Language (XSDL)

Robert C. Leif

Newport Instruments

5648 Toyon Road

San Diego, CA, 92115-1022

rleif@rleif.com, www.newportinstruments.com

Annual Advancing Practice, Instruction, and Innovation through Informatics (APIII 2009) Conference

Electronic Poster 712

Abstract

Content: In 1998, I proposed that the International Society for Advancement of Cytometry should replace their present Flow Cytometry Standard with an implementation in DICOM and work with the DICOM developers. After this suggestion was rejected, I suggested an XML implementation based upon DICOM to the International Society for Advancement of Cytometry Data Standards Task Force.

Technology: XML Schema Design Language is the basis for the development of a new standard (CytometryML) for cytometry metadata. The schemas have been validated with XMLSpy and Stylus Studio. The relevance of the composition of the major schemas has been tested by the automated creation of XML pages and the manual entry of data into them.

Design: Since only data-types including data structures and objects are described in schemas, object oriented design principles were followed in their design, which, wherever possible, was based on DICOM. Strong typing, minimization of the coupling of independent higher-level schemas, and maximization of the cohesion of individual schemas were design principles. Container files for multiple XML pages that included a table of contents were employed instead of mimicking a DICOM directory structure.

Results: XML schemas that describe the metadata XML documents for two related types of container files have been created. The series container includes the metadata files that are constant for a set of measurements, such as the instrument description. The instance container includes the metadata and binary data that is specific to an individual or small closely related group of measurements, such as the instrument settings and staining protocols. Schemas for digital microscopes and flow cytometers, as well as image and list-mode (waveform) data have been created.

Conclusion: It is possible to translate DICOM data-types into conventional, readable, modular XML

schemas. Both the XML metadata and binary data files that describe instances can be placed together in a container based upon a ZIP file. The reuse of the well tested DICOM model and descriptions resulted in a great decrease in the design and documentation effort and increased the probability of success.

Pathology = DICOM + XML

- DICOM WG 26 is doing Classical pathology (transmission microscopy) and whole slide imaging
- The International Society for Advancement of Cytometry (ISAC) Data Standards Task Force is developing in XML Schema Definition Language (XSDL) the Advanced Cytometry Standard (ACS), which includes fluorescence, confocal, and other microscope modalities, which are part of the future of pathology.
- Pathology and Cytometry are a Continuum!
 - ISAC primarily does research & Pathology uses the technology.

Why Extensible Markup Language (XML)?

- Since the DICOM data representation is unique, special interface software must be written to interface it to other software and to hardware.
- XML is ubiquitous
 - XHTML (Web applications) is XML.
 - A very large amount of data is stored in XML.
 - XML has been interfaced to virtually all commercial software including databases.

XML Schema Design Language (XSDL)

- Readable
- Strongly Typed
 - Range checking (No 750 year olds)
- Data Structures
 - Enumerated types & Choices
- Object-Oriented
 - complexType are classes
 - Extension
 - Restriction (templates or generics)
- Significant usage (Microsoft and others)
 - Create Type declarations for header files

DICOM Organization

1 Patient (Medical Record)

X Studies

Y Series

Z Instances

The items below the dotted line are the part of the CytometryML schemas that will be discussed.

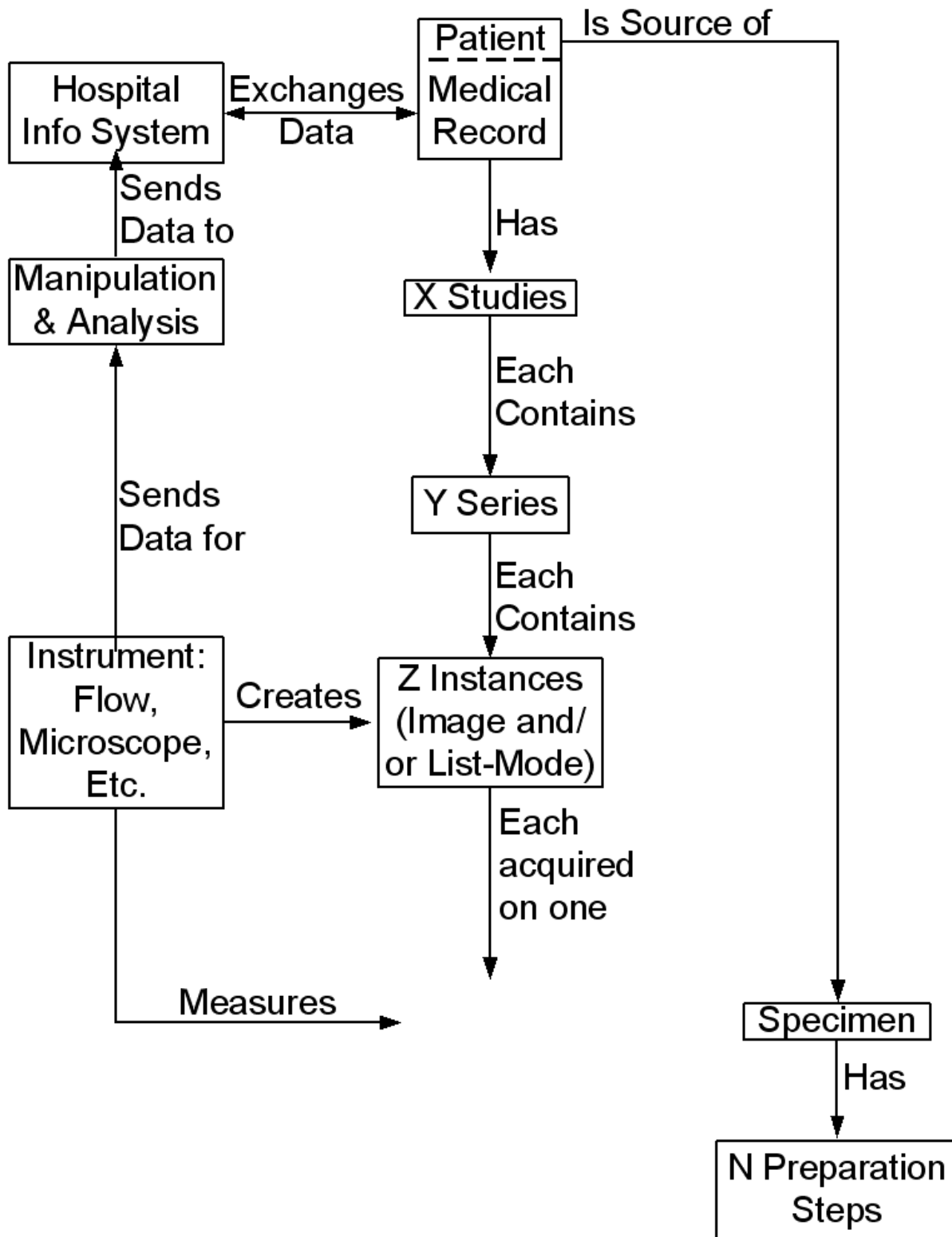


Figure 1 shows the movement and organization of data in a combined cytomics clinical data information system. It is an extension and adaptation for cytometry of DICOM Supplement 122 [2]: Specimen Identification and Revised Pathology SOP (Service-Object Pair) Classes. The numbers X, Y, and Z include 1. All Z instances are described by a single series. The patient is the source of the specimen; and the patient's medical record is the source and recipient of the patient data. If the data is for other than clinical studies, the patient and hospital information systems would be replaced by entities from the appropriate domains.

Container Files

- Zip file
- The Metadata_Type describes the contents of the container file. It consists of 2 elements: the Header and Table of Contents (ToC).
- Header
 - All file references in the header part of the metadata refer to the container zip file.
- ToC
 - The ToC describes the files that are in the container file or provides a URI, Uniform Resource Identifier Reference, that points to the files that are external.
- Two types of container files: Series and Instance

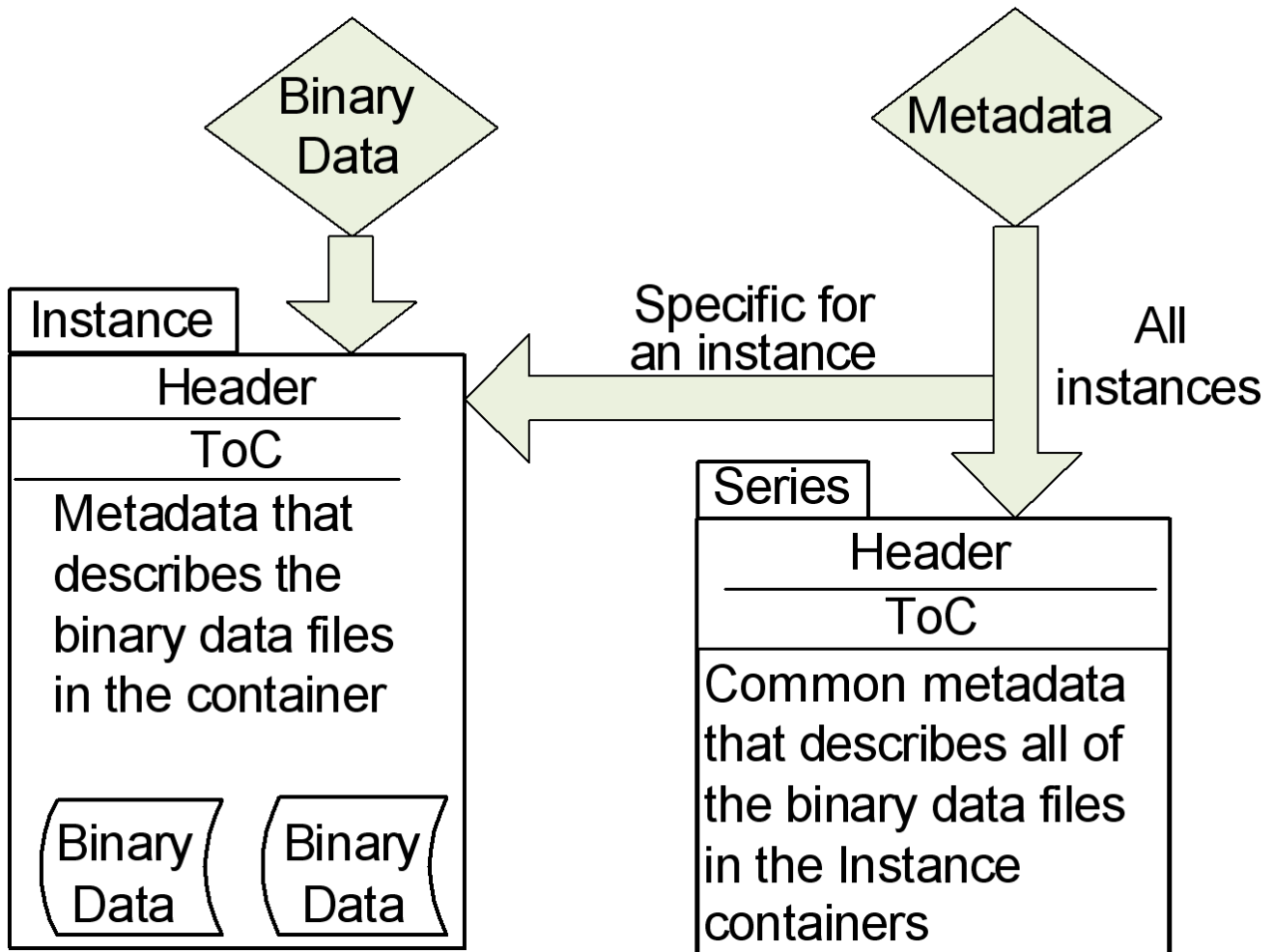


Figure 2 is a diagram that shows the placement of binary data and metadata into the instance and series containers. The Series_Data_Type (right) and the Instance_Data_Type (left) and their corresponding elements each contain Header Information and a Table of Contents (ToC).

CytometryML Container Header

. Abbreviations: S=Series, I= Instance, U= unbounded

Table 1: Present in the Series and Instance Headers

Item	Description and Comments	minOccurs	maxOccurs
Creation Date & Time	This is the time stamp for both XML pages	S=0 I=1	S=1 I=1
Requested by	In a medical application, this could be the physician who requested the measurement In other cases it could be a computer.	S=0 I=0	S=1 I=U
Performed by	This could be the operator of the cytometer. If the cytometer were a complete robot, it would be a computer.	S=0 I=0	S=U I=U
Responsible Individual	This lab manager or could be the pathologist in charge of the lab.	S=0 I=0	S=U I=1
Modality	Flow, Image, Other	S=1 I=1	S=4 I=4
Container File Level	The level indicates whether the container is a series or an instance.	S=0 I=0	I=1
File name	This is the value of the file name.	S=0 I=0	S=1 I=1
File extension	This is the value of the file extension.	S=0 I=0	S=1 I=1
File URI	This is the complete path to the file	S=1 I=1	S=0 I=1
UID	This is a DICOM UID which is a number that uniquely describes an object.	S=0 S=0	S=1 I=1
One level higher URI	For a series and an instance, these are respectively a study and a series.	S=0 I=1	S=1 I=0
File Content	This is the type of the file, which in both cases is ZIP	S=0 I=0	S=1 I=1
Description	String for information that has not been provided in any of the other header elements	S=0 I=0	S=1 I=1
Anonymized	Data in order to be anonymized must have all information that could serve to identify the patient removed.	S=0 I=0	S=1 I=1

Abbreviations: S=Series, I= Instance, U= unbounded

Table 2 Present Only in the Instance Header

Item	Description and Comments	minOccurs	maxOccurs
Instance Number	The number of the instance in the series.	I=1	1=1
Series Information	The location, series number in the Study_File, a short description, and limited information about the Series Container file	I=1	I=1
Role	Series member, Single measurement, Control, Variable (Value and Name), Modified data, or Other.	I=1	I=1

Table 3 Present Only in the Series Header

Item	Description and Comments	minOccurs	maxOccurs
Number of Instances	Total number of instance files, 1 to 64K	1	1

CytometryML Container ToC

Table 4: Present in the Series and Instance Table of Contents

Item Pointed to	Description and Comments	Element Type	minOccurs	maxOccurs
Specimen	The metadata in a specimen file includes: the type (human, veterinary, environmental, or other), the accession number, UID, field (biological research, medical, environmental, or other), an identifier, the organ and its part, and the container.	specimen:Specimen_Type	S=1 I=1	S=1 I=1
Compensation	The present Gating-ML schema includes compensation	instance:File_Format_Modifiable_Instance_Data_Type	S=0 I=0	S=1 I=1
Gating	The present Gating-ML schema includes compensation	instance:File_Format_Modifiable_Instance_Data_Type	S=0 I=1	S=1 I=1
Analysis	This is a document that needs to be created	instance:File_Format_Modifiable_Instance_Data_Type	S=0 I=1	S=0 I=1
Test Definition	The steps taken to prepare (stain or tag) the specimen. The preparation could be performed in 2 parts. One at the series level and a subsequent one at the instance level.	other:XML_PDF_or_Other_File_Type	S=0 I=0	S=1 I=1
Result File	This is a document that needs to be created.	other:XML_PDF_or_Other_File_Type	S=0	S=1
Other File	This is a place holder for the omissions that will inevitably occur in a design.	other:XML_PDF_or_Other_File_Type	S=0 I=0	S=U I=U
Additional Information	This is a place holder.	anyType	S=0 I=0	S=U I=U

Abbreviations: S=Series, I= Instance, U= unbounded

Abbreviations: S=Series, I= Instance, U= unbounded

Table 5: Present Only in the Instance Only Table of Contents

Item Pointed to	Description and Comments	Element Type	minOccurs	maxOccurs
Binary Data	The binary data can be either from an image or a list-mode (waveform) file	toc:Binary_Data_Info_Type. Both the list-mode files and image files can be in multiple formats	1	100
List Mode	The metadata in a list-mode file describes: the acquisition context, binary format, the status of the metadata contained in the binary file, the trigger parameter, the location of the channels and their numeric data formats, and the locations of the index files.	list:List_Mode_File_Type	0	1
Image Context	This is the metadata that describes an image file including its format, number of layers, compression, etc.	context:Image_Context_Type	0	1
Procedure Name	This is an enumerated type that presently includes: Immunophenotyping, DNA histogram, viability, apoptosis, counting, rare cell detection, and other.	instance:Procedure_Name_Type	1	1
Channels Settings	This is the part of the instrument settings including the order of the optical components that can be varied for each run and their detectors. Each channel corresponds to an individual parameter.	channels:Channel_Reference_Sequence_Type	1	1

The files described in a ToC can only belong to one series. The sum of the number of list mode and image binary files has been allowed to be 100, which is high and would be a very BAD practice for clinical data, because it could cause confusion in correlating the binary data with the XML metadata.

Abbreviations: S=Series, I= Instance, U= unbounded

Table 6: Present Only in the Series Table of Contents

Item Pointed to	Description and Comments	Element Type	minOccurs	maxOccurs
Instrument Information	This provides an abbreviated description of the instrument includes its modality (flow, image, etc.), name, and manufacturer.	instr:Instrument_Header_Type	1	1
Flow Cytometer Description	This is the metadata for the flow cytometer. This metadata includes the optical and fluidics parts, electronics, software and settings that are constant for the entire series.	flow:Flow_Specific_Info_Type	0	1
Microscope Description	This metadata includes the optical and positioning parts, electronics, software and settings that are constant for the entire series.	micro:Microscope_Specific_Info_Type	0	1
Instance Container File Information	These elements describe the location of an instance file containers and includes a list of the metadata and binary files in each container.	instance:Instance_Container_File_Info_Type	1	1
Series Overview	This is a file that describes the series in its entirety.	other:XML_PDF_or_Other_File_Type	0	1

There is only one instrument, which can be either a flow cytometer or a microscope. The flow cytometer schema, instrument.flow.xsd, will be extended to include image capture.

Table 7: Metadata Required Elements

Item	Required	Total	Percent
Series Header	4	16	25%
Series ToC	4	13	31%
Series Total	8	29	28%
Instance Header	7	16	44%
Instance ToC	7	12	58%
Instance Total	14	28	50%
Metadata Total	22	57	39%

At a minimum, 8 elements are needed to describe the contents of a series container and 14 are needed for an instance container. This is subject to revision by a consensus.

Schema Creation

- Reuse: DICOM, ISAC FCS, ECMA, HR-XML
- Simple object based design methodology
- Schemas = classes or packages O-O languages.
- Size kept manageable by budding off new schema(s).
- Acyclic graph. No schema was imported by any that it imports.
- All data-types that include elements based on other data-types are located below those used.

Conclusions

- It is possible to translate much of DICOM into XSDL and then into XML.
- The XML schemas can be used with other applications including Microsoft® Office.
- The DICOM design and documentation can be reused with other syntaxes, such as XML.
- DICOM can and should be extended in XML.
- Sometime in the future, DICOM could evolve into an XML based standard.

Acknowledgements

This work was sponsored by Newport Instruments. The author wishes to thank the members of the ISAC DSTF and members of DICOM Working Groups 26 and 27 for providing knowledge respectively of the ISAC ACS and of DICOM. However, any mistakes concerning either the ACS or DICOM are the author's own. The views expressed are solely those of the author and need not be that of any group of which he is a member.

Some relevant publications

- [1] R. C. Leif, "Toward the integration of cytomics and medicine," *J. Biophoton.* 2, 482-493 (2009).
- [2] J. Spidlen, R. C. Leif, W. Moore, M. Roederer, International Society for the Advancement of Cytometry Data Standards Task Force, R. R. Brinkman, "Gating-ML: XML-based gating descriptions in flow cytometry," *Cytometry Part A* **73A**, 1151-1157 (2008) and <http://flowcyt.sourceforge.net/gating/latest.pdf>.
- [3] Leif R. C., Spidlen J., Brinkman R. R., "Cytometry standards continuum," *Proc. SPIE* **6859**, 68590Q-1-8 (2008).
- [4] J. Lee, J. Spidlen, K. Boyce, J. Cai, N. Crosbie, M. Dalphin, J. Furlong, M. Gasparetto, E. M. Goralczyk, B. Hyun, K. Jansen, T. Kollmann, M. Kong, R. C. Leif, S. K. McWeeney, T. D. Moloshok, W. Moore, G. Nolan, J. Nolan, J. Nikolich-Zugich, D. Parrish, B. Purcell, Y. Qian, B. Selvaraj, C. Smith, O. Tchuvatkina, A. Wertheimer, P. Wilkinson, C. Wilson, J. Wood, R. Zigon, R. Scheuermann, and R. R. Brinkman, "MIFlowCyt: the minimum information about a Flow Cytometry Experiment," *Cytometry Part A.* **73A**, 926-930 (2008)
- [5] R. C. Leif, "An XML Cytometry Standard Based on DICOM," *Proc. SPIE* **7264**, 72640H (2009).
- [6] J. Spidlen, R. Brinkman, R. C. Leif, 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)," Flow Informatics and Computational Cytometry Society (FICCS), Available at: <http://wiki.ficcs.org/ficcs/Requirements?action=AttachFile&do=get&target=Requirementsv070920.pdf> (2007).
- [7] R. C. Leif, "CytometryML, Binary Data Standards," *Proc. SPIE* **5699**, 325-333 (2005).
- [8] CytometryML schemas. Available at: <http://www.newportinstruments.com/cytometryml/cytometryml.htm>
- [9] R. C. Leif, J. Spidlen, R. R. Brinkman, "A Container for the Advanced Cytometry Standard (ACS)," *Proc. SPIE* **7182**, 71821Q (2009).
- [10] XML Schema Part 1: Structures Second Edition W3C Recommendation 28 October 2004. Available at: <http://www.w3.org/TR/xmlschema-1/> (2004).
- [11] XML Schema Part 2: Datatypes Second Edition, W3C Recommendation 28 October 2004 Available at: <http://www.w3.org/TR/2004/RECxmlschema-2-20041028/> (2004).
- [12] P. Warmlesley, *Definitive XML Schema*, Prentice Hall, Available at: <http://www.phptr.com> (2002).
- [13] O. S. Pinykh, *Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival-Guide*, Springer Publishers, Berlin & Heidelberg, ISBN 354074570X, 9783540745709 (2008).

- [14] DICOM Supplement 148: Web Access to DICOM persistent Objects by means of Web Services, Extension of the Retrieve Service (WADO Web Service), working draft in progress (2009).
- [15] Extensible Markup Language (XML) 1.0 (Fifth Edition), W3C Recommendation 26 November 2008, World Wide Consortium(W3C[®]), Available at: <http://www.w3.org/TR/2008/REC-xml-20081126/> (2008)
- [16] D. L. Parnas, "On the Criteria To Be Used in Decomposing Systems into Modules," *Communications of the ACM* **15**, 1053-1058 (1972).
- [17] D. L. Parnas, P. C. Clements, D. M. Weiss, "Enhancing reusability with information hiding," ITT Proceeding of the Workshop on Reusability in Programming, cse.msu.edu. (1983).
- [18] B. W. Boehm, "A spiral model of software development and enhancement," *IEEE Computer* **21**, 61-72 (1988).