Requirements for a data file standard format to describe cytometry and related analytical cytology data

Version 0.070920
September 20, 2007

Abstract:

The flow cytometry data file standard (FCS) provides a specification to describe flow cytometry data. In 1984, the first Flow Cytometry Standard format for data files was adopted as FCS1.0, this standard was modified in 1990 as FCS2.0, and in 2004 as FCS3.0. This standard has kept pace with many years of technological evolution. Even though it is supported by virtually all analytical instrument and third party software suppliers, we believe that there are significant reasons for an update, especially considering the capabilities of modern (and future) instruments and use cases involving third party interpretation of flow cytometry data. This document summarizes the requirements that should optimally be met by a data file standard format for describing cytometry and related analytical cytology data.

Keywords: standard, flow cytometry, digital microscopy, analytical cytology, data file

Josef Spidlen (jspidlen@bccrc.ca)
Ryan Brinkman (rbrinkman@bccrc.ca)
Robert Leif (rleif@rleif.com)
and other members of the ISAC Data Standards Task Force

September 20, 2007

CONTENTS

1. Introduction

1.1 Overview

First developed in 1984, the Flow Cytometry Standard (FCS) [B1] specification has kept pace with many years of technological evolution. FCS is the common representation of flow cytometry data, and this standard is supported by all analytical instrument and third party software suppliers. Scientists can choose among instruments and software with no major compatibility issues for the raw fluorescence values that FCS captures. However, there are reasons as indicated in [B2], [B3] for updating the current version of the FCS standard or for adopting a new data file standard format to describe cytometry and related analytical cytology data. The development of an optimal data file standard based on these requirements may be challenging as the requirements partially contradict each other. This document summarizes the requirements of a data file standard format for describing flow and image cytometry and related analytical cytology data.

1.2 Terminology within this document

The key words “shall”, “should”, and “may” in this document are to be interpreted as described in RFC 2119 [B4] and are also compatible with the IEEE Standards Style Manual [B5].

The word shall is used to indicate mandatory requirements to be followed in order to conform to the standard and from which no deviation is permitted (shall equals is required to).

The word should is used to indicate that among several possibilities one is recommended as particularly suitable, without mentioning or excluding others; or that a certain course of action is preferred but not necessarily required; or that (in the negative form) a certain course of action is deprecated but not prohibited (should equals is recommended that).

The word may is used to indicate a course of action permissible within the limits of the standard (may equals is permitted to).

The use of the word “relevant” in this document describes the condition by which information should be provided: “relevant” information is information that is necessary for correct understanding of the context of an experiment component.

A semantic model is a machine-interpretable representation used to model an area of knowledge or some part of the world, including software, hardware, biology, etc. It provides a layer of abstraction, allowing to work formally at a higher level with underlying information and logic. Typical examples of such models are ontologies or any logic-based representations. Typical examples of such models are ontologies or any logic-based representations. Depending upon the framework or language used for modeling, different terminologies exist for denoting the building blocks of semantic models [B6].

2. Requirements

2.1 Encode the required information to interpret a cytometry experiment

2.1.1 Description

A cytometry standard shall provide a mechanism to record all the essential information for a person knowledgeable in the field to interpret a cytometry experiment.

2.1.2 Rationale

Prior to the development of a cytometry file(s) standard, the minimum required information to interpret a cytometry experiment shall be determined (e.g., [B7]). The standard(s) shall be developed to encode this information in electronic format. Such a format is needed to facilitate exchange of the minimum information between laboratory information management systems, databases and data analysis tools.

2.2 Facilitate the reproduction of a cytometry experiment

2.2.1 Description

A cytometry standard should provide a mechanism to record all the essential information to reproduce a cytometry experiment.

2.2.2 Rationale

Modern experimental science is based on the concept of reproducibility. An experimental result can not be considered to be reliably correct until it is reproduced by a second laboratory. This is one of the reasons that scientific journals have material and methods sections. There has been a historical conflict between completeness of the description of the experiment and the cost of publishing. In the digital era, this is definitely not the case. Many journals have online archives for material, such as the detailed experimental procedures and even some of the results (supplementary material).

2.3 Efficiently store cytometry list mode data

2.3.1 Description

A cytometry data file standard should provide an efficient way to store multi-parametrical list mode data and images. For example, binary information shall be encoded in a file format suitable for binary data.

2.3.2 Rationale

List mode data presently represent the primary output of flow cytometry analytical instruments and images presently represent the primary output of digital microscopes. There are also new analytical cytology instruments producing both list mode data and images.

2.4 Transparently store text-based data

2.4.1 Description

Text-based information shall be encoded in a file format suitable for text data.

2.4.2 Rationale

Encoding textual data in a plain format will facilitate access.

2.5 Facilitate support for other analytical cytology data

2.5.1 Description

Use a common (shared) mechanism/approach/methodology/terminology to store/analyze/transport data and metadata from flow cytometry as well as image and other analytical cytology technologies. (This does not imply that an initial implementation is required to support all types of cytometry data. An initial implementation could be directed to only one modality, such as flow cytometry).

2.5.2 Rationale

Flow cytometry and digital microscopy represent complimentary technologies used in analytical cytology. Any kind of shared approach has the potential to reduce development costs and facilitate data integration, which is to the benefit of both**,** developers and end-users.

2.6 Each type of information shall be uniquely identifiable

2.6.1 Description

There shall be a way to identify each type of information by a globally unique identifier. If there is an identifier assigned to an instance of information, this instance should not change. (Changes may be done via creating a new instance that receives a new identifier).

2.6.2 Rationale

A globally unique identifier is useful for unambiguous reference purposes. Any type of information (resource) may eventually be needed to be referenced from another resource. A resource may depend on another resource; thus one should not change resources that have identifiers (and thus can eventually be referenced from another resource). For example, if gating is performed using compensated parameters then a gating description resource depends on the compensation description resource. Compensation description shall not be altered as it would invalidate/alter the gating. This does not imply that each type of information must be uniquely identifiable in every setting, only that it should be possible to do so if desired.

2.7 It shall be possible to resolve data files over the Internet

2.7.1 Description

There shall be a relatively simple mechanism to resolve the globally unique identifier to obtain the data. This does not imply than anyone can always access any data file (private files, authorization restrictions; lack of internet access in remote areas), only that the specification shall support this in a clearly described manner should technical and institutional conditions favour data exchange.

2.7.2 Rationale

A resolvable mechanism (ideally via the HTTP protocol) enables/simplifies review/study of third party data.

2.8 Each type of information shall only be stored in one file format

2.8.1 Description

There shall be no competing standard ways to store the same type of information, for example, if compensation description is stored in an XML-based file, it shall not be stored as part of an FCS data file.

2.8.2 Rationale

Competing ways of storing the same type of information (i) raise the complexity of the standard unnecessary, and (ii) may eventually raise questions which information is correct if these are not quite the same.

2.9 Support for future instruments

2.9.1 Description

The standard shall be developed considering future instrument rather than restricting/bound to current hardware configurations. It shall be designed in an open way so that even unanticipated technologies can use the standard for data storage.

2.9.2 Rationale

A new standard is unlikely to be supported by the current generation of instruments. Along with technological advancements, there are new analytical instruments that have difficulties to utilize the current FCS standard to store their data, for example instruments combining digital microscopy with flow cytometry or instruments producing a detailed digital representation of the shape of the detected signal (instead of just height, width, or area of the signal).

2.10 Provide semantic meaning to all required/standardized information

2.10.1 Description

The standard shall sufficiently describe all the pieces of required and standardized information.

2.10.2 Rationale

Detailed semantic shall be provided to prevent potential misinterpretations and misusages of the standard.

2.11 Extensible by third parties

2.11.1 Description

Extensions to the standard shall be easily possible and shall not alter the conformance to the standard. These includes vendor-specific information as well as incorporation of additional information described by third-party standards (such as standards used in the clinical domain). There should be a standardized methodology for these extensions and, it shall be encouraged (not required) to take place through ISAC. Where possible, extensions should be based on, inherit or reused from other standards stating the normative source in these cases.

2.11.2 Rationale

Vendors desire a way to add proprietary or new metadata. Different environments will need to store additional information. It is essential to implement this in a coordinated way to simplify its interpretation (or discounting) by third parties.

2.12 Extensions shall be separated and removable from normative parts

2.12.1 Description

Additional information and extensions to the standard shall be easily possible; however; this additional information shall be clearly separated from normative parts. In data files, it shall be easy to remove any extensions from the normative section. There shall be no direct link from normative parts to extensions (as this would invalidate the normative part if extensions are removed, there may be a link from extensions to the normative parts if needed.

2.12.2 Rationale

Being able to easily separate extensions from normative parts in instance documents (i.e., data files) allows for “annonymization” of clinical data for research purposes as well as for removal of any kind of sensitive information prior publicizing data. Also, clear separation of standardized versus additional information simplifies robustness and independent software development.

2.13 Ensure that metadata can always be located

2.13.1 Description

There shall be a way/methodology to connect components of the standard so that these do not get separated.

2.13.2 Rationale

A mechanism that links data components is essential if they are in separate formats.

2.14 Support for use cases where new metadata are being added subsequently

2.14.1 Description

There shall be a way/methodology to add new metadata post acquisition.

2.14.2 Rationale

Performing analysis and/or providing details off line is a common use case.

2.15 Several instances of mutable information may share immutable information

2.15.1 Description

Support for use cases where several instances of potentially conflicting mutable information refer to the same information.

2.15.2 Rationale

The information represented within the standard (or some of it) may be subject of human errors and/or different opinions (for example gating or compensation). As such, there may be more than one instance of such information describing the same primary data. These may not necessary agree with each other.

2.16 Some information may be encoded by a referenced semantic model or an existing standard

2.16.1 Description

There shall be a way to individually indicate a specific piece of information is encoded by a semantic model (e.g., an ontology) or an existing standard.

2.16.2 Rationale

Semantic models and existing standards may provide addition useful knowledge that may be generally interpretable (within the context of the semantic model). As such, it has the potential to provide additional knowledge via implicit or explicit relationships, description logic and reasoning mechanisms.

2.17 It shall be possible to refer to multiple semantic models or standards

2.17.1 Description

There shall be a way to reference several semantic models or existing standards from a single instance of a data standard compliant data file. This may be done for example via the concept of a foreign coding scheme.

2.17.2 Rationale

Different semantic models and standards encode different information and multiple semantic models may be useful to reference from a single instance of the data standard compliant data file. The connection of multiple semantic models including the use of semantic web technologies in general may provide even more additional information.

2.18 Use a manufacturer-independent format

2.18.1 Description

The standard shall not be based/dependent on any particular manufacture implementation and shall not prioritize any particular hardware/software.

2.18.2 Rationale

An equal chance shall be given to anyone who wants to implement/support the standard.

2.19 Use a programming language-independent format

2.19.1 Description

The standard shall not be based/dependent on any particular programming language.

2.19.2 Rationale

The choice of programming language shall be left to implementors.

2.20 Use a file/operating system-independent format

2.20.1 Description

The standard shall not be based/dependent on any particular file- or operating-system.

2.20.2 Rationale

The choice of the file system and operating system shall be left on the end users' preferences. Also, the data file format needs to be serialize-able into a byte stream for serial digital transportation.

2.21 Use a simple format oriented on data interoperability

2.21.1 Description

The format shall be as simple as possible to support interoperability, unnecessary options shall be avoided.

2.21.2 Rationale

Simple formats/standards avoiding unnecessary options are easier to adopt. Unnecessary options significantly increase the complexity and thus the development costs for software vendors.

2.22 Describe the standard unambiguously

2.22.1 Description

The standard shall be described in an extensively explicit way avoiding any ambiguity in its interpretation.

2.22.2 Rationale

The standard shall be clear and very explicit so that any potential ambiguity is avoided to prevent incompatible interpretations resulting in non-compatible data formats. Adopting best practices from international standardization bodies increases the chance of creating a useful standard.

2.23 There shall be an open-review period before adopting the standard

2.23.1 Description

The proposal shall be opened for wide community review without restrictions (including restrictions to membership of a certain organization) prior its adoption as a standard.

2.23.2 Rationale

It is important to receive as much feedback as possible during the standard development process in order to ensure broad adoption. No stakeholders shall be discriminated and/or left out of this process.

2.24 There shall be a mechanism to formally validate the standard

2.24.1 Description

There shall be a procedure/mechanism to validate/verify the new standard. Especially, if normative parts of the standard are formally specified reusing other standard formalisms such as XML schemas, RDF schemas, UML models, etc., these shall be formally validated using appropriate tools.

2.24.2 Rationale

A formal validation is necessary to ensure the quality of the standard.

2.25 The standard shall be public

2.25.1 Description

All components of the final version of the standard shall be publicly available

2.25.2 Rationale

Public availability supports broad adoption of the standard.

2.26 An implementation of the standard shall be possible without charge

2.26.1 Description

There should be no cost for developing an implementation of the standard.

2.26.2 Rationale

Implementation fees may hinder the adoption of the standard. Although a reference implementation of a standard should be created and maintained, it is anticompetitive to raise barriers to the development of competing versions.

2.27 The implementation of the standard shall be non-restrictive

2.27.1 Description

There shall be no licensing restriction on the implementation of the standard.

2.27.2 Rationale

Licensing restrictions would hinder the adoption of the standard.

2.28 Reuse existing standards

2.28.1 Description

For areas that are outside of our expertise, the standard shall reuse, adopt, interface, or translate other broadly accepted standards that are created by experts in those areas.

2.28.2 Rationale

Reusing existing standards and technologies significantly lowers development costs (for both the standard itself and for standard compliant products), increases interoperability with other standards, prevents competing solutions, and generally highly increases the quality of the resulting standard. Methodology/templates/terminology shall be reused from professional standardization bodies such as CEN, DICOM, HL7, IEEE, IETF, ISO, SNOMED, W3C, and/or others as appropriate. Interoperability and/or joint development with existing medical standards will significantly increase the scope of new standards developed by ISAC.

2.29 Conformance to the standard shall be (easily) testable

2.29.1 Description

There shall be a relatively simple way to verify (automatically by software) whether a data file conforms to the data file standard.

2.29.2 Rationale

Being able to test conformance with the standard is essential to develop and validate standard-compliant software and hardware.

2.30 The file format shall be as self-explanatory as possible

2.30.1 Description

The file format shall be designed in a way that is as self-explanatory as possible; files should be easy to read and understand. The readability of information should be favored over file size. The hierarchy/structure should be easily understandable and elements/parts/pieces of the file standard should be named in a meaningful way.

2.30.2 Rationale

A simple and easily readable standard minimizes compliance issues. Developers tend not to read specifications carefully. Simple structure and meaningful names simplifies development process (including any debugging in case some compliance issues arise).

Bibliography

[B1] Seamer L.C., Bagwell C.B., Barden L., Redelman D., Salzman G.C., Wood J.C.S., Murphy R.F.: “Proposed New Data File Standard for Flow Cytometry, Version FCS 3.0”, Cytometry 28 (1997), Wiley-Liss, Inc., pp. 118-122, http://murphylab.web.cmu.edu/publications/64-seamer1997.pdf.

[B2] Bioinformatics Standards for Flow Cytometry Consortium: Rationale for changes to FCS 3.0; http://flowcyt.sourceforge.net/acs/.

[B3] 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 pp. 56-65 (2003).

[B4] Bradner S., The Internet Engineering Task Force: Request for Comment 2119: Key words for use in RFCs to Indicate Requirement Levels. March 1997, http://www.ietf.org/rfc/rfc2119.txt

[B5] The Institute of Electrical and Electronics Engineers: 2007 IEEE Standards Style Manual, http://standards.ieee.org/guides/style/2007_Style_Manual.pdf

[B6] Semantic Annotations for WSDL and XML Schema W3C Recommendation 28 August 2007; http://www.w3.org/TR/2007/REC-sawsdl-20070828/

[B7] Bioinformatics Standards for Flow Cytometry Consortium: MIFlowCyt - Minimum Information about a Flow Cytometry Experiment; http://flowcyt.sourceforge.net/miflowcyt/, http://www.isac-net.org/content/view/594/46/.