

Analytical Cytometry Standard Using ZIP for the Container File Format

Proposal version 071016

October 16, 2007

Abstract

The Flow Cytometry Standard (FCS) specification has been adopted for the common representation of flow cytometry data, and this standard is supported by all analytical instrument and third party software suppliers. It includes data, metadata and analysis components within the same file. However, metadata and analysis components, if included at all, are not recorded in a standardized fashion or in sufficient detail for use by independent parties.

We are proposing to address these shortcomings via series of related data standard proposals. Different components describing analytical cytometry experiments would be stored reusing standard formats such as XML or RDF, and all the components would be bundled together into an Analytical Cytometry Standard (ACS) container.

Within this proposal, we suggest using ZIP as the file format for the analytical cytometry standard container. This approach has been adopted by other groups including for purposes such as Java JAR files, OpenDocument and Office Open XML formats, Google Earth KMZ files, Mozilla Firefox Add-ons, Nokia's mobile phone themes, WinAmp and Windows Media Player skins, Magic Draw UML files and many others.

Status of this document

This document is an unapproved draft of a proposed standard that is intended for an internal review by the International Society for Analytical Cytology (ISAC) Data Standards Task Force (ISAC DSTF). As such, this document is subject to change and must not be utilized for any conformance/compliance purposes.

Josef Spidlen, jspidlen@bccrc.ca
Ryan Brinkman, rbrinkman@bccrc.ca

Table of Contents

1.	Introduction	3
1.1.	Rationale for changes to FCS 3.0	3
1.2.	Scope of this Document	3
1.3.	What is ZIP	4
1.4.	Benefits of Using ZIP	5
1.4.1.	Reusing Existing Standards	5
1.4.2.	Well Established Wrapper Format	5
1.4.3.	OS Platform Independence.....	5
1.4.4.	Availability of Tools	5
1.4.5.	High Level Quality Documentation	5
1.4.6.	Open Standard	6
1.4.7.	Format Characteristics.....	6
1.4.8.	Performance	6
2.	Normative: Analytical Cytometry Standard, Container Format Specification	7
2.1.	Filename.....	7
2.2.	File Format.....	7
2.3.	Container Contents.....	7
	Appendix A – ZIP vs. Other Archive Formats	8
	Compression Efficiency and Speed	8
	References.....	8

1. Introduction

1.1. Rationale for changes to FCS 3.0

First developed in 1984, the Flow Cytometry Standard (FCS)¹ specification has kept pace with many years of technological evolution. It has been adopted for 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.

FCS includes data, metadata and analysis components within the same file. However, metadata and analysis components, if included at all, are not recorded in a standardized fashion or in sufficient detail for use by independent parties. This is assuming that independent parties can access experimental results, as important data sets supporting publications are almost invariably unavailable. Finally, if metadata annotation takes place at any time subsequent to data capture, the all-inclusive format of FCS necessitates the generation of a new version of the file, which replicates the (hopefully unmodified) primary data.

Changes proposed in this document and related proposals (e.g., Gating-ML², Transformation-ML³, ACS containers⁴) are designed to address these shortcomings through the incorporation of technologies from standards bodies such as the World Wide Web Consortium (W3C)⁵, Object Management Group (OMG)⁶ the Dublin Core Metadata Initiative⁷, and the Unidata Community⁸, that were not available when FCS was conceived. The current proposal assumes an Analytical Cytometry Standard (ACS) container for different components describing analytical cytometry experiments. The ACS container may include XML-based⁹ components (such as based on MathML¹⁰, RDF¹¹, XHTML¹², SVG¹³, XSD¹⁴, CytometryML¹⁵, and others), as well as other types of data and formats including images, text-based documents, vendor-specific information (Figure 1). The actual checklist of what should be described is contained in the Minimum Information for a Flow Cytometry Experiment (MIFlowCyt¹⁶).

Within this proposal, we suggest using ZIP¹⁷ as the file format for the analytical cytometry standard container. This approach has been shown very useful by many other groups who have adopted ZIP for purposes such as Java JAR files, OpenDocument and Office Open XML formats, Google Earth KMZ files, Mozilla Firefox Add-ons, Nokia's mobile phone themes, WinAmp and Windows Media Player skins, Magic Draw UML files and many others¹⁸.

The support of these standards by additional software tools, journals and scientists will bring flow cytometry data to the semantic web^{19, 20} and significantly facilitate the reproduction of experiments and clinical measurements. Most importantly, these changes will allow scientists and software agents to search, automatically process, and in particular understand both flow cytometry data and metadata.

1.2. Scope of this Document

This document represents a proposal for the file format of the Analytical Cytometry Standard (ACS) container. There are many components of the ACS Container as shown

in Figure 1. Within this document we only address the file format of the container (wrapper). The formats for the components of the container are addressed by separate proposals such as Gating-ML² and Transformation-ML³.

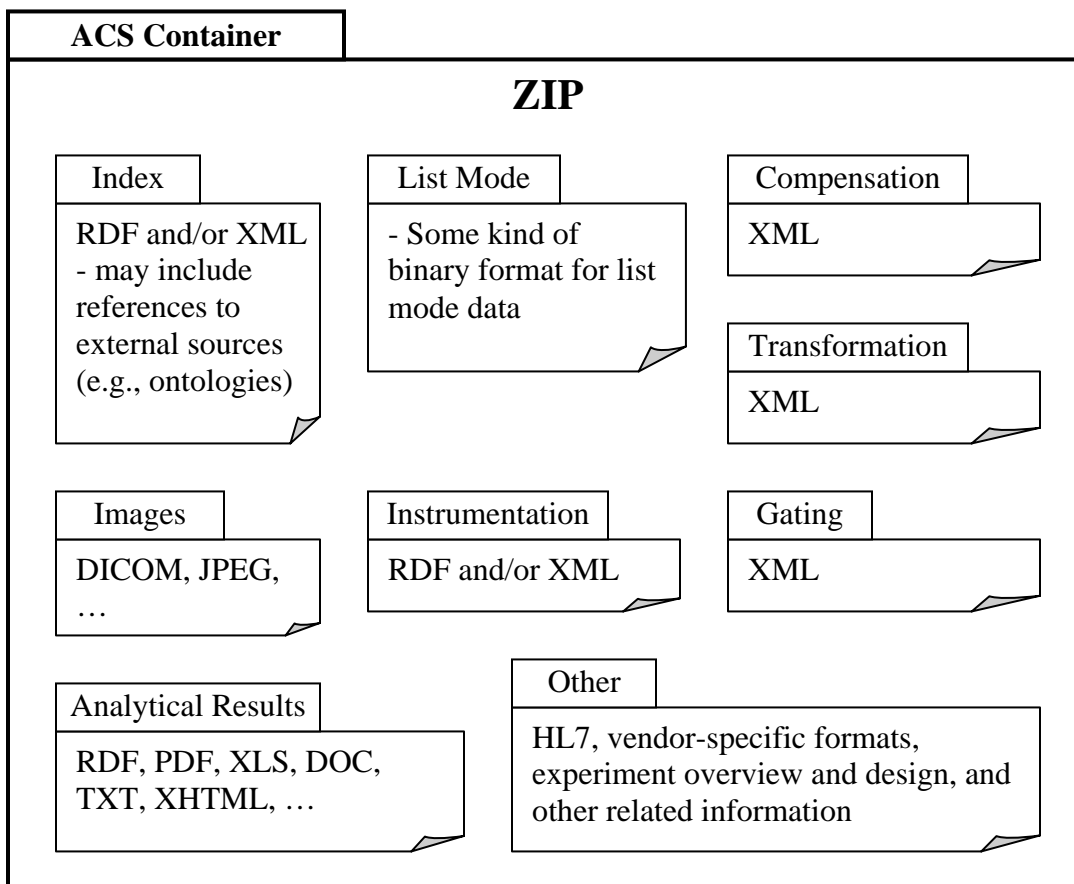


Figure 1 – Analytical Cytometry Standard. Overview of the different components of the proposed Analytical Cytometry Standard (ACS). The file format of the ACS container addressed by this proposal is highlighted in bold.

1.3. What is ZIP

The ZIP file format is a popular data compression and archival format. A ZIP file contains one or more files that have been compressed to reduce their file size, or stored as-is. The format was originally designed by Phil Katz for PKZIP. However, many software utilities other than PKZIP itself are now available to create, modify, or open (unzip, decompress) ZIP files, notably WinZip, BOMArchiveHelper, KGB Archiver, PicoZip, Info-ZIP, WinRAR, IZArc, 7-Zip, ALZip, TUGZip, Universal Extractor and Zip Genius. Microsoft has included built-in ZIP support (under the name “compressed folders”) in later versions of its Windows operating system. Apple has included built-in ZIP support in Mac OS X 10.3 and later via the BOMArchiveHelper utility.

ZIP files generally use the file extensions “.zip” or “.ZIP” and the MIME media type “application/zip”. Many software applications use ZIP as a container (wrapper) for several files in a certain structure. Generally, when this is done a different file extension

is used. Examples include Java JAR files, id Software .pk3/.pk4 files, package files for StepMania and Winamp/Windows Media Player skins, XPInstall, as well as OpenDocument and Office Open XML office formats, Google Earth KMZ files, Mozilla Firefox Add-ons, Nokia's mobile phone themes, or Magic Draw UML models.

1.4. Benefits of Using ZIP

1.4.1. Reusing Existing Standards

Reusing existing standards is one of the proposed requirements for a new cytometry data file standard. ZIP is a well matured open standard that has been adopted in many different fields over the past 18 years.

1.4.2. Well Established Wrapper Format

As enumerated in 1.3, many groups standardizing a container/wrapper format reused ZIP for their purposes even though there are alternative container file formats.

1.4.3. OS Platform Independence

Based on comparison²¹, ZIP is supported on all common operating system platforms including Windows, MS-DOS, PC-DOS, Mac OS X, Linux, BSD, Unix, and AmigaOS.

1.4.4. Availability of Tools

The variety of available software tools shows the popularity of the ZIP format. Based on comparison²¹, ZIP is the mostly supported archive format. End user software supporting reading as well as writing the ZIP file format includes the following: 7-Zip, Alpha ZIP, ALZip, Basic Zip, Beezer, BitZipper, BOMArchiveHelper, Filzip, Info-ZIP, IZArc, jzip, KGB Archiver, PeaZip, PKZIP, PowerArchiver, Squeez, StuffIt, StuffIt Expander, The Extractor, The Unarchiver, TUGZip, WinAce, WinRAR, WinRK, WinZip, XAD, ZipGenius, Zipeg, ZipZag (based on comparison²¹). Software libraries supporting the ZIP file format are available for virtually all programming languages. It is possible to create and extract zip-formatted archives with open source software, e.g., Info-ZIP²².

1.4.5. High Level Quality Documentation

The standard is well described¹⁷. In order to ensure the interoperability of the .ZIP file format by third party development organizations, PKWARE remains committed to the periodic publication of the “Application Note on .ZIP File Format Specification” specification. The publication of changes and revisions to this specification occurs on the following schedule:

Notification of Change: When feature changes affecting the .ZIP file format are identified, a preliminary release of this information will be made within 120 days following the general release of the first product in which these changes are implemented. This preliminary information will identify areas of current, or planned, changes to allow third party developers to avoid any conflicts with their development activities.

Final Publication: Final publication of feature changes to the specification will occur within 9 months of the general release of the last product in which these changes are

implemented. The availability of the final release of the specification will be timed to ensure the changes work reliably and conflict-free across all supported operating environments.

Changes made to the ZIP file format have always been backwards compatible for the past 18 years of existence of the file format.

1.4.6. Open Standard

ZIP format represents an open standard with the exception of features described in Section XIV (the "Strong Encryption Specification") that are covered by a pending patent application. Portions of the Strong Encryption technology are available for use at no charge under certain terms and conditions. See License Agreement of the Application Note on .ZIP File Format Specification¹⁷. However, encryption of ACS containers is an institution-specific requirement and not within the scope of ISAC's DSTF.

1.4.7. Format Characteristics

ZIP archive format compresses every file separately, which allows individual retrieval of files without reading through other data.

Files can be stored either uncompressed or by one of the supported compression algorithms (See the Application Note on .ZIP File Format Specification¹⁷ section IV. - General Format of a .ZIP file, subsection A - Local File Header / compression method). In practice, ZIP is almost always used with Katz's DEFLATE algorithm (see section IX. Deflating - Method 8, and section X. Enhanced Deflating - Method 9 of the Application Note on .ZIP File Format Specification¹⁷), except when files being added are already compressed or are resistant to compression.

Using the "ZIP64" format extensions available since version 4.5 (current version is 6.3.2) there are no limits to the uncompressed size of a file, compressed size of a file and total size of the archive. Since version 6.3.0, the ZIP specification contains a provision to store file names using UTF-8, adding Unicode compatibility to ZIP.

1.4.8. Performance

The ZIP format represents a fairly fast and effective compression algorithm / file format (see Appendix A – ZIP vs. Other Archive Formats).

2. Normative: Analytical Cytometry Standard, Container Format Specification

The key words “shall”, “should”, and “may” in this document are to be interpreted as described in RFC 2119²³.

2.1. Filename

The ACS Container files shall have the file name extension “.acs”.

Rationale

Reusing the ZIP file format while changing and standardizing a new file extension is a common methodology that has the advantage of using the common ZIP format while preventing users in exigently altering the contents of the container. It also expresses the semantic of the file format. This methodology has been successfully used by others (e.g., Java JAR files, Magic Draw MDZIP files, Google KMZ files, Mozilla Firefox XPI files, Nokia’s NTH files¹⁸).

2.2. File Format

The ACS file format shall correspond to the ZIP file format as specified by PKWARE’s “Application Note on the .ZIP file format” version 6.3.2 or later.

<http://www.pkware.com/documents/casestudies/APPNOTE.TXT>.

In order to ensure the continued interoperability of the .ZIP file format for all users, PKWARE publishes an Application Note on the .ZIP file format. The APPNOTE provides developers a general description and technical details of the .ZIP specification. This specification is periodically updated to include new features and information for features being introduced in anticipation of the emerging functionality requirements of the ZIP community.

Rationale

See section 1.4.

2.3. Container Contents

The content of the ACS container is not addressed/standardized by this proposal. Please see proposals for some of the components, e.g., Gating-ML² and Transformation-ML³. Additional components are expected to be standardized in the future²⁴.

Rationale

Specifying components of the container in separate documents increases the flexibility of the resulting standards and allows for a stepwise development process.

Appendix A – ZIP vs. Other Archive Formats

Compression Efficiency and Speed

Table 1 compares ZIP against other file formats. Compression effectiveness and speed are compared.

Table 1 – Comparison of ZIP vs. other archive formats²¹. The table is based on study²⁵ where ZIP and other file formats has been tested to compress text-based data, executables and raw images. For ZIP, version 2.3 from Info-ZIP²² has been used with the option -9 (maximum compression). See the reference²⁵ to find out about detailed settings of the test system and software used. The percentages in the table are the percentages of the size of the uncompressed file, i.e, the lower the percentage, the better compression. The time indicates time needed for compression on the testing system in seconds.

Name	Text		Executables		Images	
ZIP	25%	4.3s	39%	23.3s	60%	5.7s
7-zip	19%	18.8s	27%	59.6s	50%	36.4s
bzip2	20%	4.7s	37%	32.8s	51%	20.0s
rar	23%	30.0s	36%	275.4s	58%	52.7s
advzip	24%	21.1s	37%	70.6s	57%	41.6s
gzip	25%	4.2s	39%	23.1s	60%	5.4s
lha	27%	3.7s	40%	13.2s	61%	9.3s

References

1. Seamer LC, Bagwell CB, Barden L, et al. Proposed new data file standard for flow cytometry, version FCS 3.0. Cytometry. 1997;28:118-122.
2. Spidlen J, Brinkman RR, Bioinformatics Standards for Flow Cytometry Consortium. Gating-ML: Draft Standard for Gating Description in Flow Cytometry. Available at: <https://sourceforge.net/projects/flowcyt>. Accessed 07/31, 2007.
3. Spidlen J, Brinkman RR, Bioinformatics Standards for Flow Cytometry Consortium. Transformation-ML: Draft Standard for Transformation Description in Flow Cytometry. Available at: <https://sourceforge.net/projects/flowcyt>. Accessed 07/31, 2007.
4. Spidlen J, Brinkman RR, Bioinformatics Standards for Flow Cytometry Consortium. A Proposal for the Analytical Cytometry Standard. Available at: <http://flowcyt.sourceforge.net/acs/>. Accessed 07/31, 2007.
5. World Wide Web Consortium. Available at: <http://www.w3.org/>.
6. The Object Management Group (OMG). Available at: <http://www.omg.org/>.
7. Dublin Core Meta Data Initiative. Available at: <http://dublincore.org/>.
8. Unidata. Available at: <http://www.unidata.ucar.edu/>.

9. World Wide Web Consortium (W3C). Extensible Markup Language (XML). Available at: <http://www.w3.org/TR/REC-xml/>.
10. W3C Math Working Group. Mathematical Markup Language (MathML) Version 2.0 (Second Edition), W3C Recommendation. Available at: <http://www.w3.org/TR/MathML2/>. Accessed 07/31, 2007.
11. W3C - RDF Core Working Group. Resource Description Framework (RDF). Available at: <http://www.w3.org/RDF/>.
12. W3C. XHTML 1.0 - The Extensible HyperText Markup Language (Second Edition). Available at: <http://www.w3.org/TR/xhtml1/>.
13. W3C. Scalable Vector Graphics (SVG). Available at: <http://www.w3.org/Graphics/SVG/>.
14. W3C. XML Schema. Available at: <http://www.w3.org/XML/Schema>.
15. Leif RC, Leif SB, Leif SH. CytometryML, an XML format based on DICOM and FCS for analytical cytology data. *Cytometry*. 2003;54A:56-65.
16. Lee J, Spidlen J, Boyce K, et al. MIFlowCyt: Minimum Information for a Flow Cytometry Experiment. Available at: <http://flowcyt.sourceforge.net/miflowcyt/>. Accessed 07/31, 2007.
17. PKWARE. Application Note on the .ZIP file format. Available at: http://www.pkware.com/business_and_developers/developer/appnote/.
18. ZIP file format. Available at: [http://en.wikipedia.org/wiki/ZIP_\(file_format\)](http://en.wikipedia.org/wiki/ZIP_(file_format)).
19. Shadbolt N, Berners-Lee T, Hall W. The semantic web revisited. *IEEE Intelligent Systems*. 2006;21:96-5.
20. Wang X, Gorlitsky R, Almeida JS. From XML to RDF: How semantic web technologies will change the design of 'omic' standards. *Nat Biotechnol*. 2005;23:1099-103.
21. Comparison of file archivers. Available at: http://en.wikipedia.org/wiki/Comparison_of_file_archivers.
22. Info-ZIP. Available at: <http://www.info-zip.org/>.
23. Bradner S., The Internet Engineering Task Force. Request for Comments: 2119 - Key words for use in RFCs to Indicate Requirement Levels. Available at: <http://www.ietf.org/rfc/rfc2119.txt>. Accessed 07/31, 2007.
24. Spidlen J, Leif RC, Moore W, Brinkman RR. Analytical Cytometry Standard - NetCDF Conventions for List Mode Binary Data File Component Proposal. Available at: https://sourceforge.net/project/showfiles.php?group_id=175725&package_id=202340&release_id=443566.
25. Archiver comparison. Available at: <http://warp.povusers.org/ArchiverComparison/>.