 Cytometry 28:118–122 (1997)

Proposed New Data File Standard for Flow Cytometry, Version FCS 3.0

L.C. Seamer,[1]* C.B. Bagwell,[3] L. Barden,[4] D. Redelman,[5] G.C. Salzman,[2] J.C.S. Wood,[6] and R.F. Murphy[7]

[1]University of New Mexico, Cancer Research and Treatment Center, Albuquerque, New Mexico
[2]Los Alamos National Laboratory, Los Alamos, New Mexico
[3]Verity Software House, Topsham, Maine
[4]National Institutes of Health, Bethesda, Maryland
[5]Sierra Cytometry, Reno, Nevada
[6]Coulter Corporation, Hialeah Florida
[7]Carnegie Mellon University, Pittsburgh, Pennsylvania

Received 7 February 1997; Accepted 7 February 1997

In 1984, the first flow cytometry data file format was proposed as Flow Cytometry Standard 1.0 (FCS1.0). FCS 1.0 provided a uniform file format allowing data acquired on one computer to be correctly read and interpreted on other computers running a variety of operating systems. That standard was modified in 1990 and adopted by the Society of Analytical Cytology as FCS 2.0. Here, we report on an update of the FCS 2.0 standard which we propose to designate FCS 3.0. We have retained the basic four segment structure of earlier versions (HEADER, TEXT, DATA and ANALYSIS) in order to maintain analysis software compatibility, where possible. The changes described in this proposal include a method to collect files larger than 100 megabytes (not possible in earlier versions of the standard), the inclusion of international characters in the TEXT portions of the file, a method of verifying data integrity using a 16-bit cyclic redundancy check, and increased keyword support for cluster analysis and time acquisition. This report summarizes the work of the ISAC Data File Standards Committee. The complete and detailed FCS 3.0 standard is available through the ISAC office [Sherwood Group, 60 Revere Drive, Ste 500, Northbrook, IL 60062, phone: (847) 480-9080 ext. 231, fax: (847) 480-9282, E-mail: isac@sherwood-group.com] or through the internet at the ISAC WWW site, http://nucleus.immunol.washington.edu/ISAC.html. Cytometry 28:118–122, 1997. © 1997 Wiley-Liss, Inc.

Key terms: FCS 3.0; Data File Standard; flow cytometry

In 1984 Murphy and Chused introduced the notion of a standard flow cytometry data file format with the publication of Flow Cytometry Standard 1.0 (FCS 1.0) (1). The purpose of the data file standard was to provide a clearly defined and uniform file format that allowed data collected by one instrument to be correctly read for analysis by other software on another computer. The FCS 1.0 standard introduced the four-segment file structure that continues through the current proposal. These segments are HEADER, TEXT, DATA, and ANALYSIS, and are described in more detail below.

In 1990 the Society for Analytical Cytology (now called the International Society for Analytical Cytology) formed a Data File Standards committee to recommend a flow cytometry standard data file format. The committee used the FCS 1.0 file structure as the basis for an updated file standard that became known as FCS 2.0 (2). Several new concepts were introduced in the updated standard including *required* data descriptors called keyword-value pairs, guaranteeing the inclusion of the minimal information necessary to properly read file data. The second major change introduced in FCS 2.0 was support for creating multiple data-sets in a single file. A data-set is defined as all of the information relative to a flow cytometric measurement. FCS 2.0 also introduced many new FCS-standard defined keywords.

Since 1990, advances in the state-of-the-art in biotechnology, computer technology, and data communications have prompted the development of the next generation file format. The ISAC Data File Standards Committee has therefore proposed FCS 3.0. With FCS 3.0, we have

*Correspondence to: Larry C. Seamer, University of New Mexico, Cancer Research and Treatment Center, 915 Camino de Salud NE, Albuquerque, NM 87131.
e-mail: lseamer@cobra.unm.edu

attempted to address several limitations of the existing standard that have arisen since 1990. First, with the growing number of measurement parameters, high bit ADCs, the requirements of rare event detection, long kinetic files, and the inclusion of calculated parameters, list-mode data files now approach and occasionally reach the 100 megabyte file limit inherent in earlier versions. Second, previous versions of the FCS standard were limited to the 256 character ASCII text set. Flow cytometry is international in its scope; therefore data file text should be able to accommodate a broader range of international text characters. Third, computer network data transfer has become a common means of moving data. Therefore, a mechanism to detect errors and verify data integrity is needed. Fourth, as third party analysis software has found widespread use, the need has evolved to include more descriptive information regarding data acquisition and signal amplification. Finally, development of cluster analysis software to help identify and classify cell subsets has created a need for new keyword support in the FCS standard (4). The goal of the FCS 3.0 committee was to revise the FCS standard, addressing these needs while maintaining much of the existing file structure and causing as little disruption as possible to those reading or writing flow cytometry data files. Retaining backwards compatibility with previous versions was a priority when format changes were contemplated. Backwards compatibility here is defined as the ability of previous releases of analysis software to correctly read and interpret FCS 3.0 data files.

FILE STRUCTURE

FCS 1.0 introduced the basic data file structure which has been maintained in all subsequent versions including the current proposal (FCS 3.0). All FCS compliant data files are divided into four segments: HEADER, TEXT, DATA, and ANALYSIS (optional).

HEADER Segment

The HEADER segment must always be the first segment of the file. The first 10 bytes of the HEADER (and therefore of the file) give the FCS version in ASCII text, e.g., FCS 3.0, followed by four ASCII space characters (ASCII32). The remainder of the HEADER segment provides byte offsets to the other file segments. Each offset value occupies exactly 8 bytes. See Table 1 for a complete list of HEADER byte assignments). The 8-byte fields of the byte offset values create a file size limit of 99,999,999 bytes (8 characters).

The proposed FCS 3.0 standard avoids file-size limits through the following mechanism: when any segment of a data-set falls outside the 99,999,999 byte limit, 0s are placed in the HEADER for both the begin and end byte offset for that segment. The byte offset values for that segment are then found in keyword-value pairs in the data-set TEXT segment. The new byte offset keywords are $BEGINDATA and $ENDDATA, describing the beginning and end of the DATA segment, $BEGINANALYSIS and $ENDANALYSIS, describing the beginning and end of the ANALYSIS segment, and $BEGINSTEXT and $ENDSTEXT, describing the beginning and end of the supplemental

Table 1
HEADER Segment Byte-Positions

Contents	Bytes
FCS3.0	00–05
ASCII(32)-space characters	06–09
ASCII-encoded offset to first byte of TEXT segment	10–17
ASCII-encoded offset to last byte of TEXT segment	18–25
ASCII-encoded offset to first byte of DATA segment	26–33
ASCII-encoded offset to last byte of DATA segment	34–41
ASCII-encoded offset to first byte of ANALYSIS segment	42–49
ASCII-encoded offset to last byte of ANALYSIS segment	50–57
ASCII-encoded offset to user defined segments	58–beginning of next segment

TEXT segment (see TEXT segment below for a definition of the primary and secondary TEXT segments). There are no byte offset keywords for the beginning and ending of the primary TEXT segment because that segment must be placed entirely within the first 99,999,999 bytes of a data-set. An example of the new byte offset structure can be seen in the instance where a DATA segment starts at byte 257 and ends at byte 100,345,679. In such case, the HEADER fields 26–33 (begin DATA) and 34–41 (end DATA) would contain the value 0'. The actual byte offsets would be found in the $BEGINDATA and $ENDDATA keyword values in the TEXT segment.

This keyword-value system was selected over alternatives such as a free-form HEADER structure, where the byte offsets in the HEADER are allowed to occupy as many bytes as necessary, to allow previous versions of analysis software to read most FCS 3.0 data files. In the rare event that a data-set is 100 megabytes or longer, file reading programs designed for previous FCS versions will read the 0 byte offset and fail to read the data, avoiding partial or erroneous data reads. However, for the vast majority of data-sets that are smaller than 100 megabytes, byte offset structure will not prevent file reading by software designed to read older FCS files.

TEXT Segment

New to FCS 3.0 is the allowance for primary and supplemental TEXT segments. The primary TEXT segment is required and must be located completely within the first 99,999,999 bytes of a data-set. The primary TEXT segment must contain all required keyword-value pairs, as well as any number of optional keyword-value pairs. The byte offsets to the TEXT segment in the HEADER point to the primary TEXT segment.

The supplemental TEXT segment is optional and may be located anywhere in the data-set after the HEADER segment. The supplemental TEXT segment can contain only optional keyword-value pairs. The supplemental TEXT segment was added to allow data acquisition software to stream file acquisition to a previously created primary

TEXT segment, adding additional descriptive text after the primary TEXT has been created.

As before, the first character of the primary TEXT segment contains the delimiter character. The delimiter separates keywords from the keyword values. The remainder of the TEXT segments contain a series of keyword-value pairs that describe various aspects of the data-set. For example, $TOT/5000/ is a keyword-value pair, indicating that the total number of events in the file is 5000. $TOT is the keyword, 5000 is the value and ''/'' is the delimiter character. The ''$'' character flags this keyword as a standard FCS keyword. There was some discussion about creating new keyword flag characters, such as a ''#'' sign to indicate a local-user or manufacturer defined keyword. However, the committee concluded that this would add little utility since the ability to create such designations is currently available under existing FCS standards without the standard strictly defining them.

The FCS 3.0 TEXT segment allows the inclusion of international characters in keyword-values using the UNICODE characters set (5). The value for the keyword ''$UNICODE'' is a coma-delimited list of other keyword-values that are represented in UNICODE characters. UNICODE is a two-byte character set in which the first byte indicates the UNICODE ''page'' where the character is found and the second byte represents the specific character on that page. This allows for 256 pages of 256 characters per page, or, a total of 65536 characters in the UNICODE set. For example, American ASCII is UNICODE page 0.

Several keyword-value pairs have been modified to provide more detailed information regarding data acquisition. The keyword $PnE, introduced in FCS 2.0 to describe the method of log amplification for any given parameter, n, is a required keyword in the proposed standard. With the widespread use of third party analysis software, it is necessary to record the specific details of log conversion used in data acquisition. Information provided in this keyword value will allow users to calculate more accurately relative fluorescence intensities necessary to make quantitative fluorescence measurements. Specifically, the $PnE keyword value provides analysis software with the number of log decades and the linear value that would have been obtained for a signal with a log value of channel 0.

The set of $DFCiTOj keywords that described the amount of fluorescence compensation employed during data collection has been replaced by the single new keyword $COMP. The value of $COMP is a set of comma-delimited numbers representing an ''n'' dimensional square matrix, where ''n'' is the number of acquisition parameters (also found in the value of the $PAR keyword). The elements are stored in row-major order, i.e., the elements in the first row appear first. For example, in a three parameter data set, the fourth matrix element in the $COMP value represents the first element of the second row and indicates the percentage of parameter 1 that has been subtracted (or added) electronically from parameter 2. Both positive and negative matrix values are allowed. A positive (or unsigned) value indicates that compensation has been additive while a negative value indicates the more common case of subtractive compensation. Values must be present for all matrix elements (e.g., including light scattering parameters) even if no compensation has taken place. Note that only those parameters that are actually collected in the file can be included in the matrix. For example, there would be only four elements in the matrix for a two parameter correlated histogram even if other (uncollected) parameters were being compensated. Note also that the percentages refer to linear parameter values prior to arithmetic conversion (e.g., log conversion).

A new keyword, $TIMESTEP, has been added to more specifically detail the absolute measurement of time used in kinetic analyses. When time is collected, the keyword value of the time-parameter name ($PnN) must be the string ''TIME.'' The keyword value $TIMESTEP then gives the channel resolution for that parameter in seconds or fractions of a second. For example, if the time channel increments every 1/30 of a second, the keyword value for $TIMESTEP would be 0.0333.

DATA Segment

As in previous versions, the DATA segment contains the raw data in one of three modes (list, correlated, or uncorrelated) described in the TEXT segment by the $MODE keyword value. Data are written to the DATA segment in one of four allowed formats (binary, floating point, double precision floating point, or ASCII) described by the $DATATYPE keyword value. The most common form of data storage has historically been list mode in the form of binary integers ($DATATYPE/I/ $MODE/L/). The $PnB set of required keywords specify the bit width for the storage of each parameter. The $PnR set of keywords specify the channel number range for each parameter. For example, $P1B/16/ $P1R/1024/ specifies a 16-bit field for parameter 1 and a range for the values of parameter 1 from 0 to 1023. This implies that 10 bits of the 16 bit field are used to store the data. The remaining bits are usually unused and set to ''0''; however, some file writers store non-data information in that bit-space. Implementers must use a bit mask when reading these list mode parameter values to insure that erroneous values are not read from the unused bits.

ANALYSIS

ANALYSIS is an optional segment that, when present, contains the results of data processing. The ANALYSIS segment has the same structure as the TEXT segment; i.e., it consists of a series of keyword-value pairs. There are no required keywords for the ANALYSIS segment. It is often the case that analysis is performed off-line, after the data has been collected and stored in a data-set. Therefore, the ANALYSIS segment typically contains information added to a copy of the original file. For examples, the results of cell cycle analysis or immunophenotype determinations often involve more complex analyses than can be performed in ''real time'' as the data is collected and stored.

Therefore, these results can later be stored in the ANALYSIS segment.

Data-Set Verification

New to FCS 3.0 is the ability to verify data-set integrity. With the growth of local and wide area networks and their routine use to move flow cytometry data files, it has become desirable to detect errors introduced in data files. Therefore, in FCS 3.0, the last two-bytes of a data-set are set aside for the storage of a 16-bit cyclic redundancy check (CRC) value (3). Using a 16-bit CRC, the chances of a random error going undetected is 1 in 2^{16} or 1 in 65,536. The CRC value will be calculated and added to the file at the time of file acquisition. The CRC value can then be checked by analysis software at any later time. This is an optional feature; however, if an implementer chooses not to include a CRC value, ''0''s must be placed in these two bytes.

Summary of the Major Changes Found in FCS 3.0

FCS 3.0 is different from FCS 2.0 in the following ways:

1) In FCS 3.0, the HEADER has been modified to accommodate data-sets longer than 99,999,999 bytes. A data offset value that requires more than 8 bytes is now represented by placing a ''0'' in the HEADER for both the start and stop values of that segment. The actual byte offset value is found in the primary TEXT segment of the data-set.

2) The TEXT segment may now be split into primary and supplemental TEXT segments. The primary TEXT segment must contain all required keyword-value pairs and be located entirely within the first 99,999,999 bytes of a data-set. The supplemental TEXT segment may contain only optional keyword-value pairs and may be located anywhere in a data-set after the HEADER segment. The byte offset to the primary TEXT segment is found in the HEADER segment. The byte offsets to the supplemental TEXT segment is found in keyword-value pairs in the primary TEXT segment.

3) An optional 16-bit Cyclic Redundancy Check (CRC) has been added to the end of each data-set. This internal check-word allows for data-set integrity checks.

4) To enable third party or off-line analysis software to correctly read and interpret data, the keyword $PnE is now required for each parameter. The $PnE keyword describes the method of amplification used for a given parameter using two floating point values.

5) The new keyword $COMP has replaced $DFCiTOj to describe the amount of fluorescence compensation employed in the collection of the data.

6) $TIMESTEP has been added to more accurately define how a TIME parameter is measured.

7) The $UNICODE has been added to enable the specification of certain keywords in languages not representable in ASCII. UNICODE is an international standard that enables computer representation of most of the world's languages.

DISCUSSION

The changes incorporated in FCS 3.0 were necessitated by the rapid evolution in computer technology, computer communications, instrument design, and experimental complexity. In 1984 few envisioned the cost of data storage would dip below 3 cents per megabyte and flow cytometry data file size would be approaching 100 megabytes or that data would routinely be transmitted via Ethernet world-wide. FCS 3.0 reflects the current state of the art of flow cytometry while retaining a high level of backwards compatibility with previous versions of the standard.

The FCS data file standard must evolve to keep pace with technology and current practices in flow cytometry. Continued development in analytical cytometry will continue to require changes in data file format. For example, new digital data acquisition hardware, such as the DiDac system in use at National Flow Cytometry Resource at Los Alamos, will have the ability to store individual pulse shapes as a parameter (personal communication, Habbersett R). Such a parameter will include many individual integers describing a pulse. In other words, one of the measured parameters for each event in a list-mode file may itself be a single parameter histogram. Also, pressures are mounting to integrate flow cytometry data into larger files of related information. For example, a flow cytometry list-mode or histogram file could be one element in a pathology report file that may also include MRI or X-ray images, clinical laboratory data, etc. Because the FCS standard is both simple and flexible it has gained wide acceptance and is now implemented in some form by most instrument manufacturers. With carefully measured evolution the FCS file format should continue to find wide use for some time to come.

This report summarizes the work of the ISAC Data File Standards Committee. The complete and detailed FCS 3.0 standard is available through the ISAC office [Sherwood Group, 60 Revere Drive, Ste 500, Northbrook, IL 60062, phone: (847) 480-9080 ext 231, fax: (847) 480-9282, email: isac@sherwood-group.com] or through the internet at the ISAC WWW site, http://nucleus.immunol.washington.edu/ISAC.html

LITERATURE CITED

1. Dean PN, Bagwell CB, Lindmo T, Murphy RF, Salzman GC: Data file standard for flow cytometry. Cytometry 11:323–332, 1990.
2. Murphy RF, Chused TM: A proposal for a flow cytometric data file standard. Cytometry 5:553–555, 1984.
3. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: Numerical Recipes in C. 2nd ED. Cambridge University Press, Cambridge, UK, 1992.
4. Redelman D, Coder DM: Cell subset (CS) parameter to record the identities of individual cells in flow cytometric data. Cytometry 18:95–102, 1994.
5. The UNICODE Consortium: The UNICODE Standard, Version 1.0, vol. 1. Addison- Wesley Publishing Co. Inc., Reading, MA, 1991.

Appendix A
Data File Standards Committee of the International Society for Analytical Cytology

Larry Seamer, Chair
Director, Flow Cytometry Facility
University of New Mexico
Cancer Center, Cytometry
900 Camino de Salud NE
Albuquerque, NM 87131
(505) 272-6206
lseamer@cobra.unm.edu

C. Bruce Bagwell
Verity Software House
PO Box 247
Topsham, ME 04086
(207) 729-6767 x102
cbb@vsh.com

Luther Barden
Div. of Computer Research and Technology,
Building 12A Room 2015
National Institutes of Health
9000 Rockville Pike
Bethesda, MD 20892
luther_barden@nih.gov

Marc Christofferson
Becton Dickinson Immunocytometry Systems
2350 Qume Drive
San Jose California 95131-1807
(408) 954-2058
m_chr@BDIS.com.

Louise E. Magruder
Division of Clinical Laboratory Devices
FDA/CDRH/ODE
72 Gaither Road
Rockville, MD 20850
lem@fdadr.cdrh.fda.gov

George Malachowski
Cytomation, Inc.
400 E. Horsetooth Rd.
Ft. Collins, CO
(303) 226-2200

Robert F. Murphy
Associate Professor
Department of Biological Sciences and Center for Light Microscope Imaging and Biotechnology
Carnegie Mellon University
4400 Fifth Avenue, Box 52
Pittsburgh, Pennsylvania 15213
(412) 268-3480
murphy@cmu.edu

Doug Redelman
Sierra Cytometry
3150 Susileen Dr.
Reno, NV 89509

Gary C. Salzman
Life Sciences Division
Los Alamos National Laboratory
Mail Stop M888
Los Alamos, NM 87545
(505) 667-5503
salzman@lanl.gov

James C.S. Wood
Coulter Corporation
Mail Code 52-A01
11800 S.W. 147th Avenue
Miami, FL 33196-2500
(305) 380-2449 or (954) 344-1290 (voice)
(954) 344-5240 (FAX)
woodjcs@gate.net