Astronomical Data Analysis Software and Systems XXVI
ASP Conference Series, Vol. 521
Marco Molinaro, Keith Shortridge, and Fabio Pasian, eds.

FITS and Data Representations WG BoF Session

Lucio Chiappetti,[1] Jessica Mink,[2] Adam Dobrzycki,[3], and Mark B. Taylor[4]

[1]*INAF IASF Milano, via Bassini 15, I-20133 Milano, Italy;*
`lucio@lambrate.inaf.it`

[2]*Smithsonian Astrophysical Observatory, Cambridge, MA 02138, USA*

[3]*European Southern Observatory, D-85748 Garching bei München, Germany*

[4]*H. H. Wills Physics Laboratory, University of Bristol, Bristol BS8 1TL, U.K.*

Abstract. A Birds-of-a-Feather (BoF) session was held about the FITS data format and IAU Data Representation Working Group which should care about the maintenance of FITS, potentially of other astronomical data formats, and think about a possible successor. A summary of the presentations and discussion held during the BoF, and afterwards during the conference, is given.

1. Introduction (L.Chiappetti)

The introductory talk was split in two parts: one about recent activities about FITS (Flexible Image Transport System), and one about the IAU Data Representation Working Group (DRWG) being established.

1.1. FITS

The IAU FITS WG has recently released a draft of version 4.0 of the FITS standard[1], which is undergoing language editing and will be officially issued soon. All additions to the standard have already been voted by the FITS WG and therefore are *already in effect* and include (see draft for more details):

- A chapter on time WCS (Rots et al. 2015);
- A chapter on tiled image and table compression (incorporated convention);
- Incorporation of the `CONTINUE` convention for long-string keywords;
- Data integrity (checksum) keywords;
- Column limit keywords;
- Usage convention to reserve blank space at the end of a FITS header;
- Mention of the Green Bank convention;
- Reserving the `INHERIT` keyword.

[1] the draft can be retrieved from page `http://fits.gsfc.nasa.gov/fits_standard.html`

The FITS WG is going to reorganize its membership (and possibly its voting rules) in the framework of the DRWG and continue the maintenance of the FITS standard building upon discussions already initiated.

The possible change to the voting rules (proposed by P. Grosbol and A. Dobrzycki, ESO) involves a weighted voting systems where the votes of *institutional members* weigh more than those of individual members. Institutional members should represent major data centres, producing and/or archiving a large quantity of FITS files.

Concerning possible updates to the FITS standard, the FITS WG started already to consider the possibility of long keyword names (e.g. from 55 to 74 characters) with an extended character set (`a-z _-$.:` and perhaps blank space) and to evaluate the possible incorporation of more registered conventions (the project-specific ones will remain just conventions), e.g. spatial region and WCS conventions, ESO HIERARCH, hierarchical grouping, FOREIGN extension.

Other items mentioned are: dedicated metadata or MIME extensions, Unicode strings, an "index HDU" for multi-extension files, etc.

1.2. DRWG

A broader Data Representation Working Group (DRWG) has been established under IAU Commission B2 "Data and Documentation" during the recent IAU commission reform. Among its terms of reference, alongside with the continued support to FITS, there is the capability *to manage a careful and minimally disruptive transition from FITS to more modern and capable data representations.* This BoF session is an occasion to invite volunteers as members of the DRWG.

A possible scheme with an outer plenary tier, and an inner tier with thematic "Special Expert Groups" (SEGs) was presented. This is mimicked on the relationship between the FITSBITS mailing list (open to the entire community), and the FITS WG (with an authoritative role to decide about the standard), except that the outer tier would be made mainly by IAU members with a general interest in data formats. The SEGs members would be IAU members and technical associates with specific skills and competences about their particular topic.

According to a proposal by R. Shaw [NOAO], there could be *curation SEGs* and *topical SEGs*. A curation SEG will take care of maintaining an existing standard and therefore be long-term. Could have about two dozen members, possibly renewed by rotation, with a limited time commitment (except for occasional task forces). A topical SEG with about a dozen members will have a limited duration and a specific charter. It requires a significant time effort and competence from its members.

The former FITS WG will become the first SEG, and of course be a curation SEG. Other curation SEGs could look after astronomical usage of other data formats (e.g. HDF5, NDF, ASDF).

Possibilities for other SEGs are e.g. a Next Generation data format SEG (in charge to design, prototype, and formulate a proposal); a Data Provenance SEG (for the concept see e.g. Riebe et al. 2019); and an Event Stream SEG (in both cases to define IAU standards in cooperation with VO).

A meeting on one day following to the BOF session (see section 3.1) lead to the proposal of a Structured Data SEG, which should *define a grammar and semantics for structured data file formats, and provide a common framework for a possible multiplicity of different serializations.*

2. A possible solution for metadata (M.Taylor)

Based on the general discussion (see section 3) about the need of better ways to store metadata, the solution used in TOPCAT/STIL (Taylor 2005) to serialize VOTables in FITS was presented. During the discussion it was recommended to register it as a FITS convention.

VOTable has richer metadata than FITS, and is streamable. It keeps metadata and data separated and has two variants. In one tabular data is stored internally as HTML-like data, in the other it refers to an external FITS file. This solution requiring two files (although not unknown, see e.g. the original IRAF imh+pix format) is nowadays rarely used.

The so called *FITS-plus* solution, using a single (fully legal) FITS file, allows aware clients to get both FITS efficient data storage and VOTable rich metadata, while plain FITS clients will be able to use just the standard `BINTABLE`. The *FITS-plus* "trick" consists in storing the VOTable XML metadata in the FITS primary HDU serialized as a `BITPIX=8` 1-dimensional array, while the table data are present in normal `BINTABLE` extensions (the <`DATA`> tag in the VOTable are empty).

This approach allows use of existing code without the need to develop a new (FITS) standard for "extended metadata". The *FITS-plus* format as it stands is specific to VOTable metadata, but the approach could in principle be extended to other metadata models.

3. Discussion summary

The main points arisen during the discussion are summarized here.

Landry [IRSA] mentioned the problem of multiple simultaneous access for big data, Taylor (et al.) the difficulty to stream FITS when the number of rows in the data is not known *a priori*. Akhlaghi [Lyon] noted that speed, specially for multiple reads and writes, is an OS tuning issue, not a FITS one.

HDF5 (and the possibility of an HDF5 SEG) was mentioned by Glendenning [NRAO], Mink, et al. as interesting for what concerns rich metadata and hierarchical arrangement, but requiring a degree of standardization alike to FITS (about which Chiappetti reminds the "interoperability demonstration" required in the FITS WG procedure to accept a new change). Shortridge [AAO] noted that a hierarchical data format can give too much rope to hang oneself, one needs to set the Data Model right!

Fitzpatrick [NOAO] reminded that FITS is born as an archival and transport format. Working formats, particularly for existing pipelines, can be different. Legacy applications will never be updated. Chiappetti agreed that legacy should preserve the data, because the applications (and their know-how) can't be there forever. Fitzpatrick added that it is unlikely to rewrite archival data soon because of lacking resources. Raugh [Maryland] noted that an archive horizon is about 50 years, only ASCII text so far lasted that long. A similar statement was made during the O3.2 talk (Fernique et al. 2019).

Raugh and Marmo [Orsay] mentioned the Planetary Data System (see Raugh & Hughes 2019, and link therein), allowing e.g. to use XML for rich metadata, and various working formats (FITS included) according to needs. The point is information transfer, not data transfer. Shortridge and Glendenning noted that software has internal data models too, one needs an intermediate layer translating internal and external data models (and there are working examples). Data should be delivered to users in the format they want.

3.1. The "upstairs meeting"

A meeting one day after the BOF session was held in the "upstairs" meeting room provided by the conference centre (attendance: Chiappetti, Mink, Dobrzycki, Taylor, Rots [SAO/CXC], Wicenec [ICRAR], Jenness [LSST], Gabriel [ESA], Pineau [CDS], Zhelenkova [SAO RAS]), during which it was agreed to go on with SEGs, and in particular with the proposal of a Structured Data SEG, which should define a grammar and semantics for structured data file formats, using the VO-DML as starting point (Louys 2019) and provide a common framework for a possible multiplicity of different serializations (e.g. HDF and ASDF). The SEG could *only later* split in data format and data model specific groups.

About the *FITS-plus* metadata arrangement (section 2) Rots suggested using a separate extension instead of the primary HDU. About streaming FITS files, Chiappetti, citing Don Wells, commented that streaming could be achieved defining a protocol which stores header and data separately, or leaves `NAXISn` undefined, and defines a data terminator like e.g. MIME multipart files, updating the file when its final size is known.

Wicenec noted that it makes no sense to overcomplicate FITS for what some data centres are already doing efficiently. One could separate working format from transport format. And one should keep separate the data model from the format (and have a data model for keywords too! For these Chiappetti commented that the confusion originated around 1992 when defining `BINTABLE`). On the same topic Rots notes that unification of the entire community behind a single data model could be not possible nor desirable. FITS provides a grammar, not a semantics. Jenness, about the coexistence of different formats, reminded the case of NDF where different formats were converted to and from such format (Jenness et al. 2015). Virtually everybody agreed that the data model issues should be dealt with in close coordination with the VO.

References

Fernique, P., et al. 2019, in ADASS XXVI, edited by M. Molinaro, K. Shortridge, & F. Pasian (San Francisco: ASP), vol. 521 of ASP Conf. Ser., 46

Jenness, T., et al. 2015, Astronomy and Computing, 12, 146. 1410.7513

Louys, M. 2019, in ADASS XXVI, edited by Molinaro, M. and Shortridge, K. and Pasian, F. (San Francisco: ASP), vol. 521 of ASP Conf. Ser., 437

Raugh, A., & Hughes, J. 2019, in ADASS XXVI, edited by M. Molinaro, K. Shortridge, & F. Pasian (San Francisco: ASP), vol. 521 of ASP Conf. Ser., 100

Riebe, K., et al. 2019, in ADASS XXVI, edited by Molinaro, M. and Shortridge, K. and Pasian, F. (San Francisco: ASP), vol. 521 of ASP Conf. Ser., 450

Rots, A. H., et al. 2015, A&A, 574, A36

Taylor, M. B. 2005, in ADASS XIV, edited by P. Shopbell, M. Britton, & R. Ebert, vol. 347 of ASP Conf. Ser., 29