The FITS file format for the long-term preservation of digital objects in the Environment and Earth Observation domain

Stefano Allegrezza
University of Bologna, Italy

Abstract

One of the most critical problems in the long-term preservation of digital objects in any domain is the rapid obsolescence of file formats that become outdated in a short time and therefore no longer readable. This problem also affects the field of environmental and earth observation, where it is important that the data collected remain accessible for many years to come. For this purpose, in recent years the attention has been focused on an extremely versatile file format created for the astronomy and astrophysics domain but then spread to other sectors as well: the Flexible Image Transport System (FITS). What exactly is it? What are the characteristics of the FITS format that made it so interesting to be a candidate for adoption in the Environment and Earth Observation (EO) field? Is a format that was created over forty years ago still appropriate for long-term preservation? This paper aims at answering these questions by starting from an analysis of the FITS file format and highlighting its features in order to understand if it is appropriate for the archiving and preservation of data collected by many scientific projects in the EO domain.

1.0 Introduction

One of the most worrying problems in the long-term preservation of digital objects in any domain is the rapid obsolescence of file formats that become outdated in a short time and therefore no longer readable. This problem also affects the Environment and Earth Observation (EO) domain, where it is important that the data collected remain accessible for many years to come.

> "Earth Observation data are unique by nature and are fundamental for the monitoring of our environment and planet and of its changes. They are considered as humankind assets and as such need to be preserved without time constrains and kept accessible together with all the information and knowledge needed to understand and use them in future" (Albani 2012).

There are strong motivations for preserving Environment and Earth Observation data (National Research Council 1995). For instance, many observations about the natural world are records of events that will never be repeated exactly. Examples include observations of an atmospheric storm, a deep ocean current, a volcanic eruption, and the energy emitted by a supernova. Once lost, such records can never be replaced. Observed data provide a baseline for determining rates of change and for computing the frequency of occurrence of unusual events. They specify the observed envelope of variability. The longer the record, the greater our confidence in the conclusions we can draw from it. Data records may have more than one life and can be reused. As scientific ideas advance, new concepts may emerge – in the same or entirely different disciplines – from the study of observations that led earlier to different kinds of insights. New computing technologies for storing and analyzing data enhance the possibilities for finding or verifying new perspectives through reanalysis of existing data records. Thus, the relative importance of data, both current and historical, can change dramatically, often in entirely unanticipated directions. Finally, the substantial investments made to acquire data records justify their preservation: because we cannot predict which data will yield the most scientific benefit in years ahead, the data we discard today may be the data that would have been extremely valuable tomorrow.

Long-term preservation of EO data and of the ability to discover, access and process them is a fundamental issue and a major challenge at programmatic, technological and operational levels (Albani and Giaretta 2009). In order to create and administer archives that serve current and future generations of users adequately, archive owners require explicit and precise requirements from data users (Molch et al. 2012). Furthermore, curating and preserving EO data is important to justify funding for long-term preservation in today's challenging economic climate, archives must be able to present a convincing case for the value of the data in the long term (Conway et al. 2014).

Because of the rapid technological change, the long-term storage of EO data comes with its own set of risks and difficulties. The corruption of the bitstream, as well as existing hardware and operating environments, are all dangers that render data inaccessible on a physical and logical level. However, insufficient data description, inability to discover data, and service compatibility might all limit re-use (Albani 2012). There is another major difficulty in the EO area that makes long-term preservation of obtained data particularly difficult and contributes to the high operational and maintenance expenses of long-term archiving: the excessive proliferation of diverse and heterogeneous data formats. This is caused by mainly three reasons:

> "the lack of an agreed standard in the EO community, reason for which the formats tended to be specific for the sensors each mission carried on

> board; legacies from old ground segments architectures, which tended not to reuse elements previously developed; the non-mature status, until recently, of the information technologies and standards used to describe and package the data, preventing the creation of a unique format able to satisfy at the same time the requirements for the long-term preservation of the data and their handling in the processing centres" (Pinna and Mbaye 2014).

To overcome this problem digital preservation specialists, knowing that digital preservation requires their continuous migration to other formats over time, suggest reducing the number of formats to manage; and, better, using file formats that give the broadest guarantees of long-term sustainability.

For this purpose, in recent years the attention has been focused on an extremely versatile file format that was created for the astronomy and astrophysics domain but that is now spreading to other fields: the Flexible Image Transport System (FITS). This paper aims at evaluating the FITS file format in order to understand if it is suitable for archiving and preserving data collected by many scientific projects in the EO domain – such as the ERA-PLANET project that has generated a substantial amount of data and knowledge on different aspects related to the environmental quality and sustainability – and to make it easier to share them with stakeholders and policy makers and to support decision making.

2.0 State of the art

Over the past fifteen years, leading EO organizations have launched projects to define the content to be preserved over time (Ramapriyan et al. 2017). For example, the European Space Agency (ESA) formed a Long Term Data Preservation (LTDP) Working Group in 2007 with the aim of defining and promoting a coordinated approach to the preservation of the European EO data assets. One of the outputs of this working group was the "Earth Observation Preserved Data Set Content" (EO PDSC), a document providing guidance to data holders on preservation. There have been several versions of this document, the latest having been published in 2012 (ESA and CEOS/WGISS 2015; ESA 2105). In late 2011 NASA developed its Earth Science Preservation Content Specification (PCS) (NASA 2011) and has been using it as a requirement for its new missions. For missions that had been in progress or completed before the PCS was developed, it is used as a checklist to capture and preserve as many relevant content elements as possible based on best efforts. Other projects have also addressed these issues but the issue of electronic formats has only been partially addressed.

A very large number of file formats and solutions have been proposed in the EO domain. The major players and many research groups have proposed their own file format for data archiving, sometimes without any care for long-term preservation, although in the last 15 years several projects were funded to address the long-term preservation issue of scientific data in general and a dedicated ISO standard has been developed, ISO 19165-2:2020, which aims to provide details about content describing the provenance and context specific to data from missions that observe the Earth using space-borne, airborne or in situ instruments (ISO 2020). Some of the most adopted file formats include:

- HDF (Hierarchical Data Format), a set of file formats (HDF4, HDF5, HDF-EOS 5) designed to store and organize large amounts of data. Originally developed at the National Center for Supercomputing Applications (NCSA), HDF is currently supported by the HDF Group, a non-profit corporation whose mission is to ensure continued development of HDF5 technologies and the continued accessibility of data stored in HDF (https://www.hdfgroup.org);
- CDF (Common Data Format), a file format (together with a library, and a toolkit) that was developed by the National Space Science Data Center (NSSDC) at NASA starting in 1985;[1]
- NetCDF (Network Common Data Form), a self-describing, machine-independent file format and a set of software libraries that support the creation, access, and sharing of array-oriented scientific data. It is commonly used in climatology, meteorology and oceanography applications (e.g., weather forecasting, climate change) and GIS applications. The project is hosted by the Unidata program at the University Corporation for Atmospheric Research (UCAR);[2]
- GeoTIFF (Tagged Information File Format), an extension of TIFF that includes georeferencing or geocoding information embedded within a TIFF file (such as latitude, longitude, map projection, coordinate systems, ellipsoids, and datums) so an image can be positioned correctly on maps of the Earth;
- SAFE (Standard Archive Format for Europe), a file format designed to act as a common format for archiving and conveying data within ESA Earth Observation archiving facilities. SAFE has been designed to be used in an archive system compliant with the Open Archival Information System

1 NASA, "Space Physics Data Facility," last accessed December 13, 2021, https://cdf.gsfc.nasa.gov/html/FAQ.html.

2 Unidata, "Network Common data Form (NetCDF)," last accessed December 13, 2021, https://www.unidata.ucar.edu/software/netcdf.

(OAIS) standard. SENTINEL-SAFE format wraps a folder containing image data in a binary data format and product metadata in XML;[3]
- ASDF (Advanced Scientific Data Format) is a proposed replacement to the FITS standard for astronomical images and other astronomical data. The metadata is contained in a YAML (Human-readable data serialization format) header followed by binary or ASCII data;[4]
- GRIB (GRIdded Binary or General Regularly-distributed Information in Binary form), a concise data format commonly used in meteorology to store historical and forecast weather data. It is standardized by the World Meteorological Organization's Commission;[5]
- PP-format (Post Processing Format), a proprietary file format for meteorological data developed by the Met Office, the United Kingdom's national weather service;
- GFS (Global Forecast System), the format used by a global numerical weather prediction system containing a global computer model and variational analysis run by the United States' National Weather Service (NWS);[6]
- CEOS (Committee on Earth Observation Satellites), a standard format published in 1988, used for radar data and originally expected to be used with tape media. The format does not specify a naming convention (https://ceos.org).

There are also many other file formats, such as CCRS, EOSAT, AVSAR, AIRSAR, DORADE, Cf Radial, LGSOWG, UNIVERSAL Format (UF), FORAY NetCDF Format, OGC KML, Fast-L7A and the IWRF time series format.

From the point of view of long-term preservation, the fact that in the field of EO there are so many and diverse file formats is a serious problem. In fact, you have to consider that you need to manage all these formats (e.g., you need to migrate them over time), and managing a few formats – or better just one – is definitely easier than managing several dozens of file formats.

To overcome this problem digital preservation specialists suggest to reduce the number of formats to manage and to prefer formats that give the broadest guarantees of long-term preservation. The perfect solution would be a versatile file format (e.g. we can use it in different application areas), and

3 ESA, "SAFE," last accessed December 13, 2021, https://earth.esa.int/SAFE/index.html.
4 STSI, "ASDF Standard," last accessed December 13, 2021, https://asdf-standard.readthedocs.io/en/1.6.0.
5 COSMO, "GRIB Data Format used for tge COSMO-Model System Features of GRIB," last accessed December 13, 2021, http://www.cosmo-model.org/content/model/documentation/grib/grib_ features.htm.
6 NCEI, "Global Forecast System," last accessed December 13, 2021, https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast.

a file format that do not need to be migrated for at least a few decades. In this regard, in recent years the attention has been focused on the Flexible Image Transport System (FITS).

3.0 The Flexible Image Transport System

FITS is a widely used file format for storing images and data in astronomy and astrophysics (Wells and Greisen 1979; Greisen et al. 1980; Wells et al. 1981; Greisen and Harten 1981; Hanisch et al. 2001). It was created in the early 70s to solve the problem of the great variety of electronic formats existing in that period. Since every institution in the late Seventies had its own way to keep the collected data, it became necessary to establish a standard in order to avoid the waste of time and resources usually spent each time to develop a customized software to convert the received data to the format used at the home institution and so forth.[7]

The year of birth of the format is dated 1979, but the official year of birth of the format is considered 1981, when the specifications of the file format where published in the "Astronomy and Astrophysics Supplement Series" journal thanks to an article written by Donald C. Wells, Eric W. Greisen, and Ronald H. Harten (Wells et al. 1981). With the help of another author, Preben Grosbol (Grosbol 1988), during a seven year period of time (1981-1988) they published three more papers in the same journal to better describe the specifications of the format: all together they are known as the 'Four FITS papers' (NASA 2014).

In 1982 the International Astronomical Union (IAU), a non-profit international association of scientists of the space and astronomical sector, formally adopted it and in a short time the FITS file format became a de facto standard for the interchange of data. However, the first official version of the FITS Standard was drafted by the NASA Office of Standards and Technology (NOST) at the Goddard Space Flight Center, and formally approved by the IAU FITS Working Group (IAU-FWG)[8] only in June 1993, under the name NOST 100-1.0. The second official version, called NOST 100-2.0, dates back

7 The main motivation for the introduction of the format at the time stemmed from the fact that the majority of scientific organizations operating in the field of astronomy used for storing scientific data, proprietary file formats and not supported by the software in use by the other organizations. It was felt, therefore, the need to overcome the situation of insufficient interoperability then widespread thanks to a common format that would enable the interchange of data.

8 The IAU FITS Working Group consists of the following regional committees: the European FITS Committee; the Japanese FITS Committee; the American Astronomical Society FITS Committee; the Australian/New Zealand/Pacific FITS Committee.

to March of 1999. This version was published in the "Astronomy and Astrophysics" journal in September of 2001. Seven years later a new version, numbered 3.0, was officially approved by IAU-FWG in July 2008 and remained valid until 2016, when the fourth version of the FITS Standard was made public and is currently in force.

Table 1 lists the various versions of the format that have occurred since its birth. It is worth noticing that the file format has been very stable throughout the years, since it has gone through only four major versions in about forty years.

Version	Name	Date	Notes
	-	1979	FITS birth
	-	1981	Official FITS birth
	-	1982	FITS adopted by IAU
0.1	NOST 100-0.1	December 1990	1st draft
0.2	NOST 100-0.2	June 1991	2st draft
0.3	NOST 100-0.3	December 1991	3rd draft
1.0	**NOST 100-1.0**	**June 1993**	**NOST Standard (1st version)**
1.1	NOST 100-1.1	September 1995	Some amendments
1.2	NOST 100-1.2	April 1998	Some minor amendments
2.0	**NOST 100-2.0**	**March 1999**	**NOST standard (2nd version)**
2.1	IAUFWG 2.1	April 2005	IAUFWG standard
2.1b	IAUFWG 2.1b	December 2005	IAUFWG standard (64 bit integers support added)
3.0	**IAUFWG 3.0**	**July 2008**	**IAU-FWG standard (3rd version)**
4.0	**IAUFWG 4.0**	**July 2016**	**IAU-FWG standard (4th version)**

Table 1. The evolution of FITS

The FITS format specifications are described in a document called "The FITS Standard" maintained by NASA.[9] The latest version, consists of a text of about 75 pages divided into sections which fully covers the whole range of the format's details, such as the file organization, headers, keywords, data

9 NASA, "FITS Standard Document," last accessed December 13, 2021, https://fits.gsfc.nasa.gov/fits_standard.html.

representation, standard extensions, the world coordinate system, etc. Currently the IAU-FWG maintains the format adding updates and improvements.

Although FITS was developed by astronomers as the standard file format for the interchange of images between hardware platforms and software applications that did not share a common file format, currently it has become the standard format for archiving and preservation of astronomical images of many scientific organizations, such as the images captured from radio telescopes and other astronomical equipment or those returned by orbital satellites, from the spacecraft and from the planetary probes in the course of their missions (for example, NASA uses it to store the images of the space missions).

3.1 How FITS organizes data

A FITS file consists of a sequence of one or more blocks, each of which is referred to as HDU ("Header and Data Unit"): the first is referred to as "Primary HDU" and the following are referred to as "Extension HDU" (NASA/GSFC 1997). Each HDU in turn consists of two sections: a header (called "Header Unit" or simply "Header") and an optional section containing data (called "Data Unit" or "Data Array"). In the case of the "Primary HDU" the header takes the name of "Primary Header Unit" and the data section takes the name of "Primary Data Unit" (or also "Primary Data Array"), while, in the case of an "Extension HDU" the two sections are called "Extension Header Unit" and "Extension Data Unit" (or "Extension Date Array"), respectively (see Figure 1). The simplest FITS file contains a header and a single data stream. The simplest file in FITS format consists of a single Primary HDU.

FITS can contain data arrays that range in size from 1 to 999. In the case of size "2" it is called the matrix and the data are typically two-dimensional digital representation of an image in raster format.

The format structure is flexible and allows to add any number of Extension HDU depending on the amount of information you need to store. Currently the format specifications include three standard types of Extension HDU (NASA 2014):

- *Image Extension HDU*, used to store arrays of pixels in size between 0 and 999 (exactly the same type of data stored in the Primary HDU). In this case the Header Unit begins with the string XTension = 'IMAGE').
- *ASCII Table Extension HDU*, used to store information in a table using data encoded in ASCII. ASCII tables are generally less efficient than binary tables, but can be constructed to be easily read by humans, and are

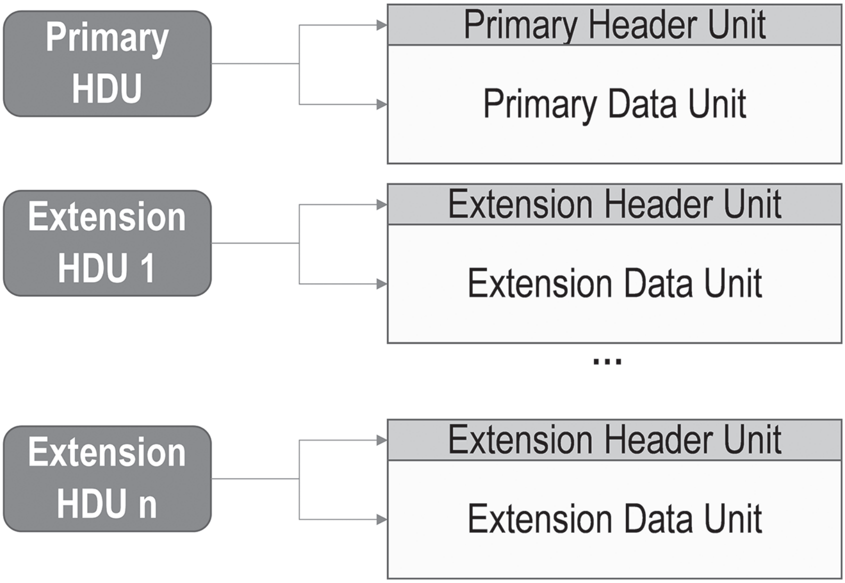

Figure 1. FITS file structure

capable of storing numeric information with arbitrary size and precision. In this case (header begins with XTENSION = 'TABLE').

- *Binary Table Extension HDU*, used to store information in tabular form using a binary representation. In this case the Header Unit begins with the string XTension = 'BINTABLE'. Each cell of the table can be a matrix, but the size of this array must be constant within a column. Thanks to this type of extension it is possible to use the FITS format to store, for example, entire relational databases.[10]

Each Header Unit consists of one or more records, known as keyword records (or even "card image"), each of which has a length of 80 characters and is encoded using the set of characters between the restricted ASCII character "SP" and the ASCII character "~" (126 decimal or hex 7E) (see Figure 2).

For historical reasons, the size of each Header or Data Unit must be an exact multiple of 2,880 bytes (corresponding to 36 alphanumeric strings with a length of 80 bytes each). If necessary, any unused space at the end of a Header or Data Unit can be completed with fill characters up to this dimen-

10 In addition to the structures mentioned above, there is another type of HDU called "Random Group" that is used almost exclusively for applications of radio interferometry, and whose use is not recommended for other types of application outside of it.

```
SIMPLE  =                    T         / file is compliant to FITS standard
BITPIX  =                   32         / Number of bits per pixel
NAXIS   =                    2         / Number of axis
NAXIS1  =                 1000         / lenght of data axis 1
NAXIS2  =                 1500         / lenght of data axis 2
DATE    = '2021/07/14'                 / creation date
ORIGIN  = 'ESA'                        / FITS file originator
AUTHOR  = 'John Brown'                 / creator
OBJECT  = 'Temperature Measurements'   / name of object
HISTORY   'First version'
COMMENT   'Example of Header Unit'
END
?„»?Šó?[1]Ù>Ðó¢?h?qß<?•^?~É?¯U×>†$¾Û<j=š ?  Çô?>z>Ò"Ô>Ù¼ë>*-O?%8¤?)£>šM?+?^?\?7"=?#;Ç?
[1]i>¦?†‹ê?g!ß?j^À?q?Bž~?\[Ž?'´¤?¤½Þ? @[?±I=?o|›?<^?è_4?Ý®à@?PK?<ƒ<?\>x?KDÏ?zØ¡?Ã##?¸
‹?¸3\?°äP@ht@E£@RË^@y>O@§¦,@¡@ÐµÕ@òhmA[1]éq@ÚÉ•@Ú¼sA7õõAzb€AfxFAV9LAD[¿Aó1A,†A2I›A4ò'
A}(A±¬›A$TÚAq³AHjZAMe+A}¬AŠéAž[1]$A¯  óA³/AÀAAÑúÂAë•|B(LÍB1B)ä£BAûTÇA¥,úAKE¦A.0¢Aß,Aö
@â¤Ï@äZ@¨Áš?¢?L™š>_u'?;dâ>®á‡>RT>ãr"?Ké?D"?°¢>ß3?V™8?Ql?€¦†?ë[>mëw?  <ì?¬>X?v?bœÔ>à3?˜
?;‰?is?Ž»?°é³?'+?›m?gt1>ËÉS>‹pw?GXt?€#4? #;?°â7?¿8?1X0?DRç?Œ¸b?ÜkV?ËuÔ?¬(™?Šùþ?œŸ¬?ê
[1]Ò?äy:?Ì~¦?ç¿€?Éì|?¹dÐ?ÀÍƒ?¬|p@Ÿ@5ü?èõsdjklaakkal
[…]
Ë±@Ôà@Q£@'Œ@†vì@šÈÝ@¡Òž@¨ŽŒ@œ§9@±åÝ@è"=@ëÊ@þ´@áq|AÀA'¬œA†$ÀAËFÂA¿†õAš¨AÈA^
ßA²¿oAÍ±Añ:IB $B[1]…B™7B„õB+>iB\^B¦B{Ú3BþBÙ[1]A¯ ,A^dKAd°AÑoAã9@óµ¸A¼Ÿ@Ý;!?ß?Þ°Ž?H
j…?q³Ë>Ð©?šDÃ?Xaá@ô>àü´?¨^Å?Àü>h³?Š^>Û^Y>£¶Ä?Ò>Ð5»?_„?^oÁ?‰‰?Y,?'Å>õ^¨?)Ö?¡j*?«j2?¢
g?ƒàé?¥?ž©Ö?A^-?AÃ?Ü×ò?°£G?©®ü?›V?[1]?›ûÄ@¹€?Ïùê?7Š?Œál?Yjó?³zW>ëL:=ƒK?kJç?£ŽS?Ø ?¢
?\?†°1?"RÍ?»•Á?Ï‰"?µ¾¤?·  4?"ih?¸RD?†Û¨?…v¦?$ëá?ÒÏd?Óqf?Õä$?õòÕ?¿± ?£m?ãR¿½;Ð1?‹¸‹?¯¾p?¸
ƒ•?‡tO@ Ý>Î\¯?{&Î???90Ü?l‹?‰o0? $ô>-}?«6¶?µçÆ?‡"?³×^@1¬m>,#½?,Ì¡>Y¦G?°üä?ÃÒ?IÆt?„
ò?P®€?<²?é#?W?P`{>ÚÒ7>«Þi?,Ýi?&?n+µ>‰c3>‹|¯>èW`>²êw?,³Ô?Y³>xB?[íG?qy>ù|>?H…t>å8|¾Ê÷¨
¿F[>:¿~?•yt?Ëxò?Köž?!O?tâN?8D¼¿Ç½ÿ+8?Œ
```

Figure 2. A simple FITS file

sion: in the case of a Header Unit the ASCII character "SP" (32 decimal or 20 hex) should be used, while in the case of the Data Unit the ASCII "NUL" (0 decimal or 00 hex) will be used.

Both the Header and the Data Unit consist of logic blocks that need to be the size of 2,880 bytes or multiples thereof, whereby if the last logical block of one unit does not reach that dimension many characters (the ASCII character "SP" in the case of the header unit or the ASCII "NUL" in the case of the Data Unit) filling up to the exact size of 2,880 bytes or a multiple thereof must be added. In turn, each logic block contains 36 records, called keyword records, each 80 bytes in size, possibly supplemented with fill characters to reach that length. As can be seen, this is an extremely linear logical organization.

3.2 The strengths of the FITS file format

Currently FITS is the most used file format for scientific data in astronomy and is also widespread in the field of astrophysics.

One of the strengths of the format is certainly its extreme simplicity: the structure of a file encoded according to the FITS format is, in fact, extremely straightforward. Essentially a file in FITS format consists of two main components, the Header unit containing the metadata and the Data unit containing the actual data. This organization can be repeated indefinitely (there is no limit to the number of HDU Extension that can be added), making the format scalable. FITS has been designed specifically for scientific data and therefore provides the possibility to use the metadata to completely describe the object represented and the context of production.

Another key for its success is its flexibility. Despite the letter "I" in the name of the format (Flexible "I" mage Transport System) is the initial of the word "Image", actually FITS not only allows the exchange and storage of images but is of much more general use.[11] In fact, thanks to the Extension HDU mechanism, it is possible to store digital objects of various kinds (not only images but also spectra, photon lists, data cubes, multi-structured data such as multi-table databases etc.). There are no limits to the numbers of Extension HDU you can add to a FITS file (so there are no limits to the amount of data you can put in a FITS file). In addition, each of the extensions may contain different types of data objects. So, for example, it is possible to store x-ray and infrared exposures in the same file. As the term 'flexible' suggests, it is a very flexible file format for transporting, analyzing, archiving and preserving scientific data files (Greisen et al. 1981). In this sense FITS may be considered a general-purpose storage format for data, and thus can be a good solution also in the EO field.

Although FITS is the most commonly used digital file format in the astronomical and astrophysics field, in recent years, thanks to its flexibility, it has also been used in other fields, such as the cultural heritage sector. For example, the Vatican Apostolic Library adopted the FITS file format as the default format for storing and preserving images from its massive manuscript digitization project of the 80,000 most important manuscripts in the world (Allegrezza 2011; Ammenti 2012).

It was the first institution that decided to use the FITS file format for something different from scientific purposes: the preservation of cultural heritage (Manoni et al. 2018). Given the great success obtained with the digitization and permanent storage of several thousands of manuscripts, as well as the usage of state-of-art features in terms of online accessibility by adopting the International Image Interoperability Framework (IIIF), this seems to be the way forward for every serious cultural institution aiming to follow the best practices of the Vatican Apostolic Library. FITS also begins to be used in the health sector to store and preserve data from imaging systems (e.g. computerized tomography in nuclear medicine).[12]

11 In the specifications of the format there is no constraint which limits its use to images, therefore it may well be considered a general-purpose storage format for data.

12 The FITS format is widely used as an image format in the field of astronomy and astrophysics, but its use in digitization projects was certainly not envisioned by its

Another important factor for the widespread adoption of FITS is the ease of writing the code for its use, given by the fact that all the documentation has always been publicly available. As a result, a large amount of software has been developed throughout time, primarily for two purposes: displaying material and converting to and from the FITS data format. The importance of this feature cannot be overstated, because one of the main causes of data obsolescence is the direct link between a digital format and the private corporation that created it, which, once off the market, leaves future users without the necessary tools to run it.

4.0 Evaluation of FITS as a digital preservation file format

In order to evaluate this format against digital preservation requirements, it is useful to consider a series of requirements a file format should have to be considered a "preservation format". In particular, we will evaluate FITS file format against the sustainability factors proposed by the U.S. Library of Congress (Library of Congress 2021), with a few adjustments:

- *non-proprietary*: FITS format is managed by the international scientific community (in particular, the astronomers and astrophysics community) and, as such, it is not proprietary. The retention of specifications is currently borne by the IAU FITS Working Group;
- *standardization*: FITS is a standard format recognized since 1993; in addition it is an open standard, being a format that has both the properties of standardization and openness;
- *disclosure*: FITS is an open format, being based on a specification called "The FITS Standard" that is freely available; there are no royalties (and never will be) to pay for the use of the format, so anyone can create applications that can create and manage FITS files;
- *documentation*: FITS format is fully documented in its specification, "The FITS Standard";
- *adoption*: The FITS format is currently the most commonly used format in astronomy; there are also numerous applications for creating and viewing files in FITS format and for their conversion from and to other image formats;

creators. This is probably due to the fact that among the many unresolved problems involved in preserving cultural heritage in digital format (fragile storage media, hardware and software obsolescence, obsolescence of electronic formats, etc.) the FITS format seems to promise a solution to at least one of the main obstacles encountered in digital preservation, the obsolescence of electronic formats.

- *licensing & patents*: FITS format has no restrictions on its use neither now nor in the future; there are no royalties (and never there will be) to pay for the use of the format, so anyone can use it and create software applications able to manage FITS files;
- *transparency*: one of the main features of the FITS format is the fact that metadata are stored in human and readable ASCII format, so anyone can easily examine, using a simple text editor, the header unit of the file to get information on the file itself; furthermore, the information contained in the Data Unit are stored in a very simple way, without any compression; due to its simplicity it is possible, and will be in the future, to read a file in FITS format using basic software tools;
- *self-documentation*: FITS file format is self-documenting; the keywords that appear in the Header Unit aim to exhaustively document the file and provide information such as size, origins, history data and anything else the creator wishes to enter; in addition to standard keywords it is also possible to use other keywords coined specifically to better describe the particular type of data that will be stored in the file;
- *self-containment*: all the information necessary to be able to correctly represent (on screen or print) a FITS file, are included in the file itself;
- *technical protection mechanisms*: the FITS format does not provide technical protection mechanisms; for example, encryption is not provided, nor are access control mechanisms such as usernames and/or passwords.

To better evaluate FITS format, it is possible to consider other requirements, such as the following:

- *robustness*: due to its simplicity and to the fact that it is not compressed, the corruption of certain sequences of bits does not produce as a rule the loss of information content of the file and the degradation of the image remains generally within acceptable levels;
- *stability*: FITS format is very stable: over more than forty years only four versions have been released;
- *backward and forward compatibility*: the format is compatible backward and forward. Any changes to the format that make it no longer a valid FITS file are not allowed;
- *device independence*: being a multi-platform format, a file in FITS format can be viewed, printed or otherwise reproduced in a reliable and consistent way regardless of platform hardware and software used;
- *accessibility*: it is possible to define a keyword to define the "alternative text" that can be read by the visualization software to describe the image, and this is very useful in cases where the image is enjoyed by persons with impaired vision;

- *unmodifiability*: FITS can be modified using specially designed software (FITS editor). However, if necessary, its modifiability cannot be ensured by the use of technical measures such as using checksums or digital signatures, etc.;
- *security*: in the current state of knowledge, FITS cannot contain viruses or other forms of malware;
- *efficiency*: FITS is non-compressed, so the size of files encoded in this format are often quite high.

In summary, FITS has many features that make it a very attractive format for digital preservation purposes.

5.0 Further developments

Although FITS is currently the de facto standard for data storage and transmission in astronomy and other emerging fields, in recent years some scholars have begun to consider FITS outdated, based on the consideration that the choices made at the time of its writing, though common at the time, are now limiting the usefulness of the format itself. In their view, today's astronomy relies more than ever on a diverse set of data models with complex metadata (Thomas et al. 2015). As a result, in 2015 the Advanced Scientific Data Format (or ASDF) was released under the patronage of the Space Telescope Science Institute (Greenfield et al. 2015). Its aim is to be, just like FITS, a hybrid text and binary format, containing both human editable metadata for interchange. It is designed for extensibility and structurally complex, and raw binary data, fast to be loaded and used. The overall objectives of ASDF are very ambitious:[13]

- it has a hierarchical metadata structure, made up of basic dynamic data types such as strings, numbers, lists and mappings;
- it has human-readable metadata that can be edited directly in place in the file;
- the structure of the data can be automatically validated using schema;
- it is designed for extensibility: new conventions may be used without breaking backward compatibility with tools that do not understand those conventions. Versioning systems are used to prevent conflicting with alternative conventions;

13 Except from: "ASDF Standard. Introduction," last updated May 10, 2021, https://asdf-standard.readthedocs.io/en/1.6.0/intro.html. The latest version of ASDF standard (v. 1.6.0) is available at https://asdf-standard.readthedocs.io/en/1.6.0.

– the binary array data (when compression is not used) is a raw memory dump, and techniques such as memory mapping can be used to efficiently access it;
– it is possible to read and write the file in as a stream, without requiring random access;
– it is built on top of industry standards, such as YAML and JSON Schema to take advantage of a larger community working on the core problems of data representation. This also makes it easier to support ASDF in new programming languages and environments by building on top of existing libraries;
– since every ASDF file has the version of the specification to which it is written, it will be possible, through careful planning, to evolve the ASDF format over time, allowing for files that use new features while retaining backward compatibility with older tools.

The basic structure of ASDF files is straightforward, consisting of three sections:

– an *header* (mandatory) that indicates that the file of interest is in ASDF data format, stating clearly which version of the standard is used;
– a *tree* (optional), i.e. a YAML segment that provides a single structured view of all the data in the file. It is the main part of most ASDF files;
– zero or more *binary blocks* (optional) that are just a cluster of binary data, allowing for a simple and flexible type system.

According to its creators, ASDF is intended to go much beyond the cases previously handled by FITS by providing more descriptive keyword names, greater flexibility towards data values, an intrinsic hierarchical structure, and the possibility of sharing references to the same objects even between different elements. The flexibility guaranteed by the two data formats allows for a great deal of interchangeability: if a specific dataset must be produced in FITS, it is possible to embed ASDF in FITS, thanks to the use of the EXTNAME = 'ASDF' extension. Vice versa, it is also possible to store a FITS file in ASDF, given the fact that the structure of the latter is more complex than the former.[14] Giving these premises, this fairly young digital data format seems to have what it takes to be a worthy heir of FITS, but only time can tell if this will be the case.

14 See Appendix A to ASDF Specification: "Appendix A. Embedding ASDF in FITS," last updated May 10, 2021, https://asdf-standard.readthedocs. io/en/1.6.0/asdf_in_fits.html.

6.0 Conclusions

FITS is a versatile and general-purpose file format and can archive and preserve many types of research data, including bitmaps, ASCII text, binary tables, multidimensional arrays such as data cube (3D) (e.g. temperature varying with time) or hypercube (4D, 5D, …) (e.g. temperature varying with time and altitude). It is extensible: user communities can develop specific conventions that refine the FITS standard in order to adapt it to specific contexts. Thanks to its versatility and extensibility, somebody proposed renaming FITS as "Flexible Image and Table Systems" (instead of "Flexible Image Transport Systems").

FITS meets a number of desirable requirements for a format suitable for digital preservation, and it is also one of the few formats that has stood the test of time because, more than forty years after its inception, it is still widely used. FITS was designed with an eye towards long-term archival preservation, and the maxim "Once FITS, always FITS" that has been coined for it clearly establishes that developments to the format must be backward compatible and that FITS files will never be made obsolete by more recent versions.

Furthermore, the Italian Unification Body (UNI, Ente Nazionale di Unificazione) is in the process of publishing the standard "Management Processes for the long-term preservation of digital images using the FITS format" (code name: UNI1606754). This new standard, which is likely to be released in 2022, will certainly contribute to its greater adoption. At ISO (International Standardization Organization) there is an ongoing proposal for its standardization as well.

Thus, FITS should be seriously considered as a format for long-term archiving and preservation of scientific data, at least in some areas of the EO field.

References

Albani, Mirko. 2012. "Long term preservation of earth observation data: The challenge and the cooperation activities". In *IEEE International Geoscience and Remote Sensing Symposium*, (*IGARSS 2012), 22-27 July 2012*, 7279-82.

Albani, Sergio, and David Giaretta. 2009. "Long-term Preservation of Earth Observation Data and Knowledge in ESA through CASPAR". *International Journal of Digital Curation* 4, no. 3: 4-16.

Allegrezza, Stefano. 2011. "Analisi del formato FITS per la conservazione a lungo termine dei manoscritti. Il caso significativo del progetto della Biblioteca Apostolica Vaticana", *DigItalia*, no. 2.

Ammenti, Luciano, 2012, "BAV and the FITS (Flexible Image Transport System) format 40 Years of Experience in Long-Term Digital conservation". In *European Week of Astronomy and Space Science (EWASS), 1-6 July 2012 Rome, Italy*. https://www.vatlib.it/moduli/Ammenti_EWASS2012.pdf.

"Appendix A. Embedding ASDF in FITS." Last updated May 10, 2021. https://asdf-standard.readthedocs. io/en/1.6.0/asdf_in_fits.html.

"ASDF Standard. Introduction." Last updated May 10, 2021. https://asdf-standard.readthedocs.io/en/1.6.0/intro.html.

Conway, Esther, S. Pepler, Wendy Garland, D. Hooper, F. Marelli, L. Liberti, E. Piervitali, K. Molch, H. Glaves, and L. Badiali. 2014. "Ensuring the Long Term Impact of Earth Science Data through Data Curation and Preservation". *Information Standards Quarterly* 25, no. 3: 28. https://doi: 10.3789/isqv25no3.2013.05.

COSMO. "GRIB Data Format used for tge COSMO-Model System Features of GRIB." Last accessed December 13, 2021. http://www.cosmo-model.org/content/model/documentation/grib/grib_ features.htm.

ESA (European Space Agency), and CEOS/WGISS. 2015. "Earth Observation Preserved Data Set Content (PDSC)", https://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Preserved%20Data%20Set% 20Content_v1.0.pdf.

ESA (European Space Agency). 2015. "Long Term Preservation of Earth Observation Space Data. Preservation Guidelines". https://earth.esa.int/eogateway/documents/20142/37627/ EO-Data-Preservation-Guidelines.pdf.

ESA, "SAFE." Last accessed December 13, 2021. https://earth.esa.int/SAFE/index.html.

Greenfield, Perry E., Michael Droettboom, and Erik M. Bray. 2015. "ASDF: A new data format for astronomy". *Astronomy and Computing* 12: 240-51. https://www.sciencedirect.com/science/article/pii/S2213133715000645.

Greisen, Eric W., and Ronald H. Harten.1981. "An Extension of FITS for Groups of Small Arrays of Data". *Astronomy & Astrophysics Supplement Series* 44, 371.

Greisen, Eric W., Donald C. Wells, and Ronald H. Harten. 1980. "The FITS Tape Formats: Flexible Image Transport Systems". In *Applications of Digital Image Processing to Astronomy*, *Proceedings*, edited by D.A. Elliott. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series 264, 298. https://doi.org/10.1117/12.959819.

Grosbol, Preben. 1988. "The FITS Data Format". *Bulletin d'Information du Centre de Donnees Stellaires*, no. 35: 7. https://adsabs.harvard.edu/full/1988BICDS..35....7G.

Hanisch, Robert J., A. Farris, E. W. Greisen, W. D. Pence, B. M. Schlesinger, P. J. Teuben, R. W. Thompson, and A. Warnock III. 2001. "Definition of the Flexible Image Transport System (FITS)". *Astronomy & Astrophysics* 376: 359–380. https://doi:10.1051/0004-6361:20010923.

ISO (International Standardization Organization). 2020. "ISO 19165-2:2020 — Geographic information — Preservation of digital data and metadata — Part 2: Content specifications for Earth observation data and derived digital products". https://www.iso.org/standard/ 73810.html.

Library of Congress. 2021. "Sustainability of Digital Formats: Planning for Library of Congress Collections", https://www.loc.gov/preservation/digital/formats.

Manoni, Paola, Ángela Núñez Gaitán, and Irmgard Schuler. 2018. "The Vatican Library's Digital Preservation Project", Paper presented at: IFLA WLIC 2018 – Kuala Lumpur, Malaysia – Transform Libraries, Transform Societies in Session 160 – Preservation and Conservation with Information Technology. http://library.ifla.org/id/eprint/2113/1/160-manoni-en.pdf.

Molch, Katrin, Rosemarie Leone, Mirko Albani, and Eberhard Mikusch. 2012. "User needs and requirements impacting the long term preservation of earth observation data". In *IEEE International Geoscience and Remote Sensing Symposium, (IGARSS 2012)*: 7283-85, https://doi: 10.1109/IGARSS.2012.6351980.

NASA. "FITS Standard Document." Last accessed December 13, 2021. https://fits.gsfc.nasa.gov/fits_standard.html.

NASA. "Space Physics Data Facility." Last accessed December 13, 2021. https://cdf.gsfc.nasa.gov/html/FAQ.html.

NASA. 2011. "NASA Earth Science Preservation Content Specification". Greenbelt, Maryland: Goddard Space Flight Center, https://cdn.earthdata.nasa.gov/conduit/upload/10607/NASA_ESD_Preservation_Spec.pdf.

NASA. 2014. "A Primer on the FITS Data Format." Last revised October 28, 2014. https://fits.gsfc.nasa.gov/fits_primer.html.

NASA/GSFC Astrophysics Data Facility. 1997. "A User's Guide for the Flexible Image Transport System (FITS). Version 4.0, 1997", April 14, 1997. https://fits.gsfc.nasa.gov/users_guide/usersguide.pdf.

National Research Council. 1995. "Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources". Washington, DC: The National Academies Press, 11. https://doi.org/10.17226/4871.

NCEI. "Global Forecast System." Last accessed December 13, 2021. https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast.

Pinna, Gian Maria, and Stéphane Mbaye. 2014. "SAFE Archive for ENVISAT and Third Party Mission Products". July 17, 2014. https://earth.esa.int/workshops/envisatsymposium/authors/CXNL_07A03_507957.html.

Ramapriyan, Hampapuram, Dawn Lowe, and Kevin Murphy. 2017. "Towards a preservation content standard for earth observation data". In *Proceedings of the 2017 conference on Big Data from Space (BiDS' 2017) 28th-30th November 2017 Toulouse (France)*. https://ntrs.nasa.gov/ citations/20170011347.

STSI. "ASDF Standard." Last accessed December 13, 2021. https://asdf-standard.readthedocs.io/en/1.6.0.

Thomas, Brian, Tim Jenness, Frossie Economou, Perry Greenfield, Paul Hirst, David S. Berry, Erik Bray, Norman Gray, Demitri Muna, and James Turner. 2015. "Learning from FITS: Limitations in use in modern astronomical research", *Astronomy and Computing* 12, 133-45. https://arxiv.org/abs/1502.00996.

Unidata. "Network Common data Form (NetCDF)." Last accessed December 13, 2021. https://www.unidata.ucar.edu/software/netcdf.

Wells, Donald C., and Eric W. Greisen. 1979. "FITS: a Flexible Image Transport System". In *Image Processing in Astronomy*, edited by G. Sedmak, M. Capaccioli, R. J. Allen, 445.

Wells, Donald C., Eric W. Greisen, and Ronald H. Harten. 1981. "FITS: a Flexible Image Transport System", *Astronomy & Astrophysics Supplement Series*, no. 44. https://articles.adsabs.harvard.edu/pdf/1981A%26AS...44..363W.