Flexible Image Transport System: a new standard file format for long-term preservation projects?

Roma, 5.7.2012   Stefano Allegrezza

Agenda

❶ Introduction

❷ Evaluation of the FITS format

❸ Evaluation of the FITS format: general requirements

❹ Evaluation of the FITS format: specific requirements

❺ Conclusions

❶

Introduction

Introduction

Questions to answer:

- "Is FITS file format suitable for digital preservation"?

- "Can it become the standard file format to use in digitization projects"?

Why FITS?

- Today we have many image formats:
  - TIFF
  - JPG
  - JPG2000
  - DjVu
  - ….
- Now there is another "image" file format:
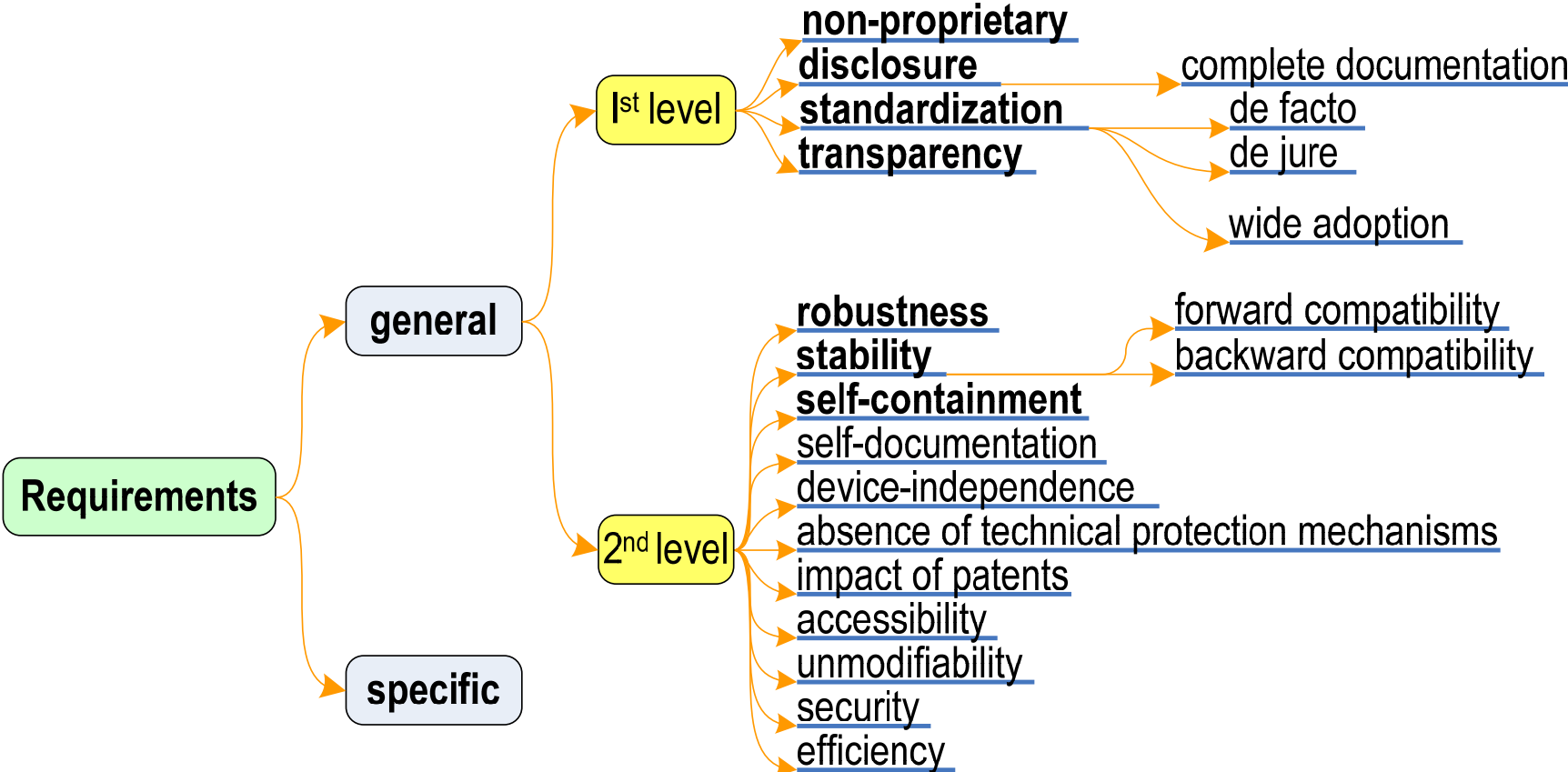 - FITS

❷ Evaluation of the FITS format

Introduction

- My report aims at presenting the results of a study carried out at the CEIDIM Research Center of the University of Macerata (Italy)
- This study identified the factors we must consider when we need to choose a format suitable for permanent preservation
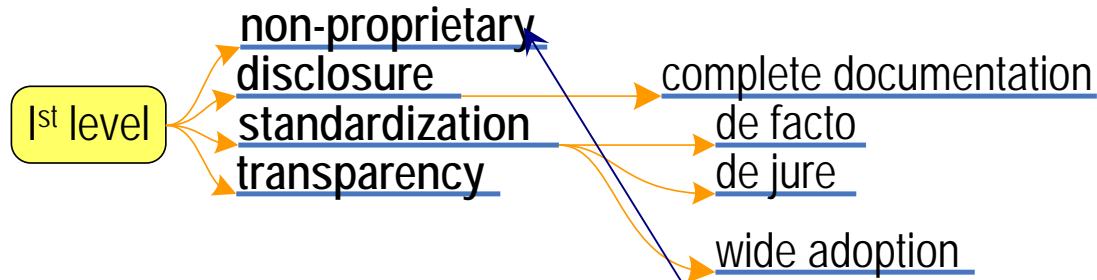
Requirements to be evaluated

❸

Evaluation of the FITS file format: general requirements, first level

Non-proprietary

Ist level
- non-proprietary
- disclosure → complete documentation
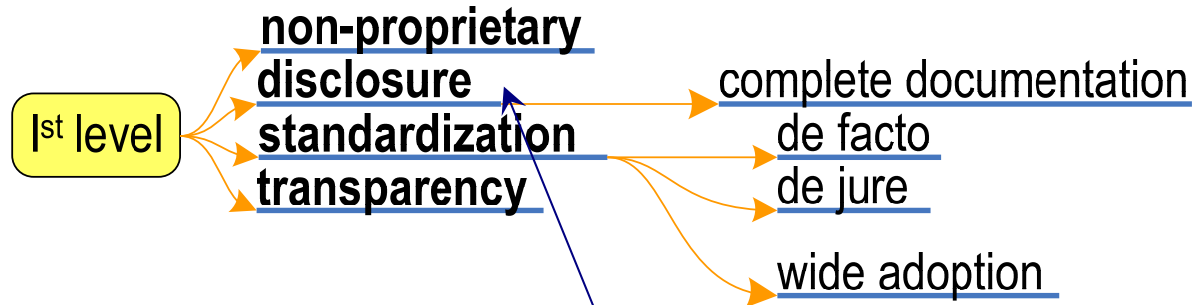- standardization → de facto
- transparency → de jure
- wide adoption

FITS format is mantained by the international scientific community (in particular, the astronomers and astrophysics community) and, as such, is **not proprietary**. The specifications are currently mantained by the **IAU FITS Working Group**

Disclosure (Openness)

Ist level

non-proprietary
disclosure
standardization
transparency

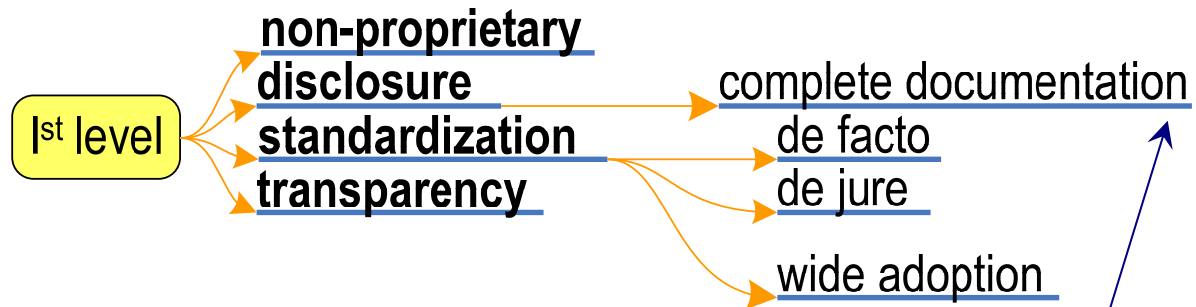complete documentation
de facto
de jure

wide adoption

FITS is an **open** format; infact, it is based on a specification, the "**FITS Reference**", publicly available.

There are not royalties (and never will be) to pay for the use of the format, so anyone can create software applications able to manage FITS files.

Complete documentation

Ist level

- **non-proprietary**
- **disclosure** → complete documentation
- **standardization** → de facto
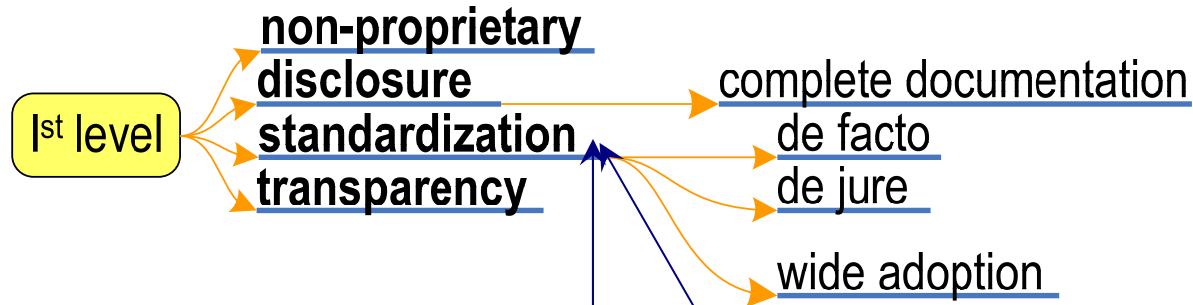- **transparency** → de jure

→ wide adoption

The FITS format is **fully documented** in the **FITS Reference**, freely available. The official reference document that defines the requirements for FITS format version 3.0 version is published in *Astronomy and Astrophysics*, Volume 524 (December 2010)

Standardizazion

Ist level

- **non-proprietary**
- **disclosure** → complete documentation
- **standardization** → de facto
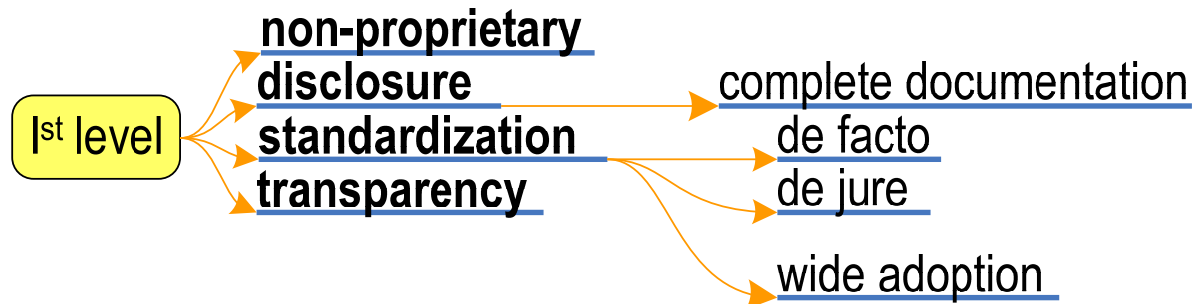- **transparency** → de jure
- wide adoption

FITS is a **standard** format recognized by the NOST since 1993.

Furthermore FITS is an **open standard** since it has both the properties of standardization and disclosure.

Wide adoption

non-proprietary
disclosure   complete documentation
Iˢᵗ level   **standardization**   de facto
transparency   de jure

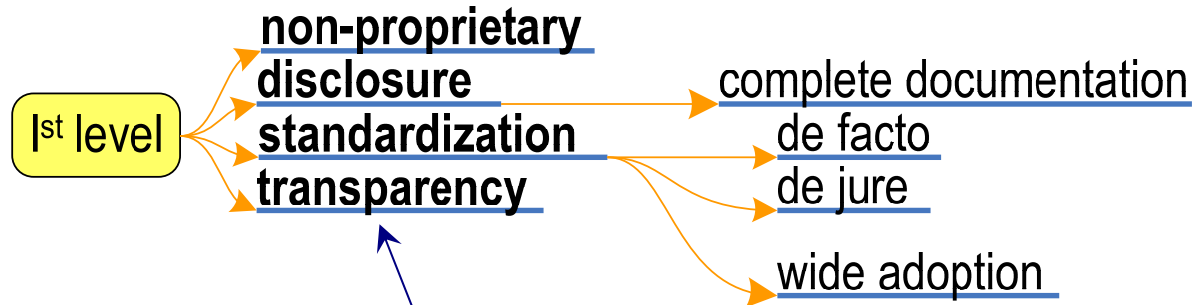   wide adoption

FITS format is currently the most commonly used format in astronomy.

There are many software applications for creating and viewing files in FITS format and for their conversion from and to other image formats.

Transparency

Ist level

- **non-proprietary**
- **disclosure** → complete documentation
- **standardization** → de facto
- **transparency** → de jure
- → wide adoption

Each FITS file consists of a header containing ASCII *card images* (80 character fixed-length strings) that carry keyword/value pairs and an image data block.

Furthermore, the information relating to the pixels of the images (contained in the Data Unit) are stored in a very simple way, without any compression; due to its simplicity it is possible (now and in the future) to read a file in FITS format using basic software tools.

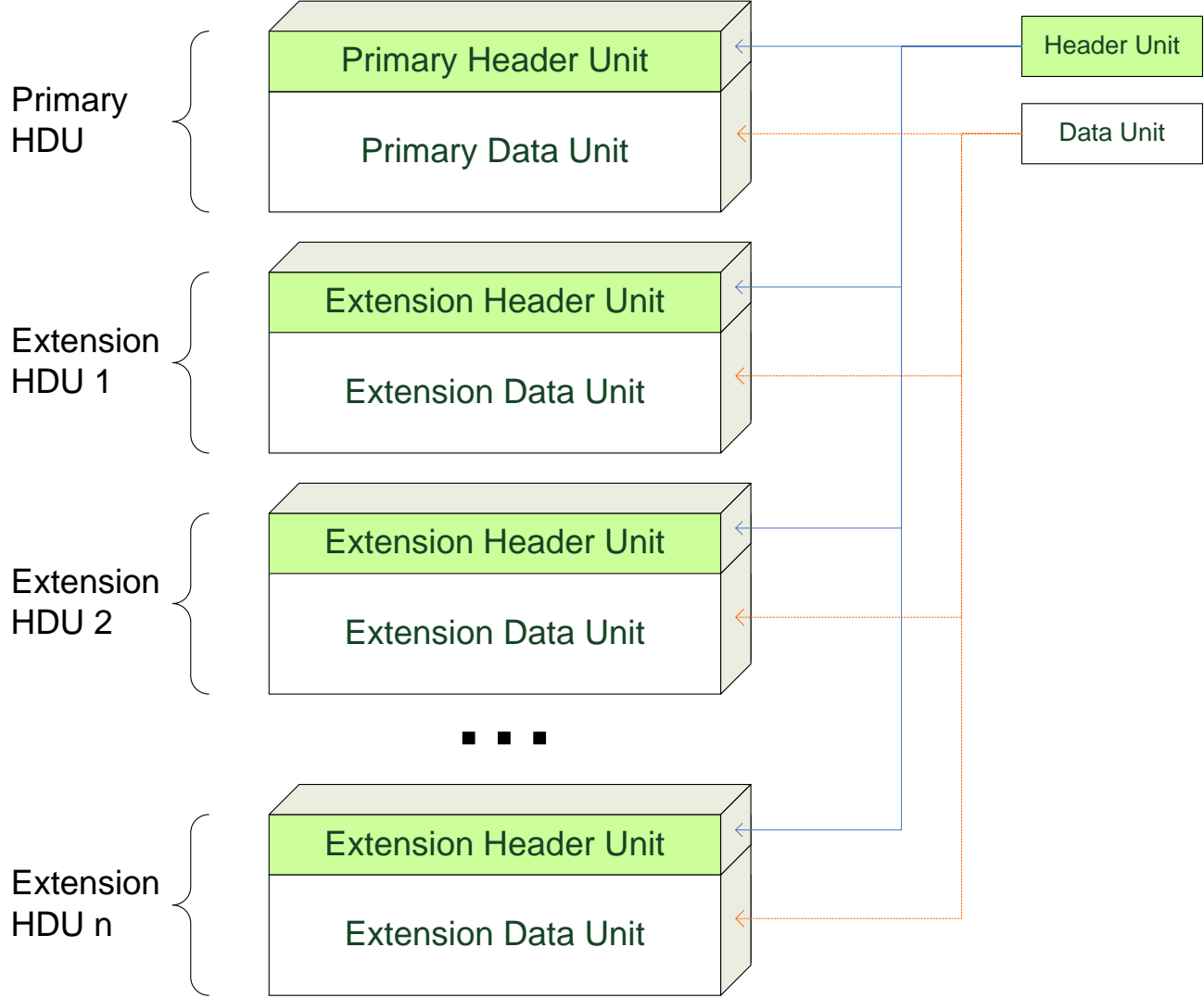# Transparency: a look at technical issues

- One of the strengths of the format can be found certainly in its **simplicity**: the structure of a FITS file is, in fact, extremely **straight**.
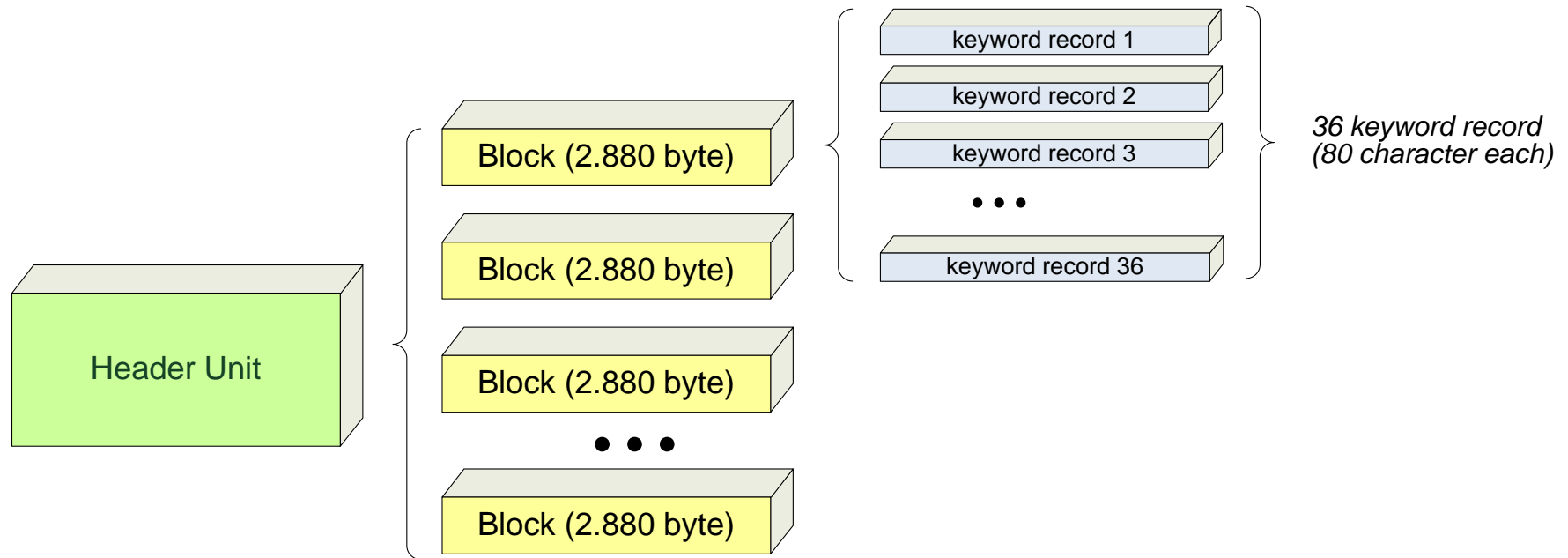
The structure of a FITS file

The structure of a FITS file

A simple FITS file

```
       1         2         3         4         5         6         7         8
1234567890123456789012345678901234567890123456789012345678901234567890
```

```
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    = '02/12/2011'         / creation date
ORIGIN  = 'Digitalia'          / FITS file originator
AUTHOR  = 'Mario Rossi'        / creator
OBJECT  = 'Manoscritto_1'      / name of object
HISTORY    'first version'
COMMENT    'Example of Header Unit'
END
```

Header Unit

```
?„»?Šó?[1]Ù>Ðó¢?h?qß<?•^?~É?¯Ux>†-$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@h t@E£@RË^@y>O@§¦‚@¡@ÐµÕ@òhmA[1]éq@ÚÉ•@Ú¼sA7ôõAzb€AfxFAV9LAD[¿Aó1A‚˜†A2I>A4ò'
A}(A±¬>A§TÚAq³AHjZAMe+A}¬AŠéAž[1]$A¯ óA³/AÀA AÑúÂÂë•|B(LÍB1˜B)ä£BAûTÇA¥‚úAKE|A.0¢A ß‚Aö
@â¤Ï@äZ@¨Áš?¢ ?L™š>_u'?;dâ>®á‡>RT>ãr"?Ké?D"?°¢>ß3?V™8?Ql?€¦†?ĕ[>mĕw? <ì?¬>X? v?bœÔ>à3?˜
?; ‰?is?Ž »?°é³? '+?>m?gt1>ËÉS><pw?GXt?€#4? #¡?°â7?¿8?1X0?DRç?Œ‚b?ÜkV?ËuÔ?¬(™?Šùþ?œŸ¬?ê
[1]Ò?äy:?Ì~¦?ç¿€?Éì|?¹dÐ?ÃÍƒ?¬|p@Ÿ @5ü?èõsdjklaakkal
[…]
Ë±@ Ôà@Q£@'Œ@†vì@šÈÝ@¡Òž@¨ŽŒ@œ§9@±åÝ@è"=@ë Ê@þ´@áq|AÀA'¬œA†$ÀAËFÂA¿†õA˜š¨A ÈA^
ßA²¿oAÍ ±Añ:IB $B[1]…B ™7B„õB+>iB\^B|B{Ú3B þ BÙ[1]A¯ ‚A^dKAd°AÑoAã9@óµ‚A ¼Ÿ@Ý¡!?ß?Þ°Ž?H
j…?q³Ë>Ð ©?šDÃ?Xaá@ô>äú´?¨^Å?Àü>h³?Š ^>Û^Y>£¶Ä?Ò>Ð5»?_„ ?^oÁ?‰‰?Y‚?'Å>õ^¨?) Ö?¡j*?«j2?¢
g?ƒàé?¥ ?ž©Ö?A^-?AÃ?Ü×ò?°£G?©®ü?>V?[1] ?>ûÄ@¹€?Ïùê?7Š ?Œál?Yjó?³zW>ëL:=ƒK?kJç?£ŽS?Ø ?¢
?\?†°1?"RÍ?>•Á?Ï‰"?µ¾¤?• 4?"ih?‚RD?†Û¨?…v¦?$ëá?ÒÏd?Óqf?Õä$?õòÕ?¿± ?£m?ãR¿½¡Ð1?<‚<?¯¾p?‚
ƒ•?‡tO@ -Ý>Î\¯?{&Î? ??90Ü? l<?‰o0? $ô>-}?«6¶?µ¢Æ?˜‡"?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
```

Data Unit

Evaluation of the format:
general requirements, second level

Robustness

robustness
stability
self-containment
self-documentation
device-independence
absence of technical protection mechanisms
impact of patents
accessibility
unmodifiability
security
efficiency

2nd level

forward compatibility
backward compatibility
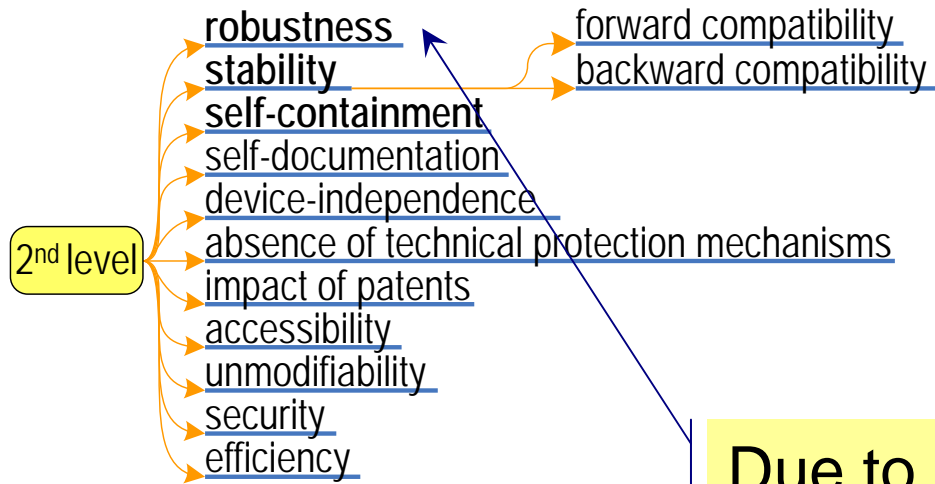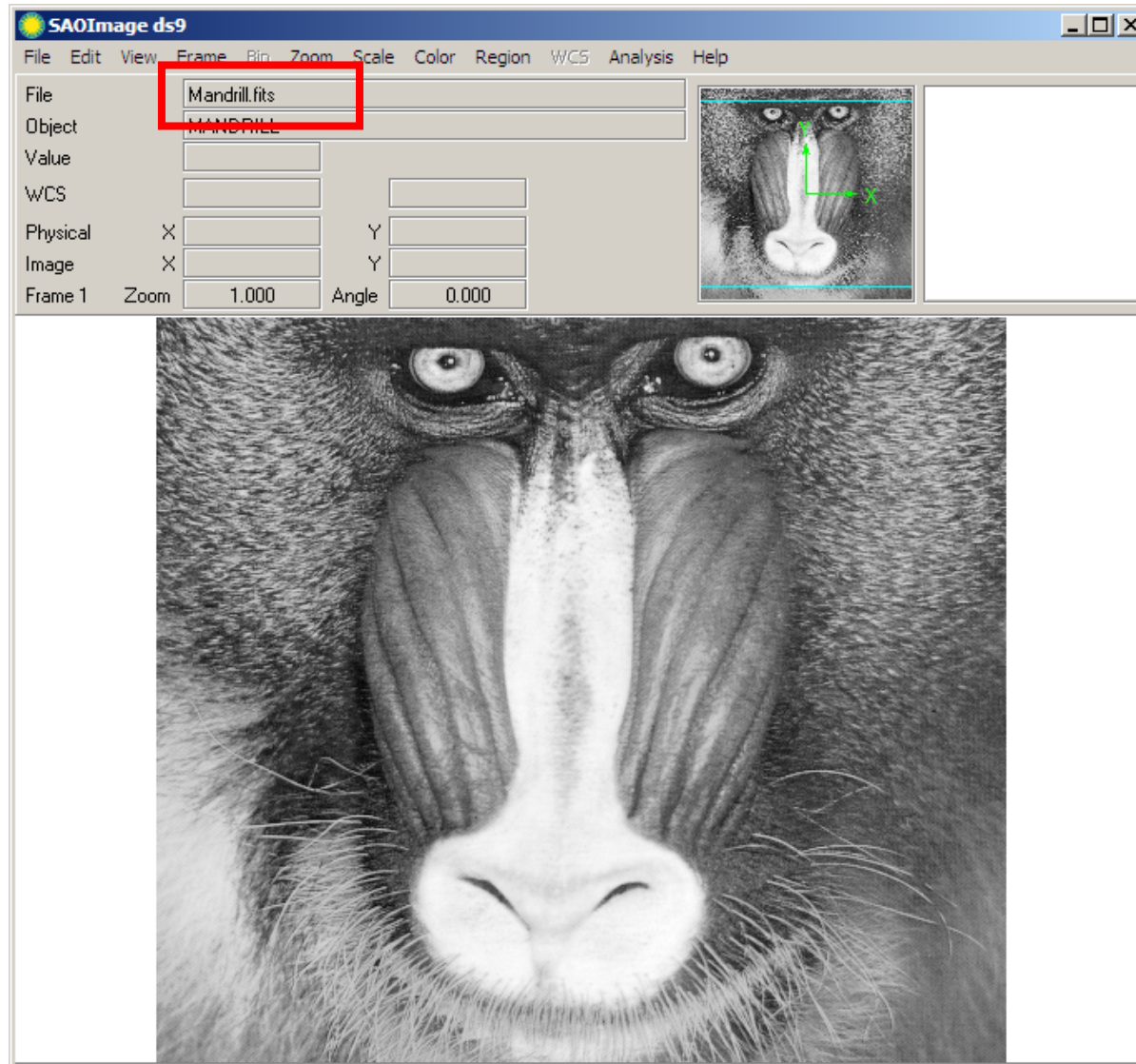
Due to its simplicity and the fact that it is not compressed, the **corruption** of certain sequences of bits does not produce, as a rule, the loss of information content and the image degradation generally remains within acceptable levels.
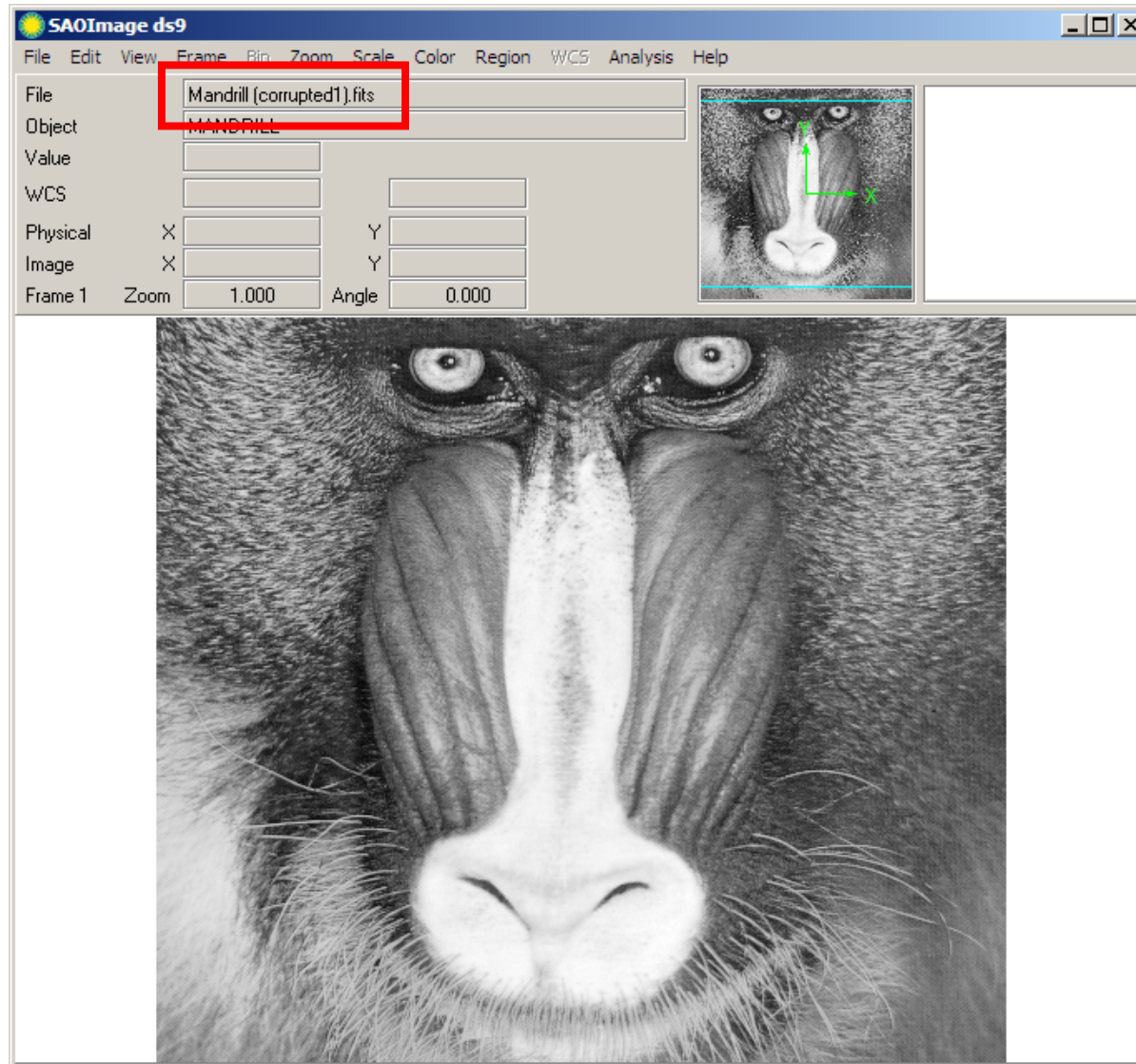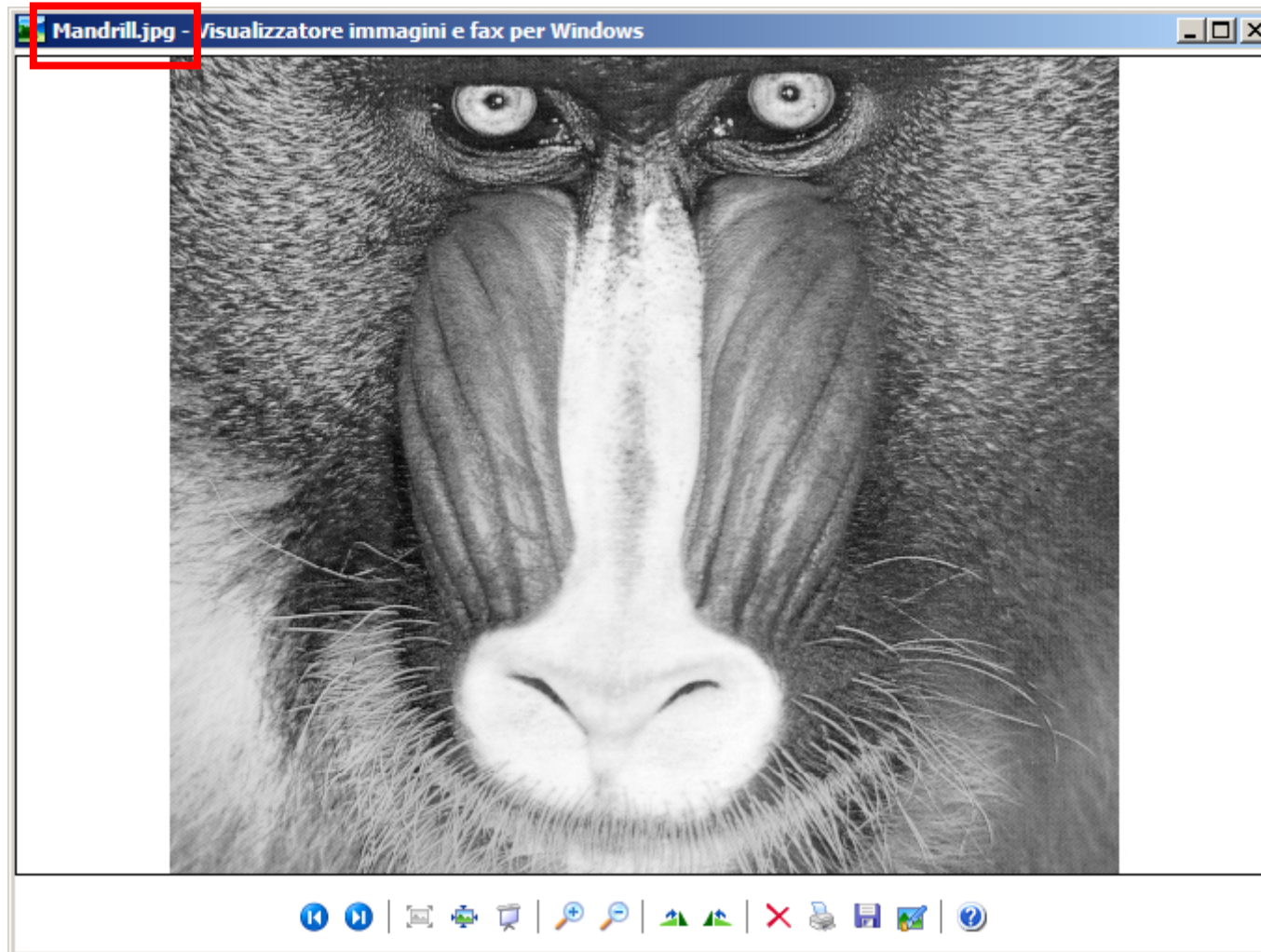
Example 1 (FITS)

Mandrill.fits

Example 1 (FITS)
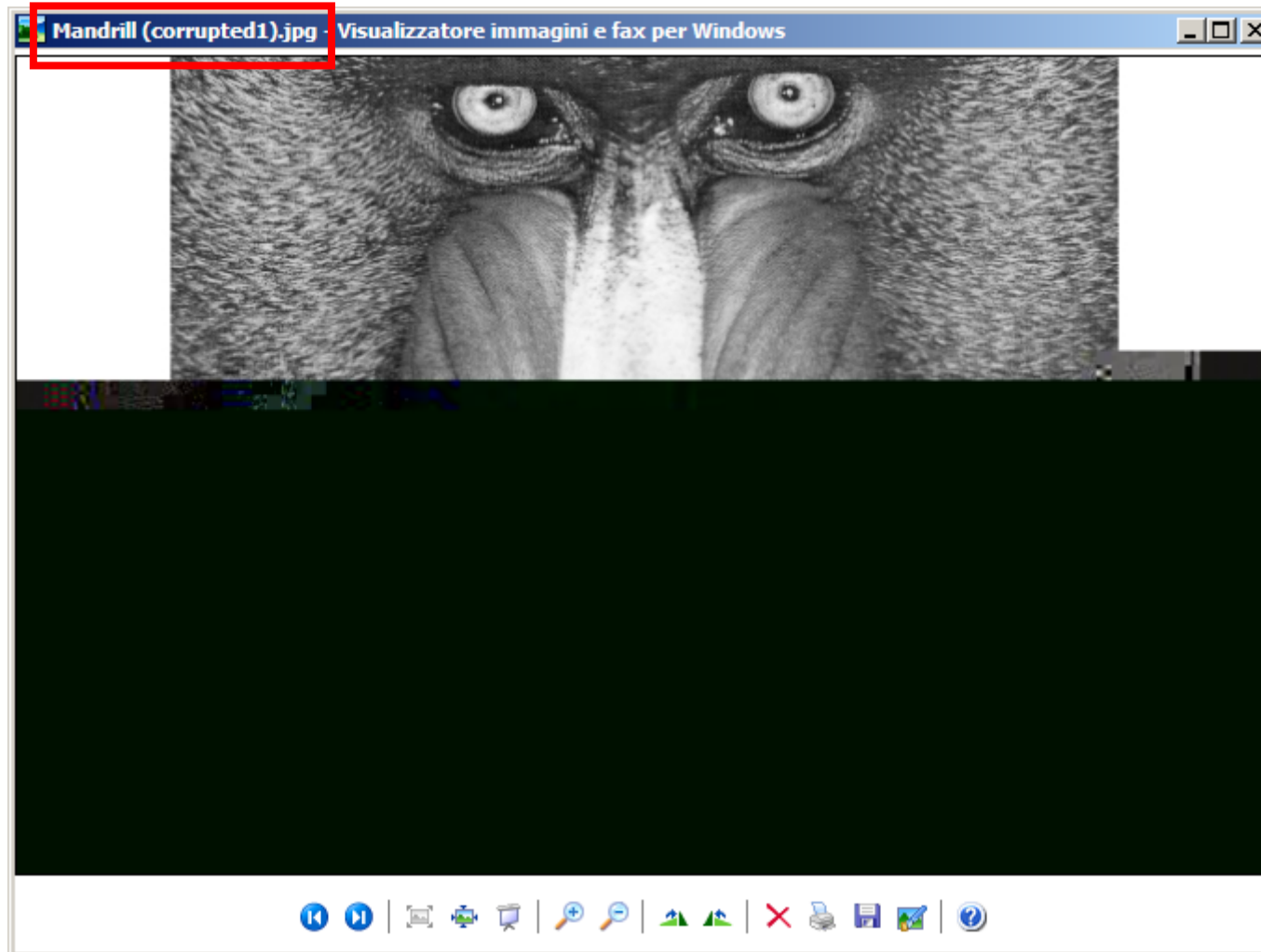

Mandrill (corrupted1).fits

Example 2 (JPG)

Mandrill.jpg

Example 2 (JPG)

Mandrill (corrupted_1).jpg

Stability

robustness
stability → forward compatibility
                backward compatibility
self-containment
self-documentation
device-independence
absence of technical protection mechanisms
impact of patents
accessibility
unmodifiability
security
efficiency

2nd level
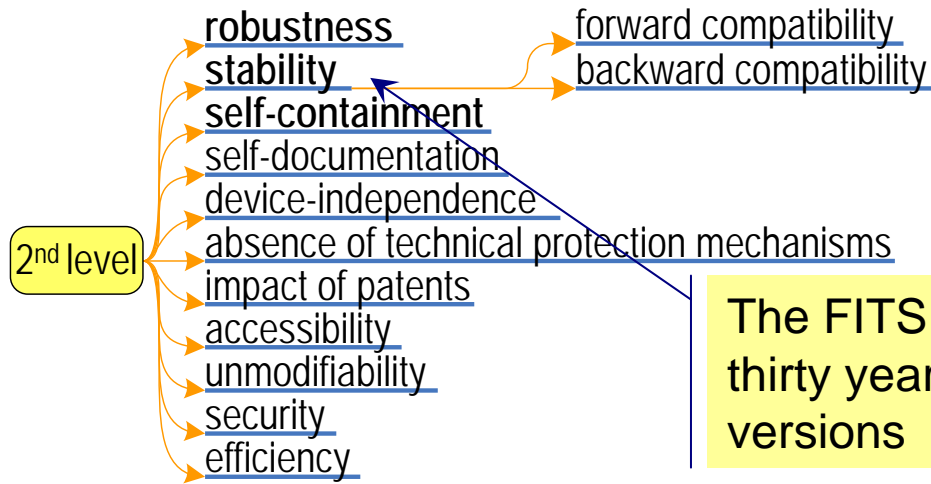
The FITS format is very stable: over more than thirty years there have been only **three** official versions

Version	Release date	Notes
-	1979	Birth of the format
NOST 100-0.1	December 1990	First draft
NOST 100-0.2	June 1991	Second draft
NOST 100-0.3	December 1991	Third draft
NOST 100-1.0	June 1993	NOST Standard NOST (first version)
NOST 100-1.1	September 1995	NOST Standard
NOST 100-1.2	April 1998	Some minor amendments
NOST 100-2.0	March 1999	NOST Standard NOST (second version)
IAUFWG 2.1	April 2005	IAUFWG Standard
IAUFWG 2.1b	December 2005	IAUFWG Standard (added 64 bit supportt)
IAUFWG 3.0	July 2008	IAUFWG Standard (third version)

Backward and forward compatibility

robustness → forward compatibility

stability → backward compatibility

self-containment

self-documentation

device-independence

absence of technical protection mechanisms

2nd level → impact of patents

accessibility

unmodifiability
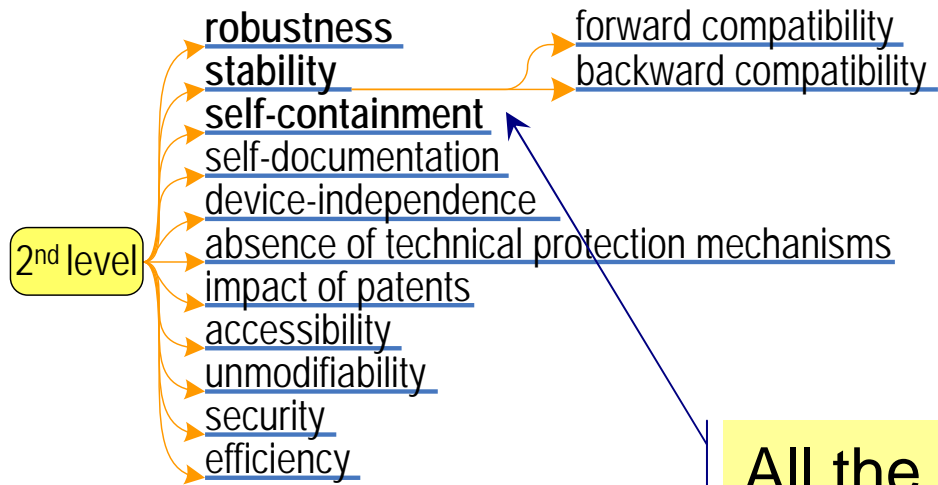
security

efficiency
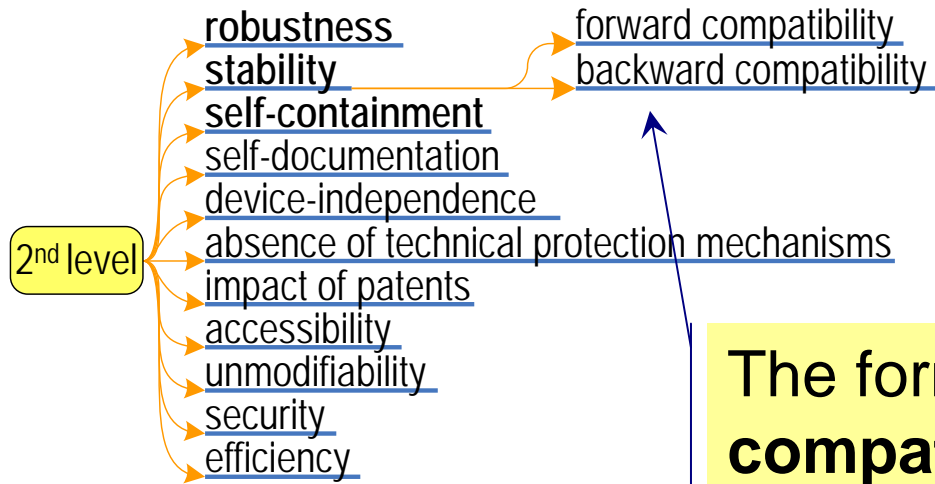
The format is **backward** and **forward compatible**.

FITS has been designed keeping in mind the aspect of long-term preservation.

The maxim "**once FITS, always FITS**" that has been coined for it clearly establishes that any changes to the format that makes a FITS file not longer valid is not allowed.

Self-containment

2nd level
- **robustness**
- **stability**
  - forward compatibility
  - backward compatibility
- **self-containment**
- self-documentation
- device-independence
- absence of technical protection mechanisms
- impact of patents
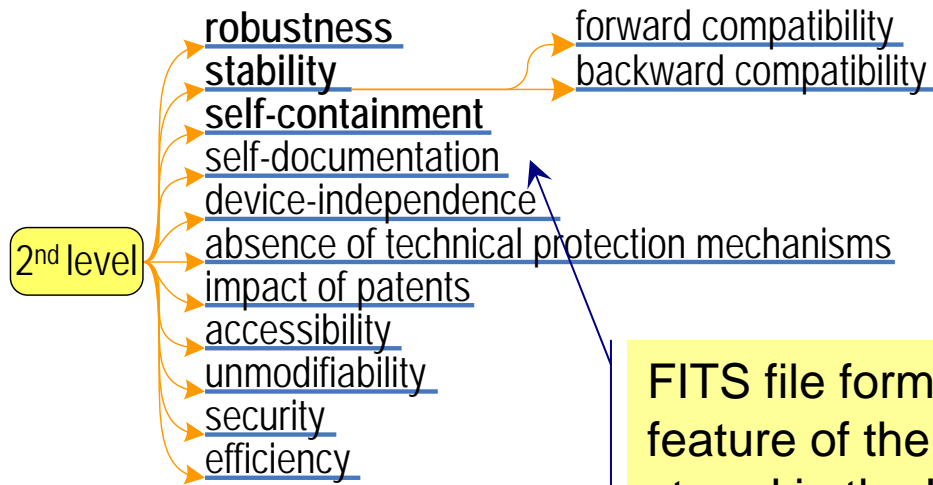- accessibility
- unmodifiability
- security
- efficiency

All the information necessary to correctly represent (on screen or print) a FITS file, are included in it the file itself.
Links to external resources are not allowed.

Self-documentation

- robustness
- stability
  - forward compatibility
  - backward compatibility
- self-containment
- self-documentation
- device-independence
- absence of technical protection mechanisms
- impact of patents
- accessibility
- unmodifiability
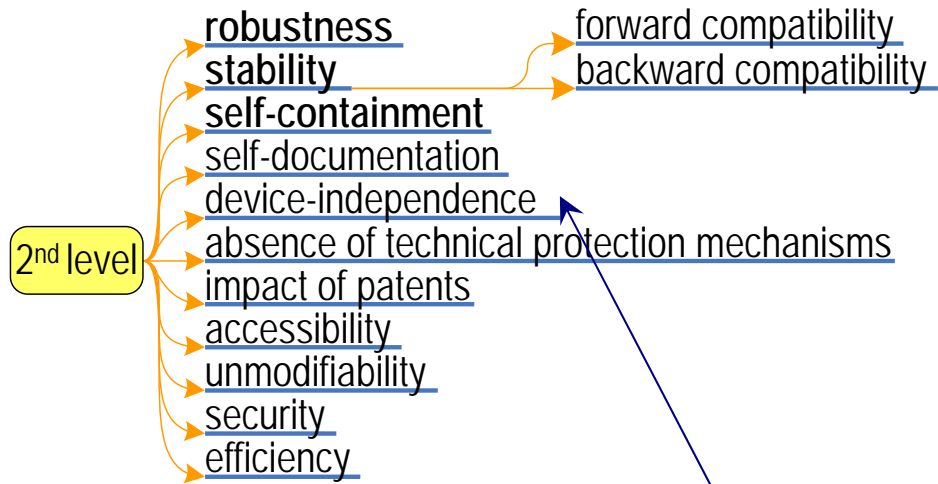- security
- efficiency

2nd level

FITS file format is self-documented. In fact, a major feature of the FITS format is that meta-data are stored in the Header Unit in a human-readable ASCII format so anyone can easily examine, using a simple text editor, the header unit to get information on the file itself.

These keywords aim to document the file and provide information such as size, origins, history data and anything else the creator wishes. Furthermore, in addition to standard keyword is also possible to create **custom keywords** to better describe the particular type of data that will be stored in the file.

Device independence

robustness
stability → forward compatibility
backward compatibility

self-containment
self-documentation
2nd level → device-independence
absence of technical protection mechanisms
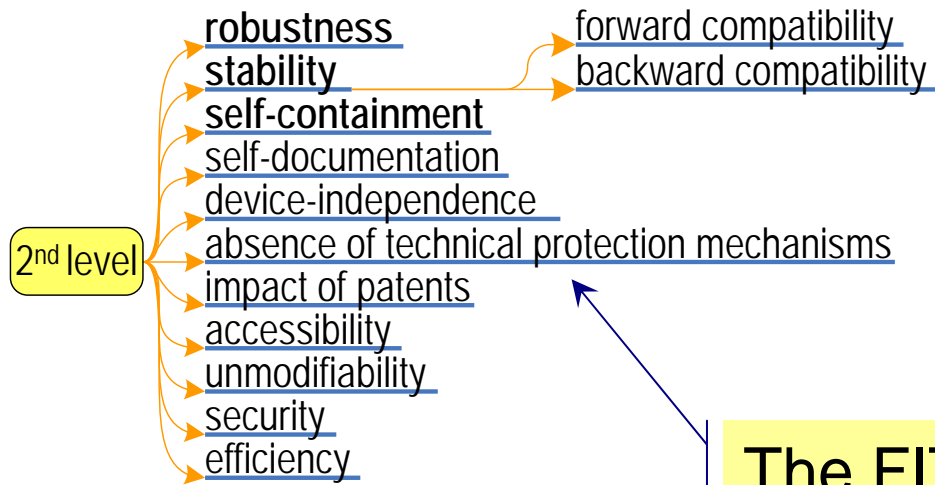impact of patents
accessibility
unmodifiability
security
efficiency

FITS is a multi-platform format, a file in FITS format can be viewed, printed or otherwise reproduced in a reliable and consistent way regardless of hardware and software used.

Absence of technical protection mechanisms

robustness   forward compatibility

stability   backward compatibility

self-containment

self-documentation

device-independence

2nd level — absence of technical protection mechanisms

impact of patents

accessibility
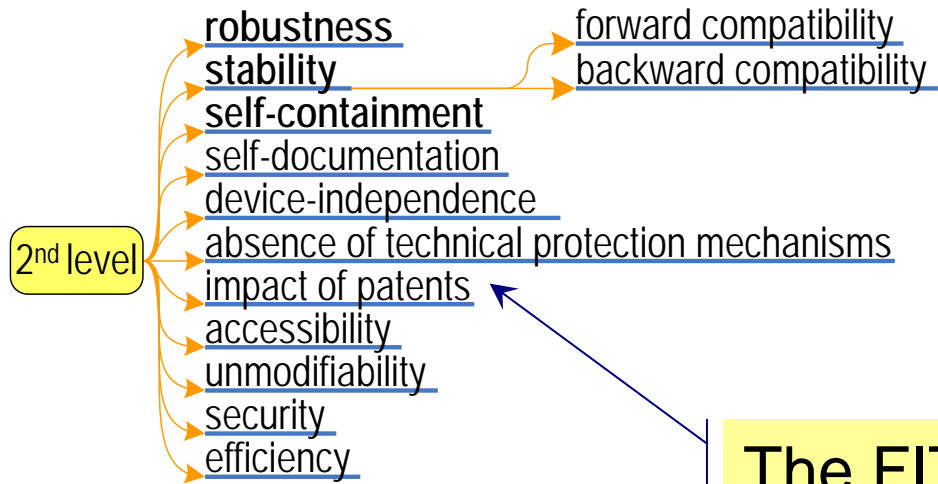
unmodifiability

security

efficiency

The FITS format does not provide technical protection mechanisms.

Is not provided, for example, encryption, as well as control of access through username and / or password.

Impact of patents

robustness   forward compatibility

stability   backward compatibility

self-containment

self-documentation

device-independence

absence of technical protection mechanisms

2nd level

impact of patents

accessibility

unmodifiability

security

efficiency

The FITS format has no restrictions on its use neither now nor in the future.

Other requirements

- **Accessibility**. It is possible to define a keyword to specify 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**. The FITS can be modified using specially designed software (FITS editor). However, if necessary, its modifiability can be ensured by the use of technical measures such as using checksums or digital signatures, etc.

- **Security**. In the current state of knowledge, FITS can not contain viruses or other forms of malware.

- **Efficiency**. The FITS format is not compressed, so files encoded in this format have often a a larger size than file encoded in other (compressed) formats.

④ Evaluation of the FITS format: specific requirements

Resolution

- FITS is capable of storing images absolutely faithful to the original, with regard both to the **resolution** and to the **color depth**.

- In fact, theoretically there is **no limit** to the resolution of the images that can be saved.

The Vatican Library is using equipment to capture high-resolution (**50 megapixels**) images, but if in the future should be available scanners and cameras with higher resolutions (eg , **500 megapixel**) there would be no difficulty in continuing to use the FITS format.

Color depth

- The BITPIX keyword specifies the number of bits used to represent the value of each data. In the case of data relating to images, the value of BITPIX is the number of bits used to represent each pixel of the image (ie, the so-called "color depth"). Possible values are summarized in the above table:

BITPIX value	Data type
8	Character or unsigned integer
16	16-bit signed integer
32	32-bit signed integer
64	64-bit signed integer
-32	32-bit floating point, single precision
-64	64-bit floating point, double precision

Color depth

- FITS is able to keep intact the quality of colors
- For example, it is possible to specify a **color depth** of **64 bits** (equivalent to more than 18 billions of billions of colors!)
- In practice,color depth is almost unlimited and certainly superior to the ability of the human eye to distinguish so subtle color differences.
- In this way it is posssible to store images without any loss of color (think of the colors of the miniatures such as the gold color very hard to reproduce) and with the certainty that the colors will be faithful to the originals stored.

Other features

- Unlike other image formats, FITS format is able to store multiple images in the same file (so it would be possible, in principle, storing an entire manuscript in one file).

- FITS is not only a format for images

- FITS is a very general format capable of storing many types of data, including bitmaps, ASCII text, multidimensional matrices, and binary tables.
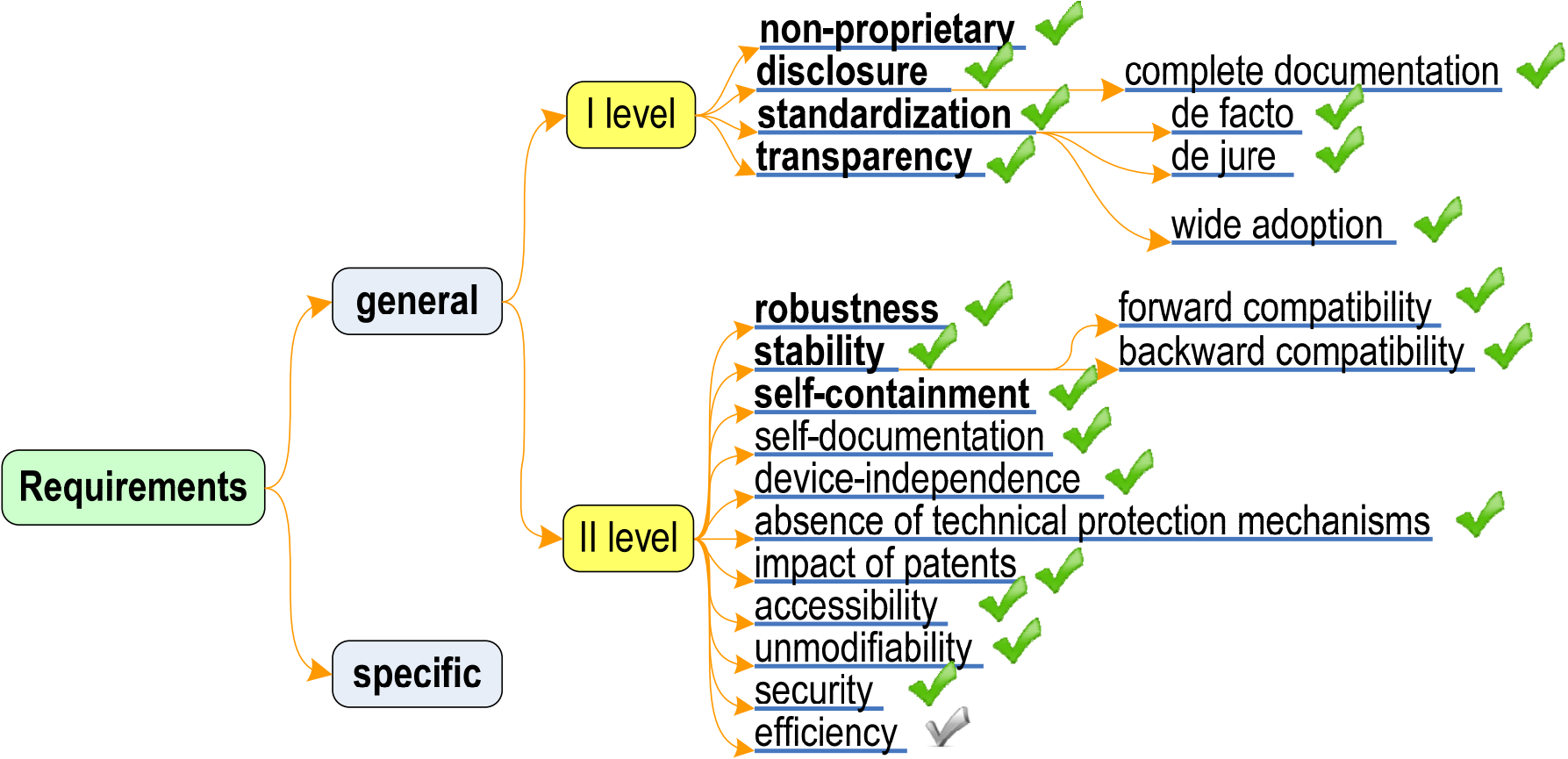
❺

Conclusions

Conclusions

Requirements
├─ general
│ ├─ I level
│ │ ├─ **non-proprietary** ✔
│ │ ├─ **disclosure** ✔ → complete documentation ✔
│ │ ├─ **standardization** ✔ → de facto ✔
│ │ └─ **transparency** ✔ → de jure ✔
│ │   → wide adoption ✔
│ └─ II level
│   ├─ **robustness** ✔ → forward compatibility ✔
│   ├─ **stability** ✔ → backward compatibility ✔
│   ├─ **self-containment** ✔
│   ├─ self-documentation ✔
│   ├─ device-independence ✔
│   ├─ absence of technical protection mechanisms ✔
│   ├─ impact of patents ✔
│   ├─ accessibility ✔
│   ├─ unmodifiability ✔
│   ├─ security ✔
│   └─ efficiency ✔
└─ specific

Conclusions

- FITS is a format that meets a large number of preservation requirements

- Other formats proposed for long term preservation of images do not always meet so many requirements.

- Example: the **TIFF** format, used for many years by both public and private companies to preserve every kind of documentation (invoices, correspondence, contracts, etc..), does not meet the same requirements as FITS format.

Conclusions

- Furthermore, FITS is one of the few formats that after more than 30 years since his birth has not became obsolete and is still widely used.

- So we can reasonably assert that FITS meets a large number of the requirements of a good format for long term preservation of images acquired by digitization projects.

- So before starting new digitization projects we must seriously begin to wonder if it is better to adopt this format instead of other formats typically adopted (such as TIFF, BMP, JPG, etc.).

A new ISO standard?

- Will the FITS format become the future standard format to be used in digitization projects?

- There is a proposal to the International Organization for Standardization (ISO) in order to obtain its recognition as a standard format for the preservation of digitized material.

- If FITS will became an ISO standard we will have an "official" format that can be used by libraries and archives to preserve digital assets obtained by digitization projects.

Conclusions

- Thank you for your attention!

stefano.allegrezza@uniud.it

stefano.allegrezza@unimc.it