Astronomical Data Formats Past, Present & Future

Jessica Mink
and the FITS Technical Group

Bob Mann
Astronomy and Computing

Special Issue of *Astronomy and Computing*

This BoF is organized in conjunction with a Special Issue of *Astronomy and Computing* on "The future of astronomical data formats", which is intended to provide a forum for peer-reviewed contributions to this debate.

http://www.journals.elsevier.com/astronomy-and-computing

Special Issue of *Astronomy and Computing*

Further papers are in preparation, and it is hoped that this BoF and subsequent discussion this week will generate more.

Final submission deadline of 1 March 2015

The special issue is being edited by Bob Mann. He and other *A&C* editors who are here this week - Tamas Budavari, Gerard Lemson, Wil O'Mullane, Andreas Wicenec -are happy to answer questions about the Special Issue or about *A&C* more generally.

Astronomical Data Formats

- Call for Action (5 minutes)
- Background (5 minutes)
- FITS (30 minutes)
- Other Formats (30 minutes)
- Discussion (40 minutes)

A Call For Action from Bob Hanisch

- FITS is now ~35 years old, an eternity in the IT world
- We are at risk of replicating the world of data format chaos that existed in the late 1970s
- Please don't criticize FITS for things that were not possible when it was designed
- The time for complaining is over; who will fill the roles of Harten, Wells, and Greisen?
- A possible way forward (c/o K. Shortridge): use VO-agreed data models as the high-level abstraction, use HDF5 as the Processing and Transport layer*
- Retain FITS as the (an) Archive* layer? Perhaps we don't have to solve all problems at once.
- Leave FITS otherwise alone so as not to distract from a more general solution
- Time scale: note reorganization of IAU Commission and Working Groups

*Terms to be defined shortly….

Astronomical Data Formats

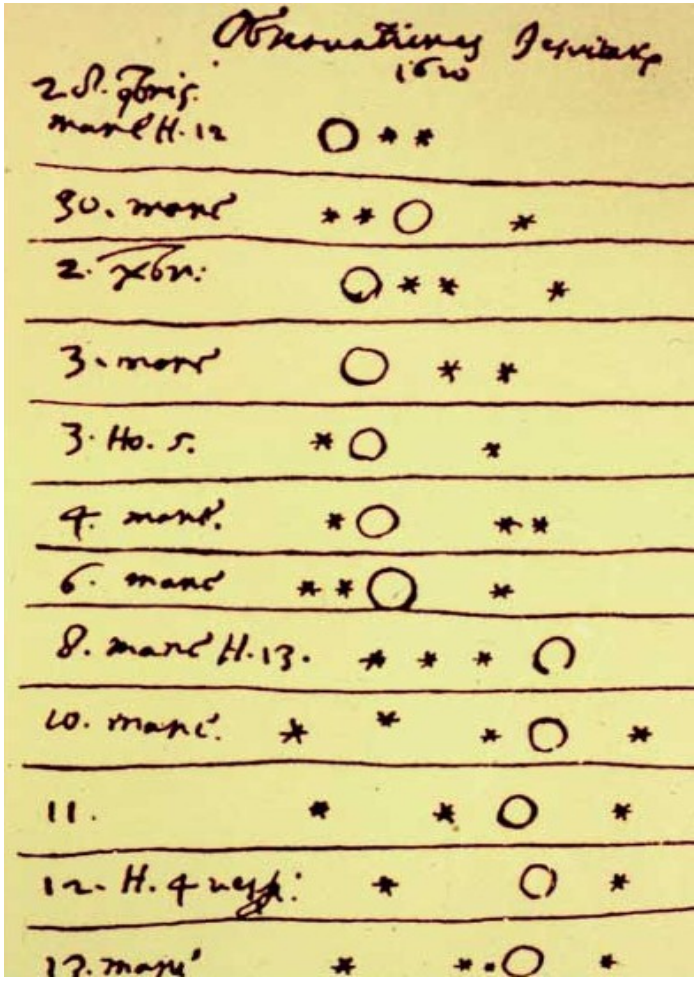
- **Recording**
  - Instrument-specific, Metadata recorded
- **Processing**
  - Software-specific, Metadata created
- **Transfer**
  - Well-documented, Metadata included
- **Archive**
  - Persistent, Metadata included

Where We Came From

Notebook

(Galileo)

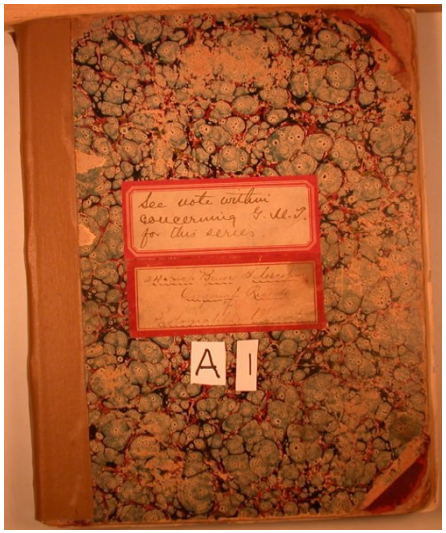
- Persistent,
  but not quantitative
- Metadata
  mixed with data

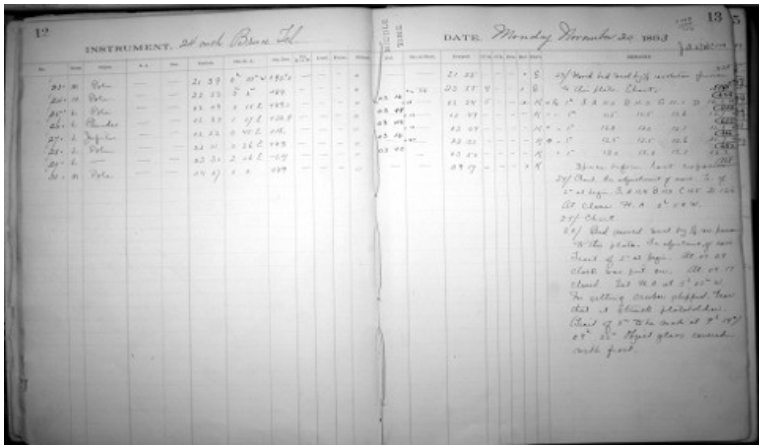


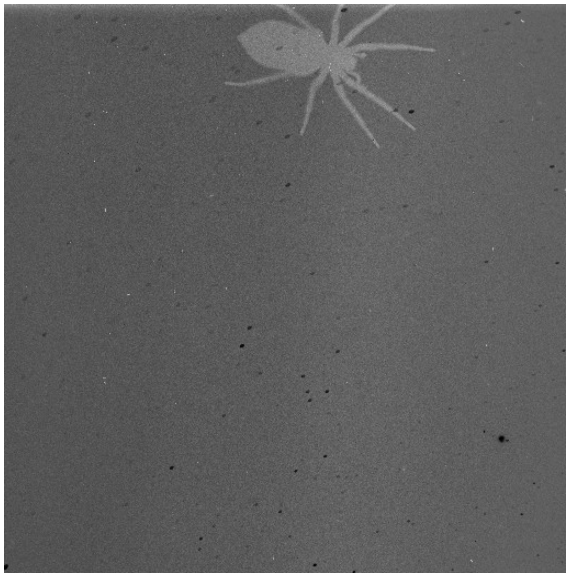
Where We Came From

Notebook with Analog Image

Persistence depends on media
Metadata may be more persistent than data





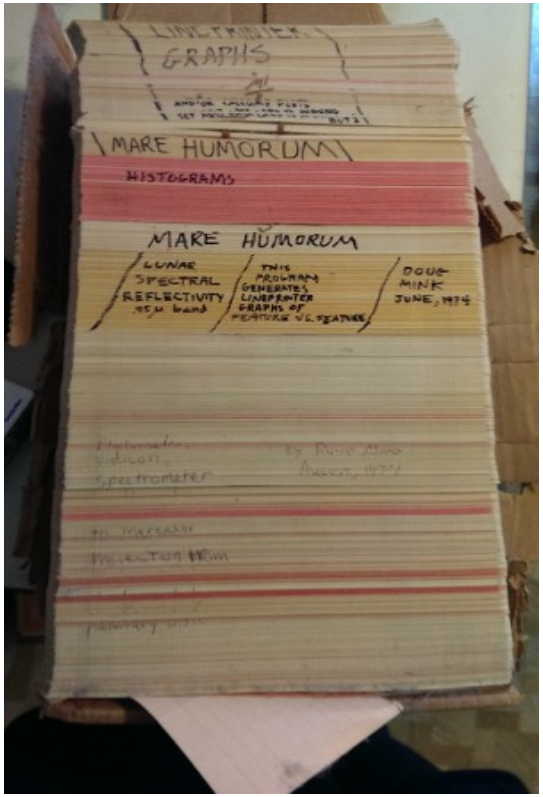


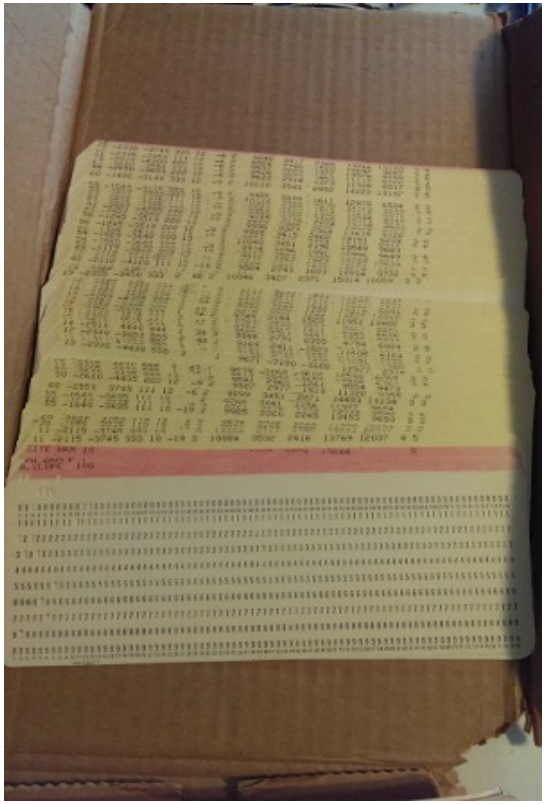
Harvard Plates, 1894

Where We Came From

Digital Hardcopy

Persistent, Metadata in software
MIT Lunar Spectroscopy 1973



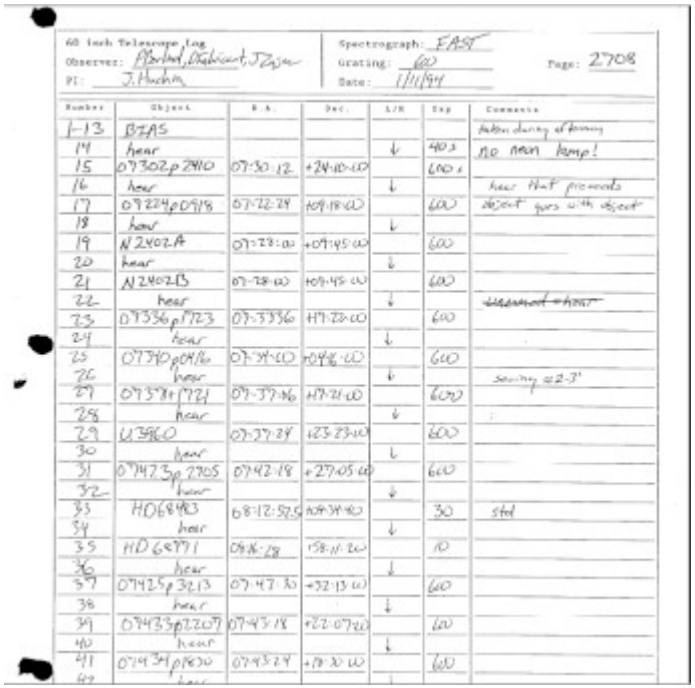


Where We Came From

Notebook with Non-standard Digital Image

Persistence depends on media

Metadata may be more persistent than data



60 Inch Telescope Log
Observer: Abbott, Dubiant, Zajac
PI: J. Huchra
Spectrograph: EAST
Grating: 600
Date: 1/11/94
Page: 2708

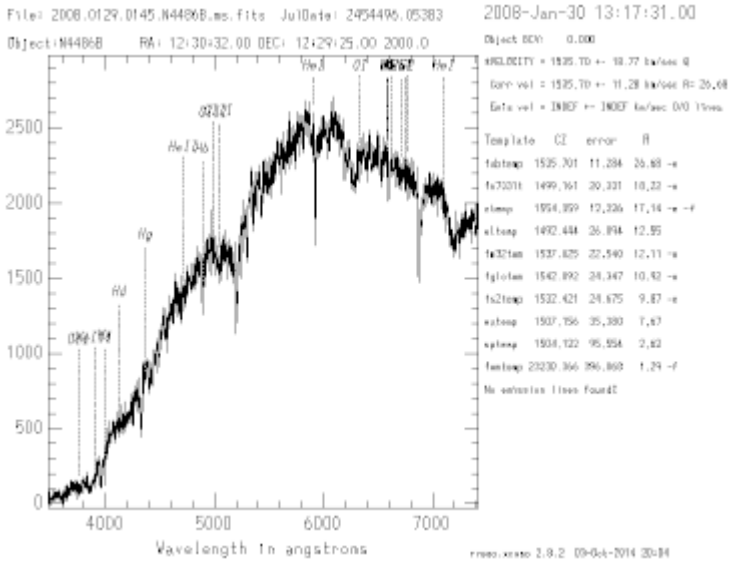
Number	Object	R.A.	Dec.	L/R	Exp	Comments
1-13	BIAS					Taken during afternoon
14	hear			↓	40 s	
15	07302p2410	07:30:12	+24:10:00		600 s	No neon lamp!
16	hear			↓		hear that proceeds
17	07224p0918	07:22:29	+09:18:00		600	object goes with object
18	hear			↓		
19	N2402A	07:28:00	+09:45:00		600	
20	hear			↓		
21	N2402B	07:28:00	+09:45:00		600	
22	hear			↓		~~second hear~~
23	07336p1723	07:33:56	+17:23:00		600	
24	hear			↓		
25	07340p0446	07:34:00	+04:46:00		600	
26	hear			↓		seeing ≈ 2-3"
27	07378p1721	07:37:46	+17:21:00		6020	
28	hear			↓		
29	U3960	07:37:28	+23:23:00		600	
30	hear			↓		
31	07473p2705	07:47:18	+27:05:00		600	
32	hear			↓		
33	HD68943	08:12:52.5	+09:34:40		30	std
34	hear			↓		
35	HD68991	08:12:28	+58:11:20		10	
36	hear			↓		
37	07425p3213	07:47:30	+32:13:00		600	
38	hear			↓		
39	07433p2207	07:43:18	+22:07:00		600	
40	hear			↓		
41	07434p1930	07:43:24	+19:30:00		600	
42	hear			↓		

Where We Are

Digital Standardized Image and metadata

Persistence depends on media
More metadata is available

```
SIMPLE  =                    T / Fits standard
BITPIX  =                  -32 / Bits per pixel
NAXIS   =                    3 / Number of axes
NAXIS1  =                 2681 / Axis length
NAXIS2  =                    1 / Axis length
NAXIS3  =                    4 / Axis length
EXTEND  =                    F / File may contain extensions
ORIGIN  = 'NOAO-IRAF FITS Image Kernel July 2003' / FITS file originator
IRAF-TLM= '10:26:23 (31/01/2008)' / Time of last modification
OBJECT  = 'N4486B  '           / OBJECT NAME
NAMPS   =                    1
DATE    = '2008-01-31T15:26:24' / UT date/time at start
DATE-OBS= '2008-01-30        ' / UT date/time at start
RA      = '12:30:32          ' / RA
DEC     = '+12:29:25         ' / DEC
EPOCH   = '2000              ' / EPOCH
ROTANGLE= '090               ' / ROT ANGLE
FASTFOC = '1280              ' / FAST FOCUS
AIRMASS = '1.20              ' / AIR MASS
RRA     = '12:30:32          ' / RRA
RDEC    = '+12:29:25         ' / RDEC
REPOCH  = '2000              ' / REPOCH
TRA     = '12:30:38.1        ' / TRA
TDEC    = '+12:30:03.7       ' / TDEC
TEPOCH  = '2000.000          ' / TEPOCH
ST      = '14:32:04          ' / SIDEREAL TIME START
HA      = '+02:00:59         ' / HOUR ANGLE
DOME    = '205               ' / DOME AZ
UT      = '13:16:01          ' / UT time at start
UTEND   = '13:19:01          ' / UT time at end
MJD     =        54495.553831 / MODIFIED JULIAN DATE MID TIME
GJDN    =       2454496.053831 / GEOCENTRIC JULIAN DATE MID TIME
HJDN    =       2454496.057228 / HELIOCENTRIC JULIAN DATE MID TIME
SITENAME= 'flwo1             '
SITELONG= '+110:52:39.0      ' / LONGITUDE, DEGREES WEST OF ZERO
SITELAT = '+31:40:51.4       ' / LATITUDE, DEGREES
SITEELEV= '2320.0            ' / ELEVATION, METERS
DETECTOR= 'FAST3             ' / DETECTOR ID
CCDSERIA= 'STA520A SN4377    ' / CHIP ID
INSTRUME= 'FAST              '
...
```



FITS
Past, Present & Future

FITS Technical Group
Lucio Chiappetti, *INAF*
Malcolm Currie, *JACH*
Jessica Mink, *SAO*
William Pence, *NASA*
Arnold Rots, *SAO*
Rob Seaman, *NOAO*

Why FITS

Needs:

- **Human and machine readability**
- **Self-documentating**
- **"Universally" readable format**
- **Presentation of standard in refereed papers**
- **Extensibility**

The History of FITS

June 1981: Publication of Standard

FITS - a Flexible Image Transport System
Wells, D. C.; Greisen, E. W.; Harten, R. H.
Astronomy and Astrophysics Supplement, Vol. 44, P. 363, 1981

A format for the interchange of astronomical images and other digital arrays on magnetic tape is described. This format provides a simple but powerful mechanism for the unambiguous transmission of n-dimensional, regularly spaced data arrays. It also provides a method for the transmission of a virtually unlimited number of auxiliary parameters that may be associated with the image. The parameters are written in a form which is easily interpreted by both humans and computers. The FITS format has been adopted for the transmission of astronomical image data by several large observatories including the Very Large Array, the Westerbork synthesis telescope, the Kitt Peak

Observatory and the Anglo-Australian Observatory

The History of FITS

Spreading use over time

Transferring →

Processing →

Recording →

Archiving

“Once FITS, always FITS.”

The History of FITS

Alternates

Re-use of format ideas by NOAO and STScI:

Processing formats with machine byte order and separate data and metadata for ease of processing which turned into Transfer and to some extent Archive formats.

The History of FITS

Metadata standards

WCS: Space, frequency/wavelength, time

Registry: Documentingn Site-specific keywords

The History of FITS

Versioning of formats

Definition of the Flexible Image Transport System (FITS), version 3.0
Pence, W. D.; Chiappetti, L.; Page, C. G.; Shaw, R. A.; Stobie, E.
Astronomy and Astrophysics, Volume 524, A42, 40 pp. (2010-12)

The Flexible Image Transport System (FITS) has been used by astronomers for over 30 years as a data interchange and archiving format; FITS files are now handled by a wide range of astronomical software packages. Since the FITS format definition document (the ``standard'') was last printed in this journal in 2001, several new features have been developed and standardized, notably support for 64-bit integers in images and tables, variable-length arrays in tables, and new world coordinate system conventions which provide a mapping from an element in a data array to a physical coordinate on the sky or within a spectrum. The FITS Working Group of the International Astronomical Union has therefore produced this new version 3.0 of the FITS standard, which is provided here in its entirety. In addition to describing the new features in FITS, numerous editorial changes were made to the previous version to clarify and reorganize many of the sections. Also included are some appendices which are not formally part of the standard. The FITS standard is likely to undergo further evolution, in which case the latest version may be found on the FITS Support Office Web site, which also provides many links to FITS-related resources.

The History of FITS

Process for approval of changes

IAU Standard

Δ

FITS Committee

Δ

Regional Committees (eliminated)

Δ

Proposed Standard

FITS WCS Time Paper Submitted

Representations of Time Coordinates in FITS
Rots, Arnold H.; Bunclark, Peter S.; Calabretta, Mark R.; Allen, Steven L.; Manchester, Richard N.; Thompson, William T.
arXiv:1409.7583 (2014-09)

In a series of three previous papers, formulation and specifics of the representation of World Coordinate Transformations in FITS data have been presented. This fourth paper deals with encoding time. Time on all scales and precisions known in astronomical datasets is to be described in an unambiguous, complete, and self-consistent manner. Employing the well--established World Coordinate System (WCS) framework, and maintaining compatibility with the FITS conventions that are currently in use to specify time, the standard is extended to describe rigorously the time coordinate. World coordinate functions are defined for temporal axes sampled linearly and as specified by a lookup table. The resulting standard is consistent with the existing FITS WCS standards and specifies a metadata set that achieves the aims enunciated above.

FITS Long-term Evolution

FITS Technical Group &

FITS Format for Document Preservation

Lucio Chiappetti, *INAF*

Malcolm Currie, *JACH*

Adam Dobrzycki, *ESO*

Jessica Mink, *SAO*

William Pence, *NASA*

Rob Seaman, *NOAO*

Estimate world FITS data holdings

- 15M FITS files @ NOAO over 20 years
  - 50M FITS HDUs *(mosaic cameras)*
- 100's of ground-based O/IR telescopes *(~ 200)*
  - NOAO ~ 10 telescopes
- 20 x 50M ~ **one billion FITS images**
  - Plus tallies for radio, space, etc.

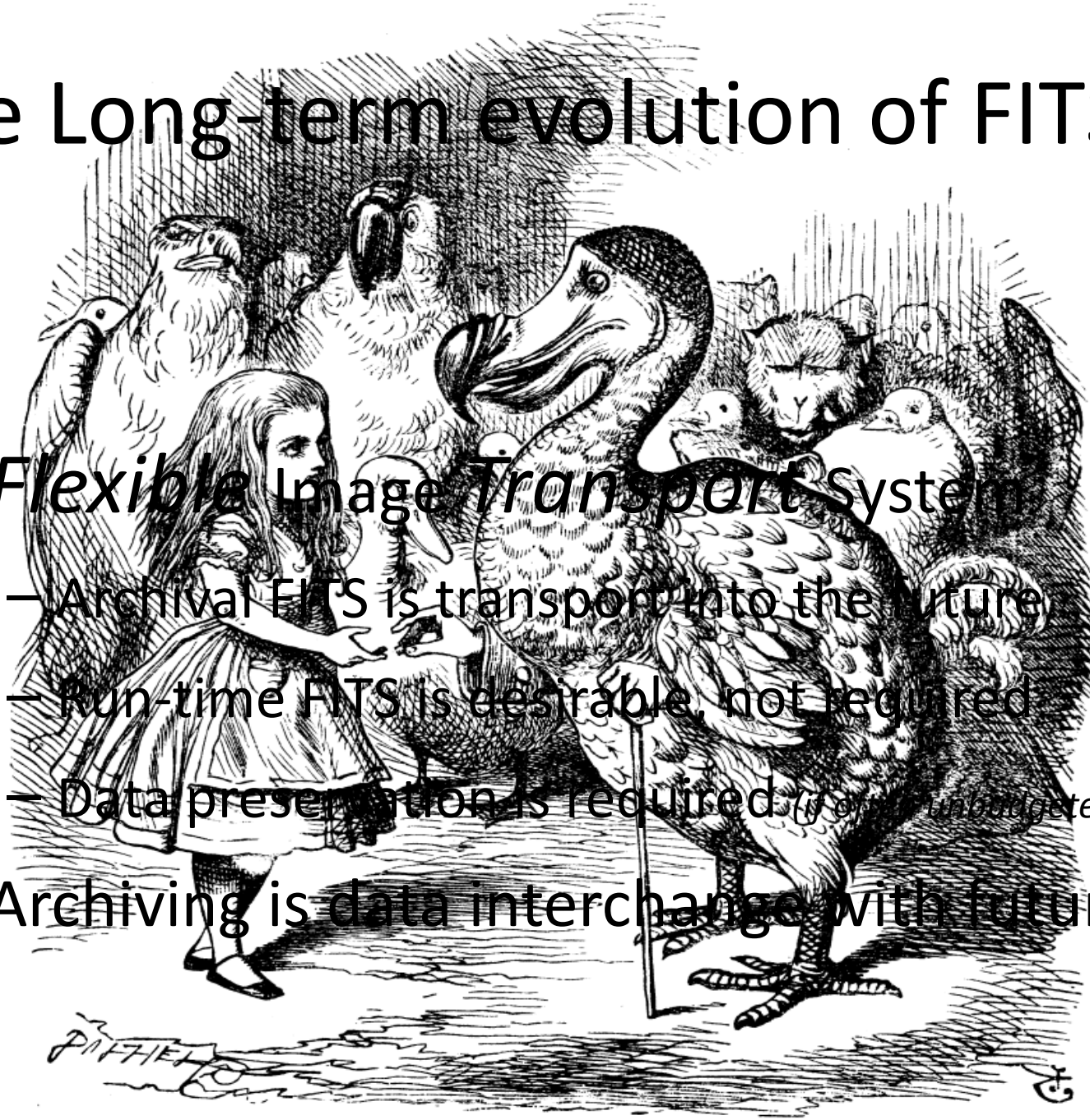
Conclusions

- FITS has been very successful
- FITS offers permanence
- Archival FITS holdings are extensive & growing
- Converting formats would be hugely expensive
    - *see poster P3-2, Data engineering for archive evolution*
- Support for FITS continues under all scenarios

Near-term Feature Enhancements

- Pence et al. paper for A&C special issue
  - Longer keyword names
  - Longer string-typed keyword values
  - Expanded character set for headers
- Others? (versioning, etc.)

The Long-term evolution of FITS



- *Flexible* Image *Transport* System
  - Archival FITS is transport into the future
  - Run-time FITS is desirable, not required
  - Data preservation is required *(if only unbudgeted)*
- Archiving is data interchange with future

FITS has a specific structure

- FITS is a sequence of Header-Data Units (HDUs)
  - an arbitrary number
- An HDU is a sequence of 2880-byte records
  - Header records first ($N_H >= 1$)
  - Data records last ($N_D >= 0$)
- Each Header record = 36 x 80-byte ASCII card images
  - Increment N_H until you reach the END card
- The number of Data records is encoded in the header
  - N_{pix} = NAXIS1 x NAXIS2 x … x NAXISn
  - $N_{bytes} = |BITPIX| \times (N_{pix} + PCOUNT) \times GCOUNT / 8$
  - $N_D = ceil (N_{bytes} / 2880)$

Everything is a table

- MEF files are now standard
  - Dataless primary HDU
  - IMAGE extensions
  - BINTABLE extensions
- Images are also BINTABLEs
  - FITS tile compression
  - Tile compression for tabular data

Tables for science metadata

- New FITS structure:
  - 1 primary header record (2880 bytes)
  - N binary tables containing data
  - 1 binary table containing metadata (or > 1)
- Assumes tile-compressed imaging data
  - efficient representation of data (perhaps lossy)
  - byte-order is a moot point
  - but also works with IMAGE extensions

Features

- Already legal FITS
- Task to convert header keywords to/from rows
  - Can include headers for backwards compatibility
- Metadata inheritance is explicit or could define separate one-to-one metadata tables
- Metadata at the end for efficient updating
- Can support proposed semantic features by adding columns to the table, modest elaboration
- FPACK supports data-type-aware compression

FITS is our Lingua Franca

lin·gua fran·ca

/ˈliNGgwə ˈfraNGkə/

noun: **lingua franca**; plural noun: **lingua francas**

a language that is adopted as a common language between speakers whose native languages are different.

- *historical*

  a mixture of Italian with French, Greek, Arabic, and Spanish, formerly used in the Levant.

Origin

late 17th century: from Italian, literally ‘Frankish tongue.’

FITS, Forever?

From more recent history:

In the early years of the IVOA there was a sense among some that "this time we would do it right and expeditiously - not like the old, slow FITS standards process." And there were those who optimistically felt that the IVOA would soon make FITS obsolete.

Well, the IVOA standards process, as it turns out, isn't going any faster than FITS did.

[Although it can be argued that the IVOA has developed more metadata standards faster than FITS did]

FITS

Question Time

Alternate Data Formats

Bob Mann, *Astronomy and Computing*

All papers from the Special Issue of *Astronomy and Computing* on "The future of astronomical data formats", are being posted on the A&C website in preprint form to enable inclusion in the debate. There are three so far:

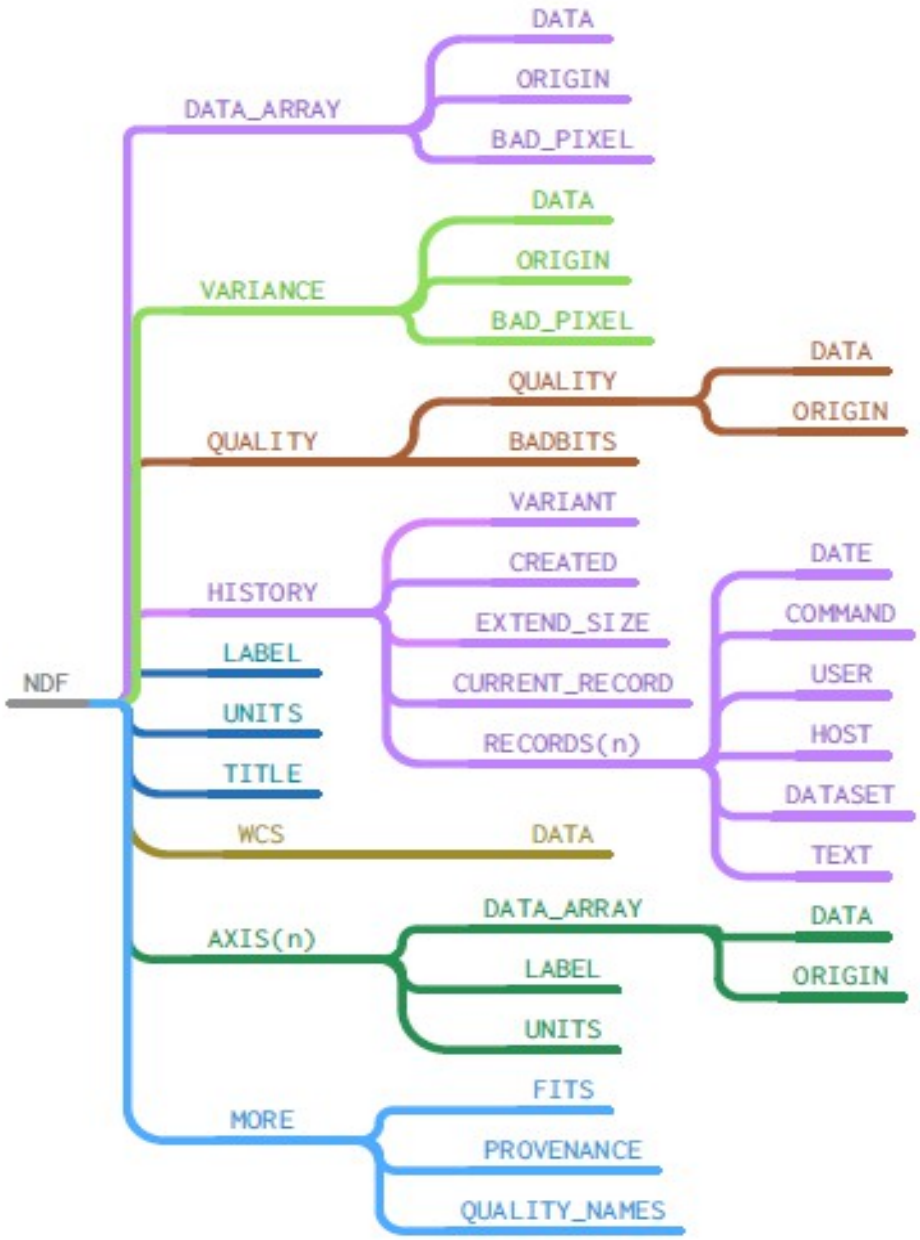
- **Tim Jenness et al - lessons learned from NDF**
- **Slava Kitaeff et al - use of JPEG2000 for astronomical imaging**
- **Brian Thomas et al - critique of FITS**

Moving NDF to HDF5

Tim Jenness, Cornell University

What is NDF?

- Extensible N-dimensional Data Format.
- Developed in 1988, still in use.
- Aimed to standardise on hierarchical data structure naming conventions for astronomy data.
- Hierarchical data model using the Starlink HDS file format on disk.
- Very successful approach with demonstrated ability to handle astronomy data from across the spectrum. Raw data still written using NDF. (See A&C paper in data formats special issue)
- No take up outside of Starlink and UK observatories.


NDF
DATA_ARRAY
DATA
ORIGIN
BAD_PIXEL
VARIANCE
DATA
ORIGIN
BAD_PIXEL
QUALITY
QUALITY
DATA
ORIGIN
BADBITS
HISTORY
VARIANT
CREATED
EXTEND_SIZE
CURRENT_RECORD
RECORDS(n)
DATE
COMMAND
USER
HOST
DATASET
TEXT
LABEL
UNITS
TITLE
WCS
DATA
AXIS(n)
DATA_ARRAY
DATA
ORIGIN
LABEL
UNITS
MORE
FITS
PROVENANCE
QUALITY_NAMES

What limits adoption?

- No one has heard of it
- NDF library written in Fortran
- What's the Hierarchical Data System? It's not FITS but is it something to do with HDF5?

- Maybe NDF should be given another chance?

HDS

- Developed in 1982 (six years before HDF)
- Now at version 4 (transitioned from BLISS to C and from VMS to unix large files)
- "Ahead of its time" for many years. Difficult to convince people that self-describing hierarchies were useful.
- No support for an efficient table format.
- No-one knows how HDS works any more. No-one paid to work it out.
- No hope of supporting arrays larger than 2 giga pixels.
- Should people use it?

HDF5

- Also hierarchical format
- Lots of users. Not a niche file format
- People trust it
- Lots of tools for reading data (h5py, pytables...)
- So, why not replace HDS with HDF5?

HDS in HDF5

- Reimplement the HDS C API in terms of HDF5
- No need to change any code that uses HDS
- NDF will "just work" but now be using HDF5 (does not solve Fortran NDF library problem, but one step at a time)
- Can then trivially extend NDF model to support tables and external links.

Does it work?

- Yes, but some compatibility wrinkles.
- HDS supports more datatype conversions than HDF5.
- HDS supports arrays of structures, HDF5 does not. Need additional layer in HDF5 hierarchy to emulate arrays.

```
/HDS_TEST/RECORDS        Group
/HDS_TEST/RECORDS/ARRAY_OF_STRUCTURES_CELL(1,1) Group
/HDS_TEST/RECORDS/ARRAY_OF_STRUCTURES_CELL(1,2) Group
/HDS_TEST/RECORDS/ARRAY_OF_STRUCTURES_CELL(2,1) Group
/HDS_TEST/RECORDS/ARRAY_OF_STRUCTURES_CELL(2,2) Group
```

- HDF5 has no concept of a group type but does support generalized attributes.
- HDS cells, slicing and vectorization of datasets can be done with hyper slabs and point lists. “Dataspaces” in HDF5 are very powerful.

Memory mapping?

- HDS supports direct memory-mapping of datasets using mmap(), except where datatype conversion is required.
- HDF5 does not seem to support mmap access at all in public API (because of filters? type conversion? different driver behavior?). Am I wrong?
- HDF5 seems to prefer you read in small chunks of data a bit at a time.

Tracing an HDF5 file

h5ls -d -r

```
/                        Group
/HDSTRACE                Group
/HDSTRACE/ADAM_DYNDEF    Group
/HDSTRACE/EACHLINE       Dataset {SCALAR}
    Data:
        (0) 0x01
/HDSTRACE/FULL           Dataset {SCALAR}
    Data:
        (0) 0x00
/HDSTRACE/NEWLINE        Dataset {SCALAR}
    Data:
        (0) 0x00
/HDSTRACE/NLINES         Dataset {SCALAR}
    Data:
        (0) "1" ' ' repeats 130 times
/HDSTRACE/OBJECT         Group
/HDSTRACE/OBJECT/NAMEPTR Dataset {SCALAR}
    Data:
        (0) "hds.h5sdf" ' ' repeats 100 times
/HDSTRACE/TYPIND         Dataset {SCALAR}
    Data:
        (0) 15
/HDSTRACE/VALIND         Dataset {SCALAR}
    Data:
        (0) 16
/HDSTRACE/WIDEPAGE       Dataset {SCALAR}
    Data:
        (0) 0x00
```

hdstrace

```
HDSTRACE  <STRUC>

   ADAM_DYNDEF    <DEFAULTS>      {structure}
      {structure is empty}

   EACHLINE       <_LOGICAL>      TRUE
   FULL           <_LOGICAL>      FALSE
   NEWLINE        <_LOGICAL>      FALSE
   NLINES         <_CHAR*132>     '1'
   OBJECT         <ADAM_PARNAME>   {structure}
      NAMEPTR        <_CHAR*132>     'hds.h5sdf'

   TYPIND         <_INTEGER>      15
   VALIND         <_INTEGER>      16
   WIDEPAGE       <_LOGICAL>      FALSE

End of Trace.
```

To Do

- Final 10% of HDS API
- Wrapper interface library that transparently selects classic HDS vs HDF5/HDS based on type of file being opened.
- Routine to walk through a tree of nodes in classic HDS file and migrate them to HDF5/HDS file.
- Conveniently, the new format would be HDSv5.
- Performance testing

Using FITS to understand astronomical data format needs

Brian Thomas
NOAO

A&C Submitted Paper

This work based on the A&C Paper (Thomas etal. 2014?)

- 37 co-authors (broad experience, POV)
- Examined primarily the limitations
- Sort of exercise can be used to extract some requirements for astronomical data formats

Draft Paper available at:

http://tinyurl.com/acfits-draft-pdf

Lets do a live poll!

- **Conclusions first!**

FITS is a useful, but aging standard

- **The question is:** What should we do about this?

- **Try a live poll** to get a sense of community mood.

  Please go answer a few questions at:

http://tinyurl.com/adfquiz

Why examine FITS?

- **FITS is a great “test particle”** for analyzing astronomical data format needs.
  - Lingua Franca of astronomical data
  - Lots of data are in FITS

    A very successful format which is widely used
- **It has many technical strengths**
  - Well documented/adopted/tested
  - Models (Image, Table, 3D data cubes, WCS)
  - Tile compression
  - Good software support
  - “Archiveabilty”

http://tinyurl.com/adfquiz

Our process

- **Form a broad group and collect issues (use cases)** on astrodataformat google group.
  - Invite a wide variety of people to participate.
  - Set and enforce ground rules to lower chance of acrimony and focus discussion.
- **Discuss and Distill down**
  - **Identify root causes** in accepted use cases.
  - **Find and group** common causes
  - **Extract** “Lessons Learned”
- **Take Time**
  - First post to submitted paper: 1.5+ yr

http://tinyurl.com/adfquiz

Important FITS limitations

- **Well known**
  - Metadata expression (8 char keywords, 68 char values, hierarchical structures not 'native', no built in associations)
  - Data model issues (WCS, associations)
  - Serialization (choice of endian, missing values)
- **New needs have surfaced**
  - The internet! Greater sharing of data, increased validation and machine understanding needed
  - Large data support
  - Virtualization/distributed support
  - More and improved data models

http://tinyurl.com/adfquiz

Lessons Learned

- One of FITS greatest strengths is sociological

A shared format is huge boon

- There are important **lessons to be learned from the limitations** however
  - Format needs to be self-describing (version and schema) to support expanded modern data interchange and archiving
  - Format must be able to express many complex metadata and associations
  - Conventions are not standards

http://tinyurl.com/adfquiz

Summary

- **Single standard for sharing data is a HUGE boon for astronomical community BUT *FITS is showing its age***.
  - IF we want to continue having this kind of shared standard, then **FITS needs to evolve sufficiently** or a new standard needs to be found.
  - **But How?** Should we choose to evolve through existing standard/conventions or apply radical surgery? New dataformat which translates useful datamodels of FITS? Perhaps start completely from scratch?

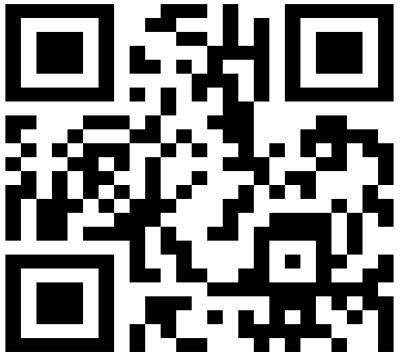
- **What happens next?**
  - Suggestion: because its a community standard, **we need to engage the community** to find a path forward
  - **Gather use cases/"lessons learned"** which also show FITS strengths
  - Glean **use cases/"lessons learned" from other data formats**
  - **Determine what is important for the future**
    - ex. is this to be just a "transport" format, or should it include "archiving". What models are important to share? Should be part of the standard, or not?

http://tinyurl.com/adfquiz

Your thoughts? Poll Results

Results at:

http://tinyurl.com/adfresults



Alternate Data Formats

Question Time

Astronomical Data Formats

Discussion