Astronomy and Computing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

Astronomical data formats: What we have and how we got here

Jessica D. Mink

Smithsonian Astrophysical Observatory, 60 Garden St., Cambridge, MA 02131, United States

ARTICLE INFO

Article history:
Received 24 April 2015
Received in revised form
29 June 2015
Accepted 1 July 2015
Available online xxxx

Keywords:
Data format
FITS
Distortion
WCS

ABSTRACT

Despite almost all being acquired as photons, astronomical data from different instruments and at different stages in its life may exist in different formats to serve different purposes. Beyond the data itself, descriptive information is associated with it as metadata, either included in the data format or in a larger multi-format data structure. Those formats may be used for the acquisition, processing, exchange, and archiving of data. It has been useful to use similar formats, or even a single standard to ease interaction with data in its various stages using familiar tools. Knowledge of the evolution and advantages of present standards is useful before we discuss the future of how astronomical data is formatted. The evolution of the use of world coordinates in FITS is presented as an example.

Published by Elsevier B.V.

1. Where we are

The astronomical community's most widespread data format, FITS (Wells et al., 1981) is 35 years old, and interest in developing a new and improved standards for formatting the larger and more varied types of astronomical data being produced by more and more complicated instruments on larger and larger telescopes is spreading (Thomas et al., 2014) and (Thomas et al., in press). Several existing options are being proposed: HDF5, a Hierarchical Data System (Jenness, 2015), and JPEG2000, a widely-used image format (Kitaeff et al., in press), among others. The problems we face are not all new, and I would like to cover some history about how we got where we are and how our present solutions developed.

2. Formatting data through its life-cycle

2.1. Genesis: origins of astronomical data

In the beginning, there is light. Most astronomical observations are of photons. In addition to a count or other measure of their intensity, we record information including direction of the source, wavelength, polarization, or frequency or energy of individual photons or some grouping thereof, and the time(s) at which they were collected. Associated metadata describing the conditions under which the data was created may be included with the data as a header, trailer, or internal labels, or may reside in a separate format, such as a logbook, digital table, or label on the data container.

At its simplest, a data format includes data structured in some way to make it retrievable. It may be a qualitative drawing in a logbook, such as Galileo's drawing of Jupiter and its largest satellites, with descriptive information about the data written right next to it. It may be a table of numbers in a published paper or monograph, with the text of the paper providing the contextual metadata and the headings on the table labeling the actual numbers.

In the nineteenth century, it became possible to record a signal from photons from the sky more directly on glass photographic plates, such as those in the Harvard Plate Collection (Grindlay et al., 2009). It is made up of photographic plates containing images of the sky, with metadata as notes in logbooks and on the paper jackets in which the plates are stored. Metadata for each plate includes the pointing direction, the time and exposure of the observation, the name of the object being observed, and who observed it. The logbook and jacket indicate what telescope was used and where it was located (See Fig. 1). Sky coverage comes from the telescope focal plane plate scale and the physical size of the plate. Not all useful parameters were (or needed to be) written out because the humans using the data knew them, so additional work has been needed to make the plate images scientifically usable (Tang et al., 2013).

Digital data acquisition formats are usually limited to one instrument or a class of instruments. Metadata is usually recorded digitally either in the same files or files associated in some way within a data structure so that processing software can learn as much as possible from its digital input. For spacecraft observations, this would be digital telemetry; for optical telescopes, this is likely to be some sort of image files. Both might have associated input files, such as pointing catalogs or fiber positions.

E-mail address: jmink@cfa.harvard.edu.

http://dx.doi.org/10.1016/j.ascom.2015.07.001
2213-1337/Published by Elsevier B.V.

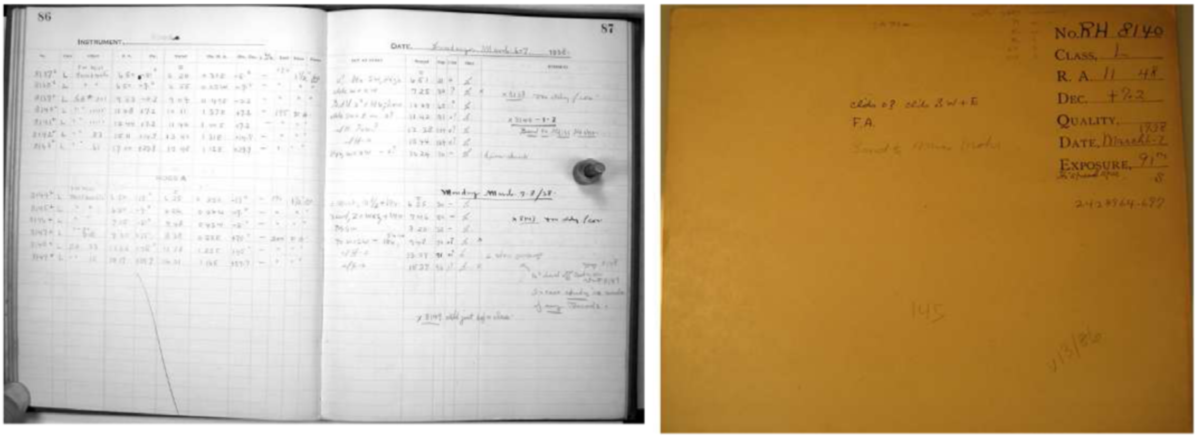

Fig. 1. Harvard plate metadata in logbook and on plate jacket.

2.2. Transfer and exchange

As we process that data, we end up with derived data which can take many different forms. To exchange that data, process it with standard software, and archive it so scientists can read it in the near or far future, the metadata has to be standardized and well-documented. It is helpful if that metadata travels easily with the actual bits of data.

2.3. Processing and analysis

But the data we have acquired is not ready to be used for science. As David Hogg noted (Bard and Hogg, 2013), "The data is like a noisy hash of the things we care about". It has to be processed from raw data, including observations and calibrations, to something that can be analyzed. Before we can look at the data, we want it formatted in a specific way which makes it accessible to the tools we wish to use.

In the course of processing, metadata as well as data is changed, with information about the processing or results of the analysis being added. It is useful if the metadata both utilizes standard definitions and is clearly associated with the data as part of the same file or data structure.

2.4. Archiving data

Final disposition of data can take several paths. It can be destroyed because it is seen to have no value or there is no space to keep it. It can be stored on media which become unreadable, such as no-longer-readable or slowly-decaying tape or disk formats. It can be presented in a publication, copies of which are preserved in multiple places, though only that part of the data content relevant to the publication will be preserved. And finally, data may be preserved in a format which is both persistent in content and readability.

Most ground-based and much space-based data from the 1970's through the 1990's was recorded on magnetic tapes or disks which are now next to impossible to read. More recent data on CDROM's and DVD's will be lost as the media degrade and readers become less common. While tapes I made during the 1970's and 1980's are no longer readable, Hollerith cards of photometry and software from my senior year in college are still human- and scanner-readable (see Fig. 2).

If they can, scientists share their data, analysis of it, and conclusions from it in presentations and papers. In the age of hardcopy journals and books, these methods were more persistent than any other method of preservation, but 21st century astronomers tend to find articles through ADS (Kurtz et al., 2000) or the ArXiv server (Vence, 2014) and read them online instead of in hard copy. As journals move toward online publication, standardized, retrievable formatting for the long term is becoming an issue here, too.

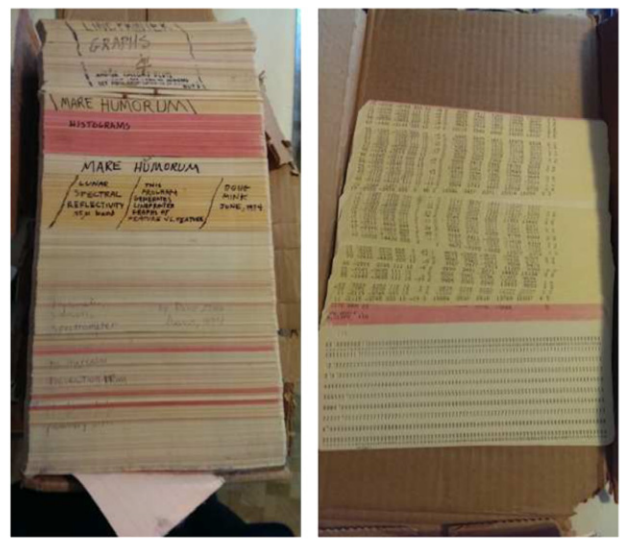

Fig. 2. The format for the photometry on these cards from 1973 is documented by the software, also in the card deck, which reads them.

We now have the storage capacity to save most of the data that we take, as well as its derivatives. So far the most permanent format we are using is printed paper, with tables and graphs containing data. Photographic plates, which degrade a bit faster over time, with metadata in separate paper logs, have lasted over a century. But now most of our data and more and more of the papers and presentations which describe its meaning are digital. We need both persistent media and persistent formats to keep today's data accessible over time.

3. Standardizing data formats

As the inventors of FITS noted in their first paper (Greisen et al., 1980):

> Under the traditional system for data interchange in astronomy, each institution exports data on magnetic tape in its own unique internal format. Thus, a group of N "cooperating" institutions would begin by creating N(N−1) format translation programs. Then, whenever one of the institutions changes its internal format, the other N−1 institutions have to change their corresponding translation programs. For obvious reasons, this traditional system has been very inefficient. It would be very

Please cite this article in press as: Mink, J.D., Astronomical data formats: What we have and how we got here. Astronomy and Computing (2015), http://dx.doi.org/10.1016/j.ascom.2015.07.001