IN53B-1630

This poster may be downloaded and commented on at : `http://eposters.agu.org/`

Review of Provenance Metadata in Solar Physics Data Archives

Joseph A. Hourclé
NASA/GSFC (Wyle)

Virtual Solar Observatory

http://virtualsolar.org/

contact info : `http://docs.virtualsolar.org/wiki/ContactUs`

We present the results of a review of solar physics data archives to evaluate the presence of preservation metadata including:

- *Archive-assigned identifiers*
- *Data checksums*
- *Provenance information*
- *Links to usage documentation*
- *Policies on data use*

Rationale

Although the FITS (Flexibile Image Transport System) allows for very robust descriptions of the data being stored within the files, many archives are not using the standard to its full ability.

Correct data use requires understanding what information is contained within the file, which means knowing when the observation was taken, where the instrument was pointed, and what physical values the data represents.

Verifying the file may also require knowing what processes have been done to derive the final values, including the specific software used, their version, information about the OS, hardware and computing environment.

Improved documentation within the files may require effort up front, but can reduce the effort needed to support usage of the data in the future. Poorly documented data can be a bottleneck preventing wider use of the data and the funding institutions from getting the maximum return on their investment and increases the risk of unusable data over the long term.

Methodology

We analyzed the data from the perspective of someone who might have found the files through the Virtual Solar Observatory (VSO) or some other search engine, and may not have or be aware of the SolarSoft library, which is how most solar physics PI teams distribute software for analyzing their data. We also considered a future researcher or archivist attempting to determine what the files contained without any other supporting documentation.

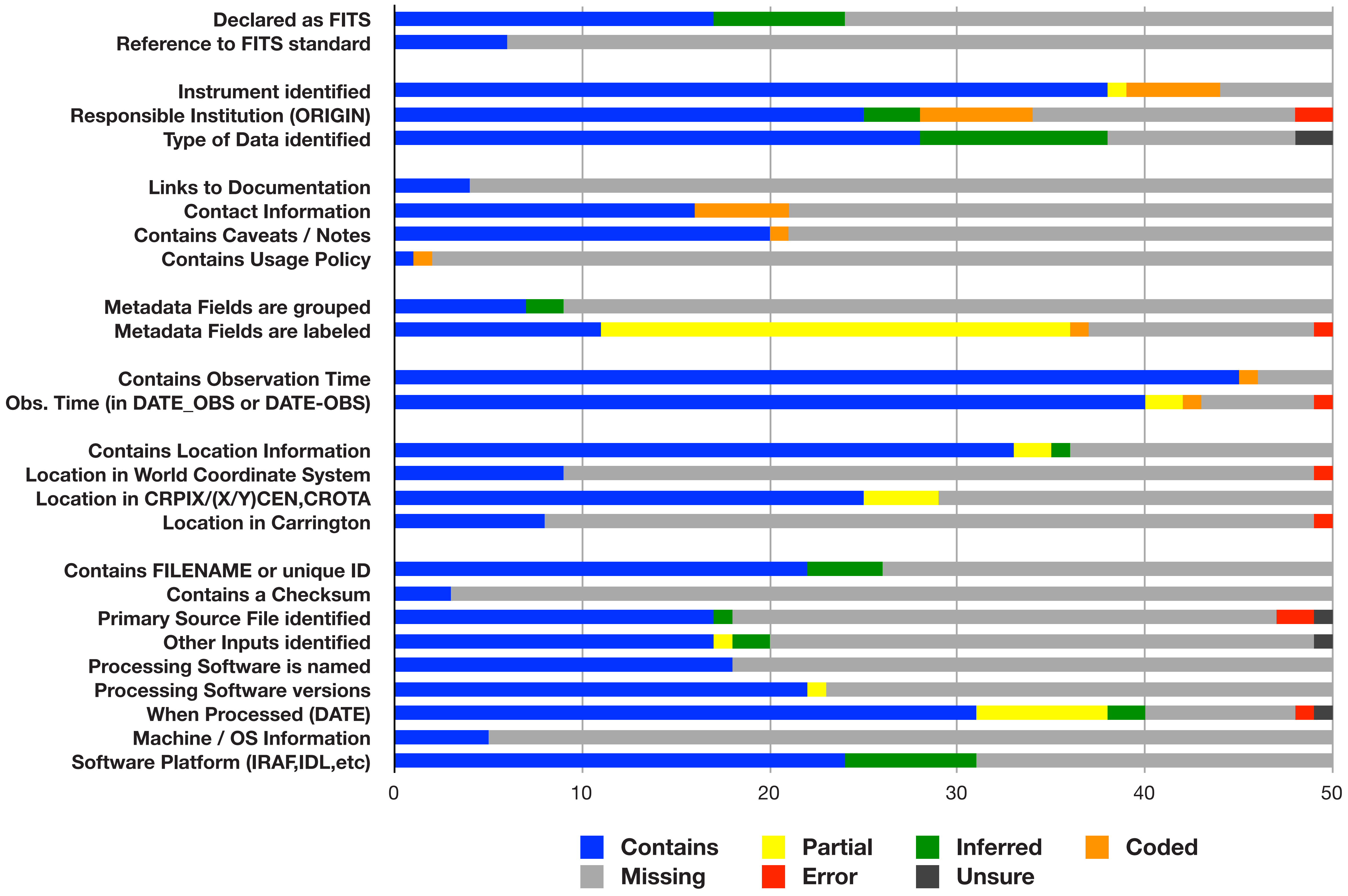
We started with a sample of FITS (Flexible Image Transport System) files from 12 different instruments from active and recent missions, both served from the VSO and those on the list of priorities to be added, and made a list of general categories of provenance and administrative metadata, and if there were de facto standards for each type of content.

We then analyzed the full set of files, making notes on how well each file seemed to cover the categories identified:

- Contains the content
- Contents were incomplete or partially compliant
- Not explicit, but might be possible to infer the values
- Content is coded; requires external documentation or knowledge
- Does not contain the content
- The content has a suspected error
  - (truncated strings, obviously incorrect dates, empty fields)
- Unclear if it contains the content
  - (has non-standard fields that may contain the content)

Analysis was done using a simple text editor and knowledge that FITS headers are organized around 80 character cards, and for some files, the '`gunzip`' utility. No external documentation was consulted in the analysis. Headers and sub-headers were considered up until the values no longer matched the 80-card values; this may mean that compressed sub-headers were missed.

As the solar community has standardized on FITS, we only analyzed FITS files. Other solar data exists in non-standard mission specific formats, ASCII tables, or NetCDF, but they were not analyzed at this time.



Results

As some of the analysis was subject to interpretation, with datasets that we were more familiar having an unfair advantage, we have aggregated the results. It's also possible that the single sample was not representative of the entire collection for that instrument. This also means we don't need 9 point type to fit it all.

We have included a couple of excellent cases; although they were good, no single file that we found included all of the metadata that we were looking for.

Notes:

'Coded' was used for abbreviations without the full name that were less well known; the determining test was a Google search for the acronym. It was also used for notes in language other than English, and when only a person's first name was given.

'Partial' was used for incomplete `DATE` fields; they conform to an older standard, and precision for time of processing is likely not a significant factor for most of the files analyzed. In some cases, the time was in the comment for `SIMPLE`.

'Partial' was also used for ambiguous (non-unique) instrument names.

Not all instruments were imagers and so some methods of specifying location were not appropriate,; in addition, WCS and FITS standards for checksums did not exist or were not in wide use when some of the instruments were commissioned.

Analysis

There was a wide range of levels of compliance. Although some institutions had great data, the level of documentation still varied within the institution, even when we look only at active efforts, and ignore data that may have been documented only as part of the preparation for final archiving.

In a few cases, there was regression; newer missions were less well documented than previous missions or browse products better documented than the 'science quality' data. Although this could be explained by active vs. final archives, to the best of our knowledge, the older data had not yet had its final archive generated.

In one case, the lower level data available for an instrument was described well, but the higher level data lacked basic metadata such as time and instrument in the FITS header (although this was encoded into the file's name). As this was non-image data, it's possible that the data was maintained in some other form such as NetCDF and that the conversion utility did not transfer the metadata. If the community wishes to standardize on FITS, we need to identify and fix these problems.

Recommendations

A deeper analysis needs to be formed to verify that metadata headers are being used consistently; for example, the '`ORIGIN`' field for some files seemed to have information about the software rather than the institution running the software. If necessary, unambiguous fields should be defined.

All files intended for archiving should have a declaration of their file format as early as possible; some explicitly state they are FITS and a reference publication, while others have no indication that they are anything other than a key-value list.

Some files store the same concepts multiple ways; although this is a major benefit to the user as they don't need to apply coordinate transformations themselves, it may be worth developing a way to denote which set of values are considered authoritative in case there is disagreement.

Future Plans

We will be composing a more detailed list of recommendations, which will be posted with the other checklists, for the SOHO mission as they generate their final archive:

`http://docs.virtualsolar.org/wiki/Checklists`

We hope to work with the Solar Probe Plus and Solar Orbiter missions to ensure well documented files are produced.

As the identifying information varied so widely, it may be useful to write a tool that would attempt to identify a file; later revisions may be able to annotate the headers, insert values from external sources, or calculate missing headers.

We believe it's also worthwhile to create external documentation standards, or at a minimum a template of types of information that is useful to scientists attempting to use the files. If possible, we would like to create a documentation repository to ensure that this information is preserved and easily discoverable.