sc | nat

Mathematics, Astronomy and Physics
Platform of the Swiss Academy of Sciences

Swiss Committee on Space Research (CSR)

How far dare you go?

SCNAT/MAP Meeting on Open Data

29 October 2018

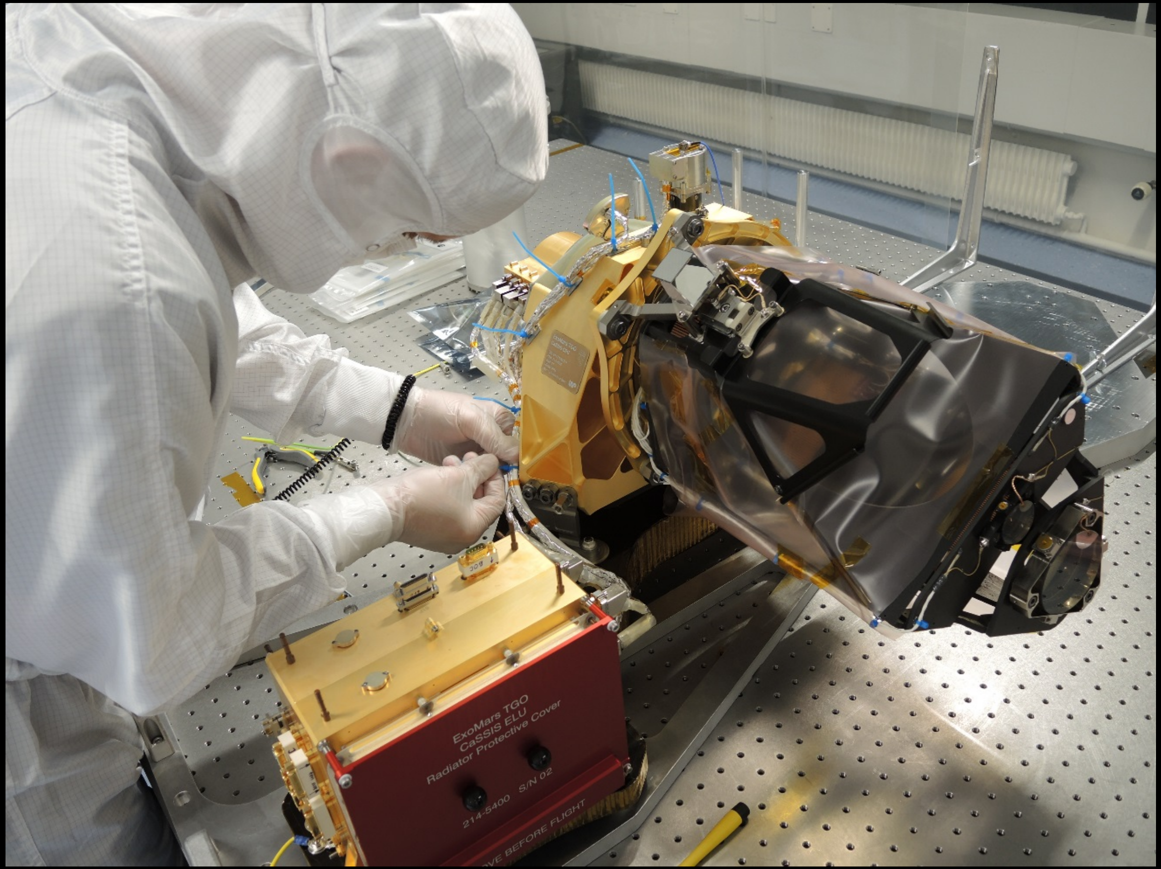
Nicolas Thomas, University of Bern

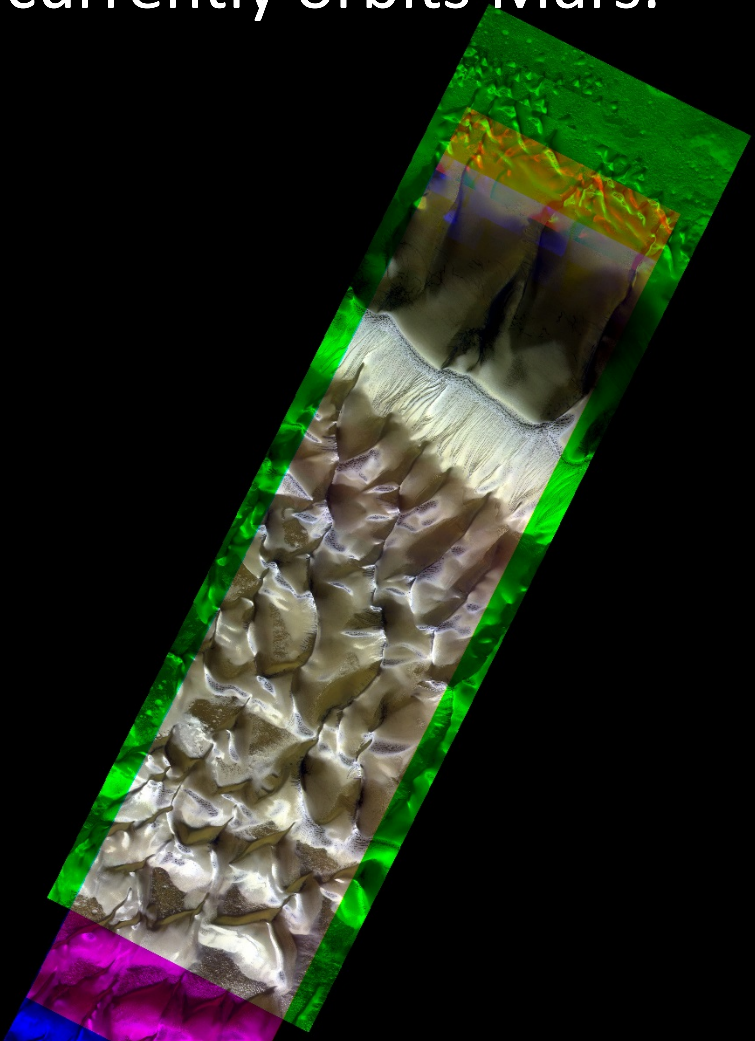
Space and open data

- Space research has long been *confronted* with “open data”.
  - NASA’s Planetary Data System has been in existence for >35 years.
  - It is mandatory to archive all data products (raw and calibrated) in PDS format and NASA missions must archive in the PDS.
    - The European Space Agency has its own Planetary Science Archive (PSA) (for more than 15 years) that performs a similar function and there is strong coordination by using PDS formats.
- Giotto (1986 fly-by of comet Halley)
  - 1st European planetary mission to have full open data
    - This was through NASA’s PDS Small Bodies node.

CaSSIS – a current example

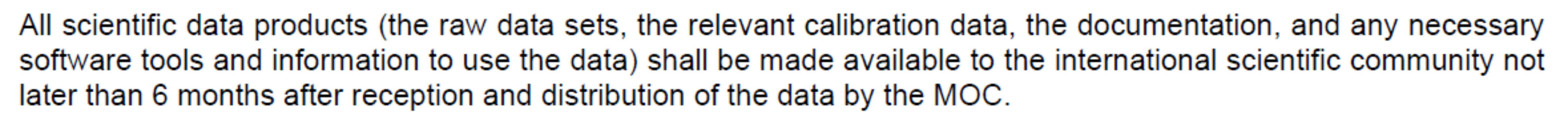
- CaSSIS is the Uni Bern built camera that currently orbits Mars.
- I will use this as an example



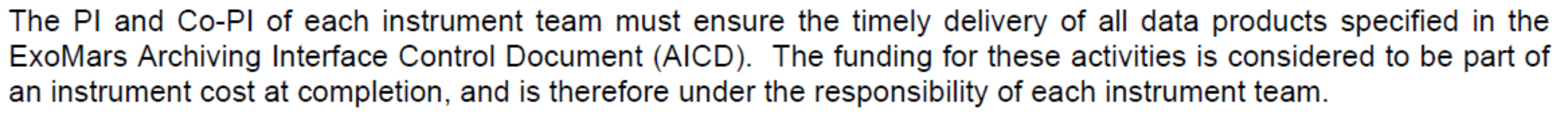


Requirements

The ExoMars Science Management Plan (EXM-MS-PL-ESA-00002; Iss. 5, Rev 4) states



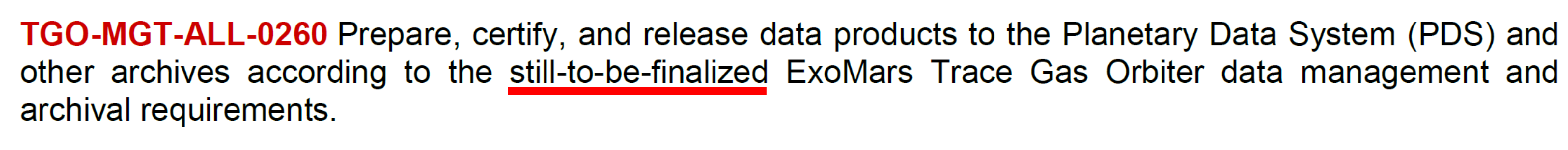
All scientific data products (the raw data sets, the relevant calibration data, the documentation, and any necessary software tools and information to use the data) shall be made available to the international scientific community not later than 6 months after reception and distribution of the data by the MOC.



The PI and Co-PI of each instrument team must ensure the timely delivery of all data products specified in the ExoMars Archiving Interface Control Document (AICD). The funding for these activities is considered to be part of an instrument cost at completion, and is therefore under the responsibility of each instrument team.

Experiment to Archive Interface Control Document (EAICD)

The Instrument Requirements Document (EXM-IRD-ESA-00003; Iss. 2; 31 Jan 2012) states

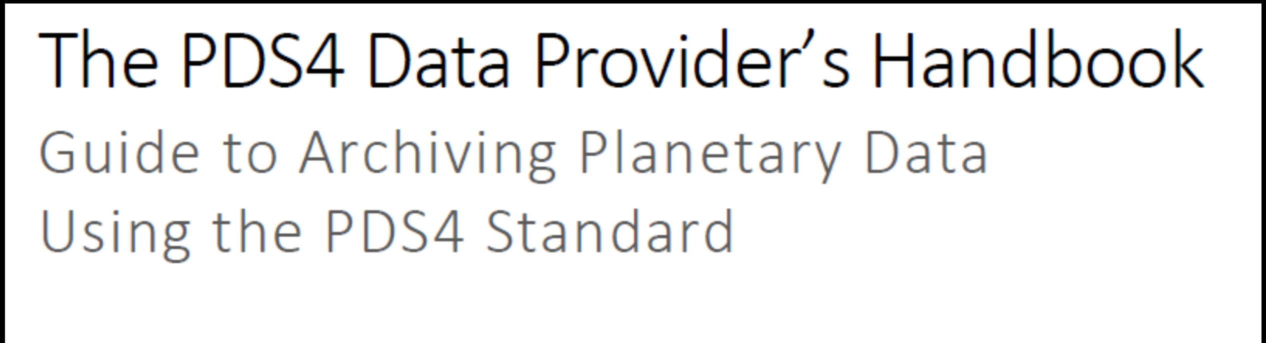


TGO-MGT-ALL-0260 Prepare, certify, and release data products to the Planetary Data System (PDS) and other archives according to the still-to-be-finalized ExoMars Trace Gas Orbiter data management and archival requirements.

I had to sign against this 3 years before launch!!

How is PDS (V4) defined?

- The Planetary Data System has a well defined format.
- But this format has just gone from V3 to V4 – a complete change.
  - V4 is an XML file describing each data record in detail plus the data record itself and it is NOT backward compatible.

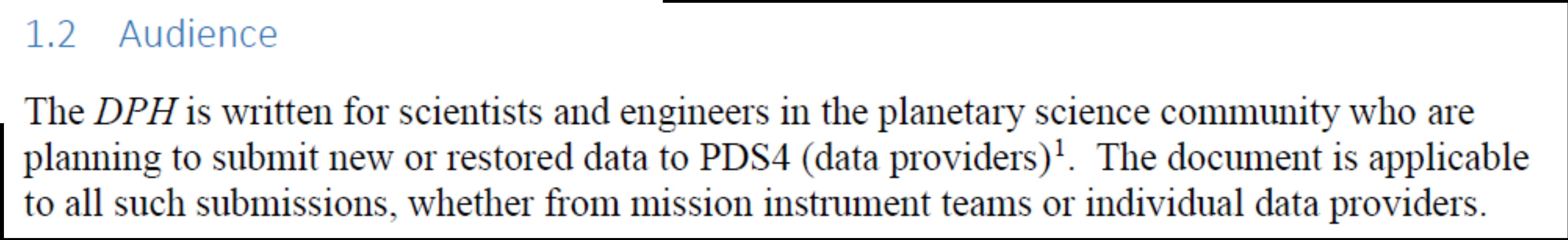


The PDS4 Data Provider's Handbook

Guide to Archiving Planetary Data
Using the PDS4 Standard

Version 1.10.1
May 10, 2018

This document is a mere 129 pages.

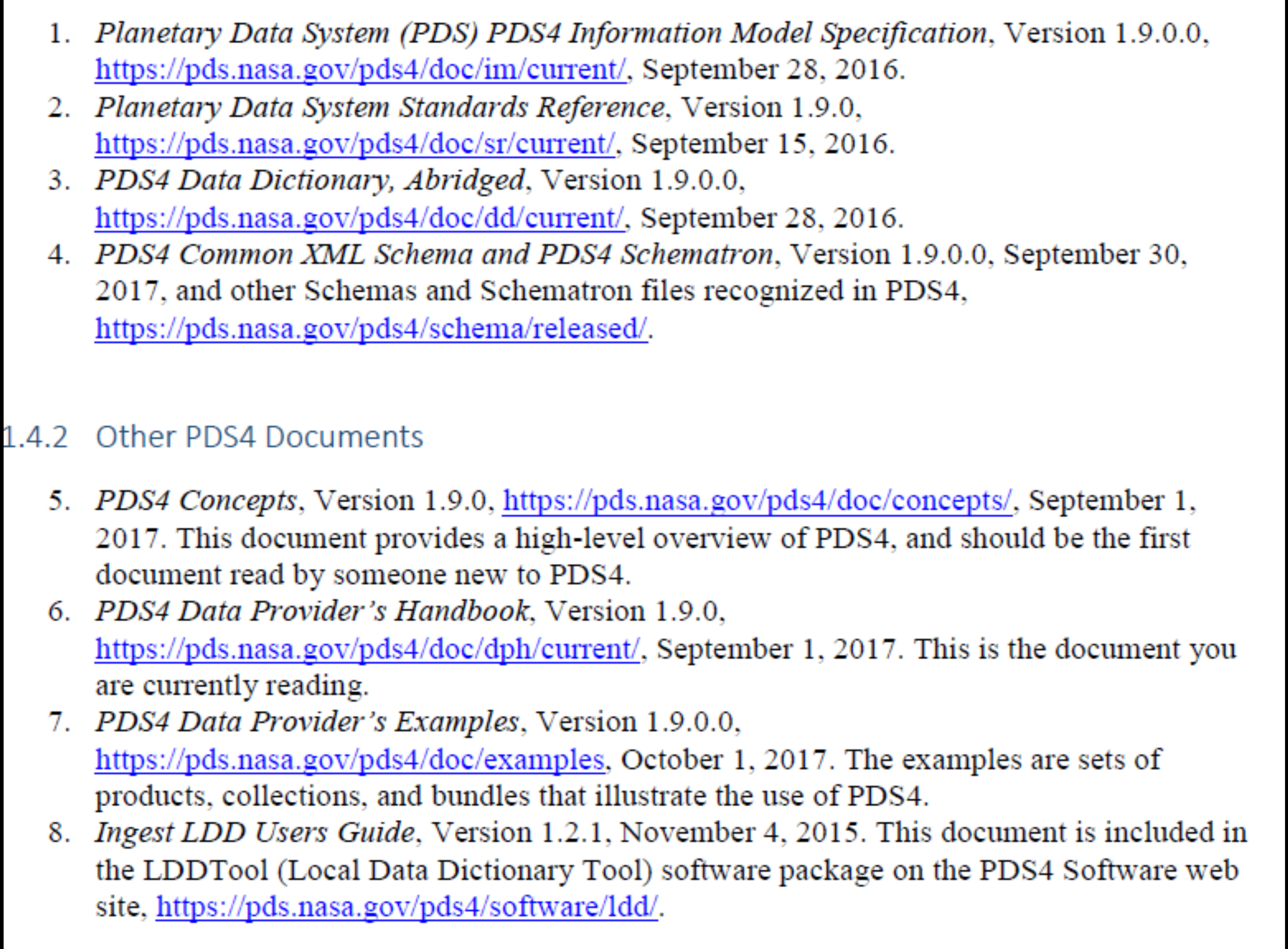


1.2 Audience

The *DPH* is written for scientists and engineers in the planetary science community who are planning to submit new or restored data to PDS4 (data providers)[1]. The document is applicable to all such submissions, whether from mission instrument teams or individual data providers.

PDS V4

The data dictionary is a trivial 644 pages.

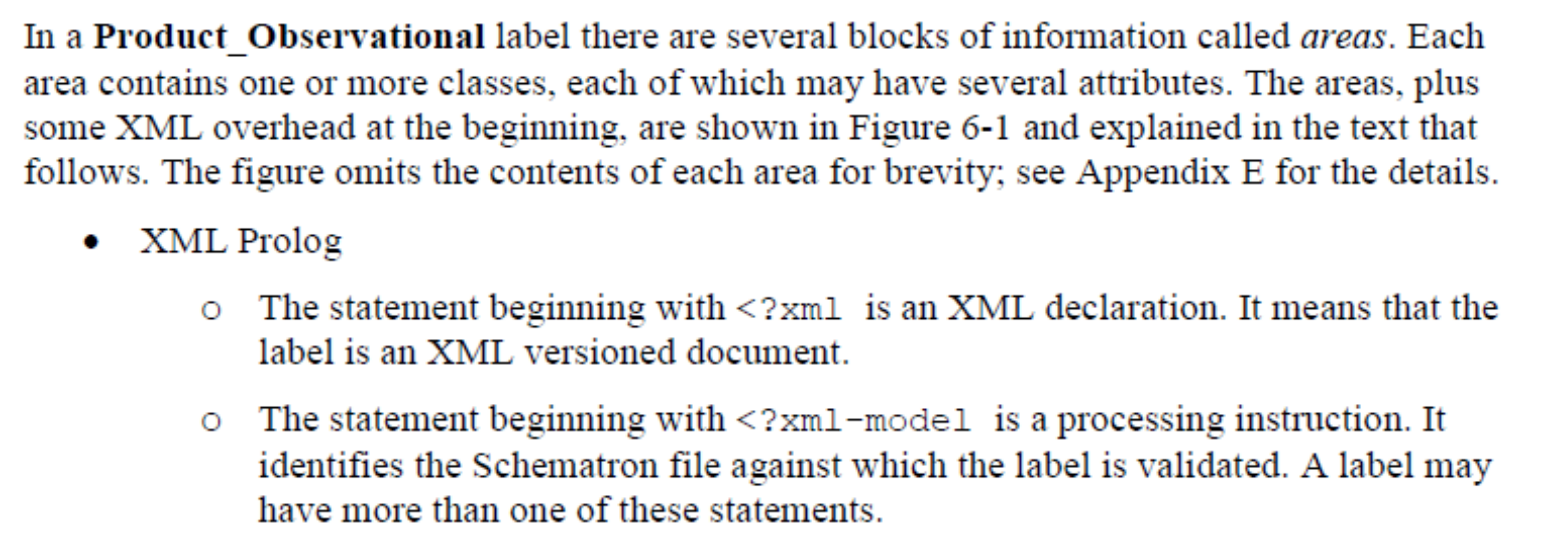


1. *Planetary Data System (PDS) PDS4 Information Model Specification*, Version 1.9.0.0, https://pds.nasa.gov/pds4/doc/im/current/, September 28, 2016.
2. *Planetary Data System Standards Reference*, Version 1.9.0, https://pds.nasa.gov/pds4/doc/sr/current/, September 15, 2016.
3. *PDS4 Data Dictionary, Abridged*, Version 1.9.0.0, https://pds.nasa.gov/pds4/doc/dd/current/, September 28, 2016.
4. *PDS4 Common XML Schema and PDS4 Schematron*, Version 1.9.0.0, September 30, 2017, and other Schemas and Schematron files recognized in PDS4, https://pds.nasa.gov/pds4/schema/released/.

1.4.2 Other PDS4 Documents

5. *PDS4 Concepts*, Version 1.9.0, https://pds.nasa.gov/pds4/doc/concepts/, September 1, 2017. This document provides a high-level overview of PDS4, and should be the first document read by someone new to PDS4.
6. *PDS4 Data Provider's Handbook*, Version 1.9.0, https://pds.nasa.gov/pds4/doc/dph/current/, September 1, 2017. This is the document you are currently reading.
7. *PDS4 Data Provider's Examples*, Version 1.9.0.0, https://pds.nasa.gov/pds4/doc/examples, October 1, 2017. The examples are sets of products, collections, and bundles that illustrate the use of PDS4.
8. *Ingest LDD Users Guide*, Version 1.2.1, November 4, 2015. This document is included in the LDDTool (Local Data Dictionary Tool) software package on the PDS4 Software web site, https://pds.nasa.gov/pds4/software/ldd/.

… and the documents are fun to read



In a **Product_Observational** label there are several blocks of information called *areas*. Each area contains one or more classes, each of which may have several attributes. The areas, plus some XML overhead at the beginning, are shown in Figure 6-1 and explained in the text that follows. The figure omits the contents of each area for brevity; see Appendix E for the details.

- XML Prolog
  - The statement beginning with `<?xml` is an XML declaration. It means that the label is an XML versioned document.
  - The statement beginning with `<?xml-model` is a processing instruction. It identifies the Schematron file against which the label is validated. A label may have more than one of these statements.

Data and verification

- Technical validation
  - You can of course write codes to convert your data from any format (e.g. as it comes out of the instrument) into PDS V4.
  - But it has to pass a validation check written by PDS/PSA to prove it is a valid file.
  - The XML headers must also be documented including a “data dictionary” supplying the meaning of all the XML attributes.
- Scientific validation
  - Further there is a review of submitted data by scientists in the field.
    - If the review board finds errors or cant actually use the data then the supplier is given actions to make changes. (I did this job on Giotto.)

Why do all this?

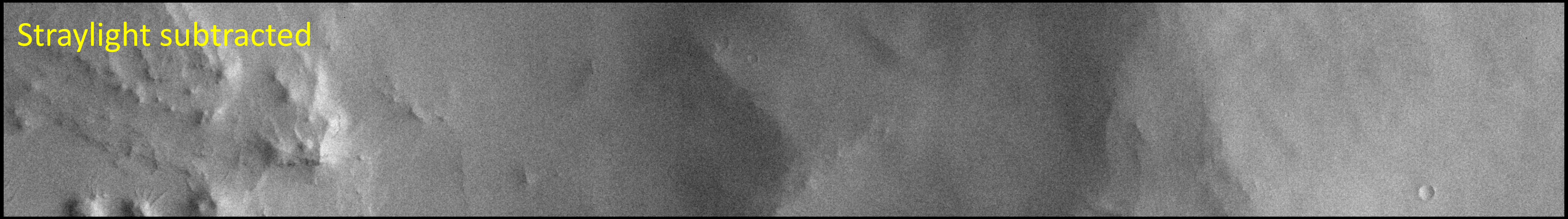
- It is pointless archiving data that cannot be used by anyone but the producers.
  - This is called a “data mortuary” and is the most likely result of undocumented data archiving in the absence of any verification and/or control.
  - And be aware that data mortuaries can occur inside your own environment.
    - I wanted to look at some old data from the Phobos 2 spacecraft that I had last used 10-15 years ago and couldn’t remember how to read it!

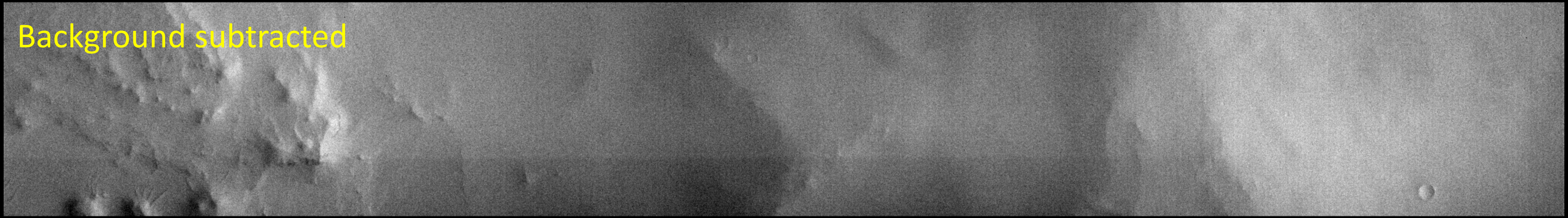
But it is not enough…

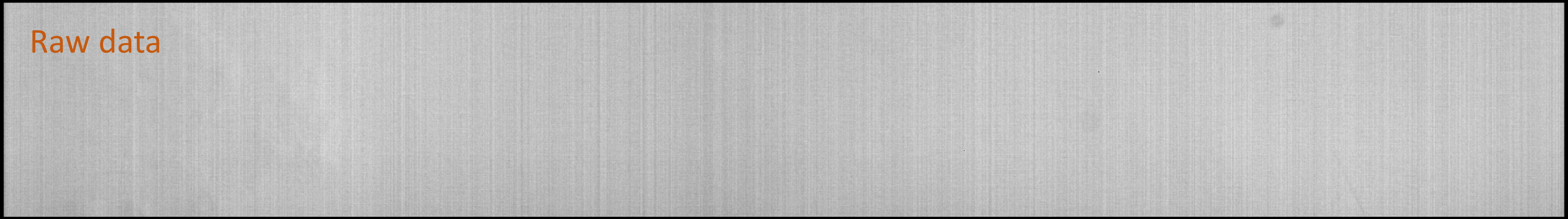
Calibration

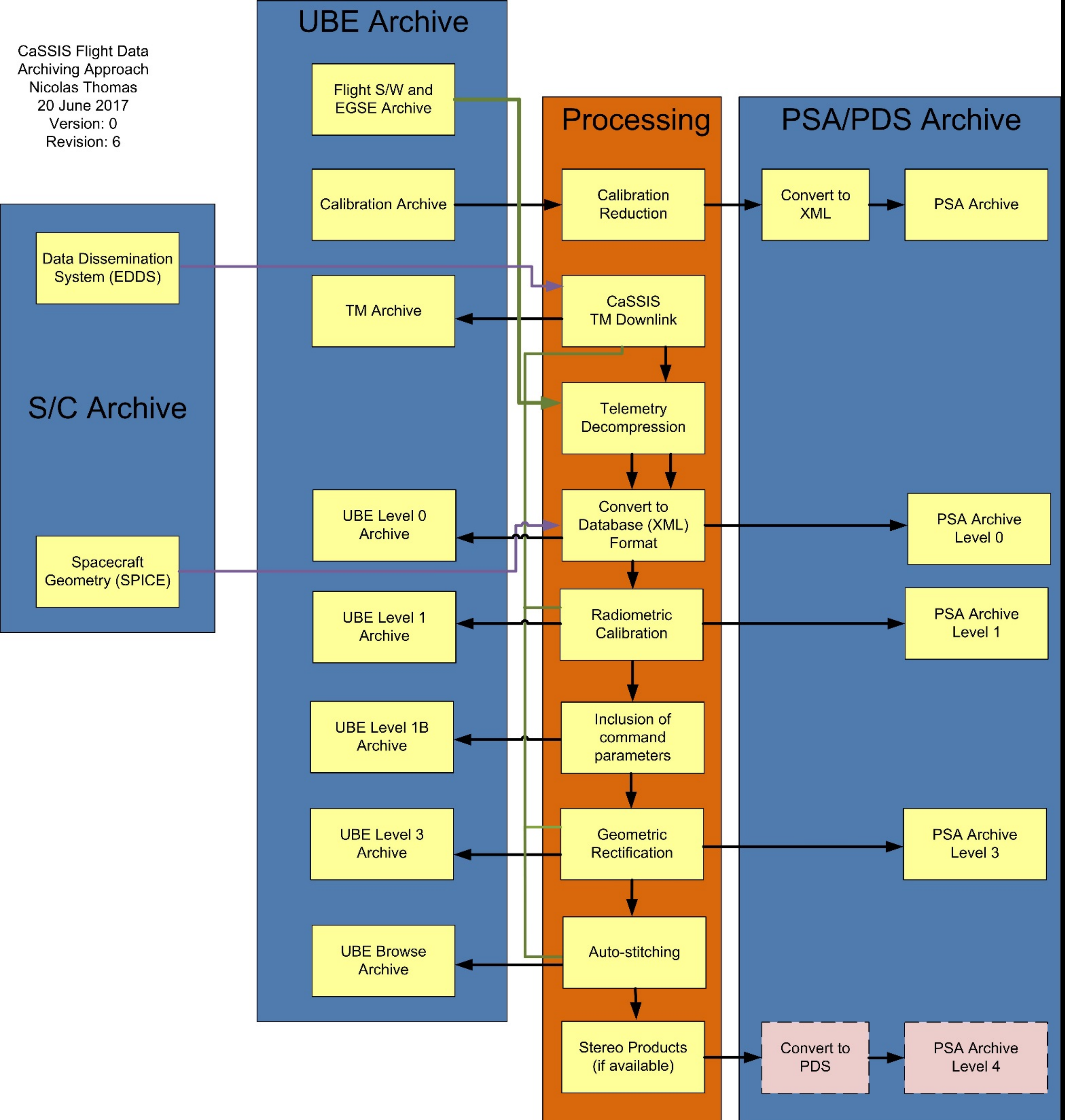
- Many instruments require calibration.
  - Removing the instrumental/systematic effects of the data acquisition.

CaSSIS Radiometry









CaSSIS Flight Data
Archiving Approach
Nicolas Thomas
20 June 2017
Version: 0
Revision: 6

S/C Archive

Data Dissemination System (EDDS)

Spacecraft Geometry (SPICE)

UBE Archive

Flight S/W and EGSE Archive

Calibration Archive

TM Archive

UBE Level 0 Archive

UBE Level 1 Archive

UBE Level 1B Archive

UBE Level 3 Archive

UBE Browse Archive

Processing

Calibration Reduction

CaSSIS TM Downlink

Telemetry Decompression

Convert to Database (XML) Format

Radiometric Calibration

Inclusion of command parameters

Geometric Rectification

Auto-stitching

Stereo Products (if available)

PSA/PDS Archive

Convert to XML

PSA Archive

PSA Archive Level 0

PSA Archive Level 1

PSA Archive Level 3

Convert to PDS

PSA Archive Level 4

Mathematics, Astronomy and Physics
Platform of the Swiss Academy of Sciences

Swiss Committee on Space Research (CSR)

This is a (simplified) flow chart of CaSSIS processing and storage.

Calibration

- Many instruments require calibration.
  - Removing the instrumental/systematic effects of the data acquisition.
- The producer can calibrate the data for the community.

Calibration

- Many instruments require calibration.
  - Removing the instrumental/systematic effects of the data acquisition.
- The producer can calibrate the data for the community.
- But PDS/PSA tries to insist that the calibration software is also delivered so that the community can re-calibrate the data if improved techniques or calibration files are established.
  - This can be a huge additional drain on resources.
  - CaSSIS calibration files were generated from data in a completely different format to the flight data and with the instrument and equipment in a completely different configuration. That means a completely new data dictionary is required, etc. etc. etc.
- On the other hand, provision/archiving of software to read and manipulate the data is helpful to the community.
  - Uni Geneva’s high energy astronomy group is of the opinion that without this, you still get a data mortuary because the community cant really access the data.

Software

- PDS tried to insist that software is delivered in an open source format using open languages (e.g. FORTRAN and C).
  - I.e. languages that don’t cost money like IDL or MATLAB.
- For us, this would have been catastrophic. High level languages are vital to save resources and time.
  - (Don’t talk to me about Python.)

- Fortunately the organizations recognized the error of their ways here.
- But this too needs specifying in some form of requirements document.

Archiving costs

- The producer calibrates anyway but there are additional costs
  - Development of the PDS V4 compatible format
  - Development of the software to accompany the format
  - Validation of the format and its local storage
  - Interfacing with the PDS/PSA (meetings, reviews, etc.)
- CaSSIS is working 24/7
  - I have a downlink engineer paid for by the Bund supervising the calibration, ensuring stuff doesn’t break and pushing the data into archives. He will also get the data dictionary on his plate. (Final delivery in January.)
- Roughly 15% of the current spending on CaSSIS is archive and “open data” related (ca. 1 FTE).

And then there is support….

- How much support should the data producer give the community to use their data/software and who pays for it?
- And should data producers referee papers submitted on the basis of their data (e.g. to assess whether it has been used correctly)?

My personal take …

- Archiving data for community use is highly valuable if it is done properly.
    - For CaSSIS I try to follow the example of a US experiment that has generated >1000 scientific papers through being open. But that experiment had (and still has) more staff than I do.
- Doing it properly
    - costs manpower
        - … and I would STRONGLY object to this being underfunded such that PhD students end up doing it.
    - requires appropriate standards
    - requires a verification system against a contractual obligation.
        - I believe such a contractual obligation requires negotiation …. another story.
    - and requires some form of curation organization for long-term preservation
- Not doing it properly
    - results in production of “data mortuaries” and is a total waste of money.
    - includes archiving the numbers used to make plots in papers!

But it could be expensive.....

- Yes, but …

- I would rather see the problem properly specified. If you then cant pay for it, then descope TRANSPARENTLY.
- Do NOT underfund and then slowly increase requirements with time without increasing funding. That is a recipe for exploitation of junior staff.

So…… ask yourself…

- Why am I preserving the data set?
- To what extent do I want it preserved?
- Am I prepared to pay for it?