# ESA's Planetary Science Archive (PSA): Ensuring The Long-term Usability Of Data

**David J. Heather [1], M. Barthelemy [2], N. Manaud [2], S. Martinez [2], H. Metselaar [2], M. Szumlas [2], J. Kissi-Ameyaw [1], and the PSA Development Team [2]**

[1] ***ESA-ESTEC***
*Keplerlaan 1, 2200AG, Noordwijk, The Netherlands*
*EMail: dheather@rssd.esa.int*

[2] ***ESA-ESAC***
*PoBox 78, 28691 Villanueva de la Cañada, Madrid, Spain*

## ABSTRACT

The European Space Agency's Planetary Science Archive (PSA) has the long-term preservation of data and knowledge from all of ESA's planetary missions as a core focus. This paper will discuss the various data handling and validation procedures put in place to ensure the long-term usability, integrity and compatibility of our archives with as wide a range of data from other planetary archiving authorities as is possible. PSA maintain and distribute a validation tool (PVV) allowing for syntactic validation of data and volumes against the Planetary Data System (PDS) archiving standards. These standards are widely accepted within the planetary science community, and have therefore been adopted for all PSA archives to ensure cross-mission compatibility. A local dictionary of keywords is maintained by PSA and used to validate all data prior to ingestion, and also to ensure that all the information that is required to understand and query relevant parameters for an instrument are present in all data delivered. In addition, scientific validation tools are now being used to ensure that the quality of measurements is flagged correctly in data product labels. It is hoped that the processes put in place will maximise the long-term usefulness of data and knowledge from all of ESA's planetary missions.

Keywords: planetary, PSA, ESA, data, archive, preservation, long-term, validation

## INTRODUCTION

The European Space Agency's Planetary Science Archive (PSA) is the central repository for data returned by all of their planetary missions. The long-term preservation of these data and associated knowledge is our core focus. All data provided within the Planetary Science Archive are therefore passed through a set of rigorous procedures designed to ensure the usability of the data not only at the time of ingestion, but also in the long-term, after the mission has closed and direct support from personnel involved with the mission can no longer be guaranteed. Currently, the PSA is hosting data from Mars Express, Venus Express, Giotto, Huygens, Halley (ground based data), and is handling data from Rosetta, SMART-1 and Chandrayaan-1 in preparation for public release.

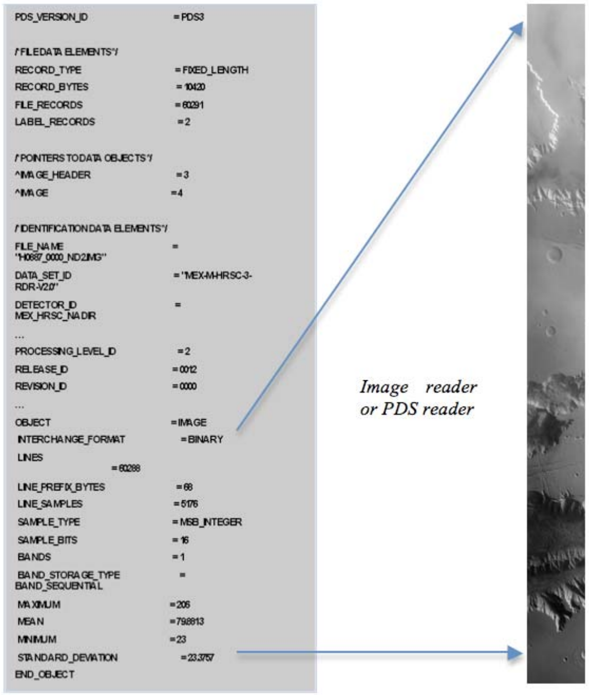
This paper will discuss the data handling and validation procedures put in place to ensure the long-term usability, integrity and compatibility of the PSA archives with as wide a range of data from other planetary archiving authorities as is possible. The standards used to define data formats and construct data sets will be introduced, and a summary of the various validation tools used to guarantee compliance presented.

## ARCHIVING STANDARDS FOR PLANETARY DATA

All PSA data are compliant with NASA's Planetary Data System (PDS) Standards [1] for formatting and labeling files, including requirements for documentation and the structuring of data sets. The PDS

Standards are widely accepted and understood within the planetary science community, having been used on data returned by the majority of previous planetary missions.  Adopting these Standards is therefore the first step in ensuring that the data stored in the PSA are as widely usable as possible and will remain so for many years to come.

Within the PDS Standards, data are organized into volumes and data sets, collecting together observations of similar type, processing level, and/or from a specific mission phase or observation campaign.  All data products are given an ASCII label, either attached directly to the product or, more commonly, provided as a separate label (LBL) file with the same root name as the product being described.  Typically, a ‘one data product, one label’ rule is followed, and is strongly encouraged within the PSA.  A subset of a label file is provided in Figure 1.  This label is from an imaging camera, and it is clear that all information required for software to be able to correctly interpret and display the data is provided in ASCII format as part of the label, thus ensuring long-term usability.



```
PDS_VERSION_ID              = PDS3

/* FILE DATA ELEMENTS */
RECORD_TYPE                 = FIXED_LENGTH
RECORD_BYTES                = 10420
FILE_RECORDS                = 60291
LABEL_RECORDS               = 2

/* POINTERS TO DATA OBJECTS */
^IMAGE_HEADER               = 3
^IMAGE                      = 4

/* IDENTIFICATION DATA ELEMENTS */
FILE_NAME                   =
"H0687_0000_ND2.IMG"
DATA_SET_ID                 = "MEX-M-HRSC-3-
RDR-V2.0"
DETECTOR_ID                 =
MEX_HRSC_NADIR
...
PROCESSING_LEVEL_ID         = 2
RELEASE_ID                  = 0012
REVISION_ID                 = 0000
...
OBJECT                      = IMAGE
 INTERCHANGE_FORMAT          = BINARY
 LINES
                  = 60288
 LINE_PREFIX_BYTES           = 68
 LINE_SAMPLES                = 5176
 SAMPLE_TYPE                 = MSB_INTEGER
 SAMPLE_BITS                 = 16
 BANDS                       = 1
 BAND_STORAGE_TYPE           =
BAND_SEQUENTIAL
 MAXIMUM                     = 206
 MEAN                        = 79.8813
 MINIMUM                     = 23
 STANDARD_DEVIATION          = 23.3757
END_OBJECT
```

Figure 1: Sample of PDS formatted label file for an imaging camera product, including all information required for any standard software to interpret and open the image.

As well as the data products themselves, the PDS Standards have strict requirements for documentation associated with each data set.  A collection of ASCII formatted ‘CATALOG’ files is required to provide fundamental information on each key aspect of the data set.  These ‘*.CAT’ files require that descriptions are provided of the instrument, the mission, the data set, any software delivered, and a number of other key data set components.  It is critical that these are carefully validated before ingestion as they are typically among the first files end-users will look through to understand a data set before undertaking any science or study with the data.  For PSA, much of the information later used to query for a given data set or data product is extracted from the catalog files, so it is also important to ensure that all such keywords and values are correctly provided prior to ingestion.

Also critical to the understanding of a data set is the Experimenter to Archive Interface Control Document (EAICD), ESA’s equivalent to the Software Interface Specification (SIS) in NASA volumes.  This is a required document in all data sets and should cover all aspects of the data and data sets being delivered to allow an end-user to complete some useful studies.  A good deal of effort is invested into the production and validation of these files well before a data itself is delivered for review.

# PLANETARY ARCHIVING WITHIN ESA – BASIC CONCEPTS

The preparation of the data in the PSA is typically the responsibility of the individual instrument teams, mostly located in Europe. PSA staff support the instrument teams throughout the entire archiving process, defining the conventions to be followed at all 'levels' in the archive (volume/data set/data product and mission/instrument/sensor). In this way, we can ensure there is consistency in all data from the same mission and for data from similar instruments, maximizing the potential for cross-instrument and cross-mission data usage.

Through the development and active mission phases, control of the archive development and of the routine deliveries is regulated with the instrument teams via the 'Data Archive Working Group' (DAWG). This group, coordinated by the PSA, is set up on each mission and contains all of the data producers and relevant science personnel required to ensure the data produced are of a good archiving quality and will remain useful in years to come. DAWG meetings and teleconferences are held regularly, with face-to-face meetings in line with all of the Science Working Team meetings.

## Initial Archive Development

The overall concept for archiving is documented in an official 'Archive Plan' put together for each of ESA's planetary missions. This is agreed and signed by all PI teams and ESA mission management. The Archive Plan is typically produced in line with the 'Archive Conventions' document that outlines all of the mission-wide keywords, values and concepts to be used by all instruments. Both of these documents are put together by the PSA in close coordination with the lead scientist of the mission (Project Scientist in ESA), ensuring all top-level aspects of the archive are defined in a useful way to allow for inter-instrument data analyses. It is at this stage that instrument co-operatives can also be highlighted. For example, a specific mission may have a suite of instruments that are very well suited to work together on a given science objective. In this case, the archive should highlight this possibility and all documentation, calibration, software and even data labeling for these instruments should be conceptualized to allow for easy cross-instrument analyses.

Once an instrument team has the initial concept and documentation, they put together the Experimenter to Archive Interface Control Document (EAICD), detailing the archive for the given instrument. This is subject to rigorous internal validation. Once agreed internally, this document is placed in to the first step of the PSA's review procedure.

## Independent Review Cycle

Independent reviews are critical to ensuring that data are useful and suitable for long-term archiving. ESA typically follow a 3-step approach to this for their planetary missions:

- *EAICD Review:* internal but independent review of the EAICD document detailing the intended delivery and data structures. Normally the document is delivered with a small sample of example data.
- *First Delivery Review:* a full-scale review of all data delivered for the first major release of data from a mission. Complete data sets are submitted for review. Scientists (potential end-users of the archive) are requested to assess the usability and suitability of the data sets for inclusion in the long-term archive.
- *Final Mission Archive Review:* a full-scale independent review at the close of the mission to assess the complete archive.

For long missions such as Rosetta, with several fly-by phases of asteroids, Earth and Mars before finally reaching the comet, full independent reviews are organized for each major encounter.

Reviews are typically face-to-face meetings, with reviewers having access to the data four to six weeks prior to the meeting. These are very resource-intensive activities, but are invaluable to ensuring the quality of the data being archived.

## PSA VERIFICATION AND VALIDATION PROCEDURES

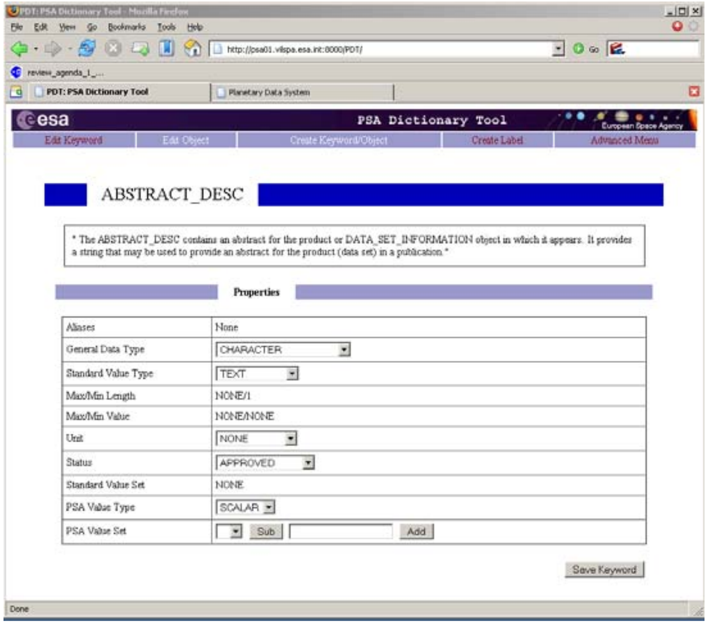
All data delivered to the PSA are subjected to a set of rigorous validation procedures. These are completed at all stages of a mission, from the first delivery through the ‘routine’ archiving phase (i.e. after the first delivery review when the pipeline for data production is frozen) until the final delivery is ingested and the mission archive is frozen. The basic procedure followed for all data deliveries and ingestions is:

- *PVV Syntax validation:* the PVV tool is run on all data sets and volumes. This ensures syntax and the presence of all required keywords. No data can be ingested without full PVV compliance.
- *PVS:* for most data set, the PSA Validation System (PVS) tool is run. This completes more qualitative checks and ensures keyword values are consistent with the data themselves.
- *Manual Checks:* a series of spot-checks are made on all data deliveries to ensure data can be read and are useable. For stand-alone data sets, these checks are made on all file types in a data set (i.e. data, documentation, geometry, calibration etc.). For teams delivering data incrementally to the same data set, spot-checks on the new data are all that is required after the initial delivery and ingestion. The manual checks required of course vary greatly between the different instruments delivering data.

The validation procedure relies upon requirements drawn from all archive documentation, with the Archive Plan and Archive Conventions at the top level, and the instrument EAICD and Review Reports providing lower-level requirements. These are input into the system primarily via the ‘PSA Dictionary’ and are then verified automatically by the PSA Volume Verifier (PVV) and PSA Validation System (PVS) tools.

## PSA Dictionary and the PDT

As already discussed, the PDS Standards are based around a ‘Data Dictionary’ containing a set of keywords that can be used to provide all of the information in the label and catalog file that are required to access and analyse the data being prepared for archive. PSA maintain their own ‘PSA Data Dictionary’, built up from the PDS version and appending many of their own ‘local data dictionaries’ to specify information pertinent only to individual ESA missions.



ABSTRACT_DESC

" The ABSTRACT_DESC contains an abstract for the product or DATA_SET_INFORMATION object in which it appears. It provides a string that may be used to provide an abstract for the product (data set) in a publication."

Properties

| | |
|---|---|
| Aliases | None |
| General Data Type | CHARACTER |
| Standard Value Type | TEXT |
| Max/Min Length | NONE/1 |
| Max/Min Value | NONE/NONE |
| Unit | NONE |
| Status | APPROVED |
| Standard Value Set | NONE |
| PSA Value Type | SCALAR |
| PSA Value Set | |

Figure 2: The PSA Dictionary Tool (PDT) used for maintaining and updating the dictionary

The PSA Dictionary also allows for ‘labels’ to be defined at all levels within the archive. These specify the required keywords for all labels belonging to a given archive, mission, instrument or sensor. For example, a ‘PSA_LABEL’ is defined at the highest level listing the keywords that are required in all data sets before they can go into the PSA. At the other end of the scale, labels are defined for individual sensors from a given instrument, such as the Ion Mass Analyzer (IMA) sensor from the ASPERA-3 instrument on Mars Express. This label will list all those keywords required for data products from that sensor. If a team produces a data product label without one of these required keywords, the PVV software will throw an error and the team will be unable to deliver the data. This allows us to ensure that all critical keywords are provided for each instrument, and also that at the first-level, all data labels in a given data set will be identical.

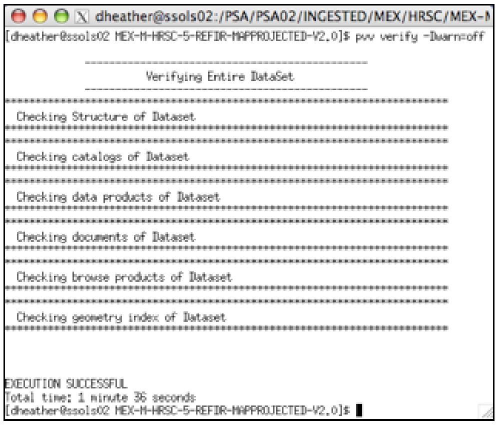
In order to maintain and update this dictionary, a web-based tool has been developed (Figure 2), allowing for new values to be added to keywords (where permitted by the PDS Standards), new keywords to be created - especially important for mission-specific elements, and also allows for checking of keyword lengths and standard value types within the dictionary.

Maintenance of the dictionary is a major job, and it has been found that the PSA and PDS dictionaries have diverged as missions developed and new requirements arose on both sides. PSA staff therefore work in close collaboration with the PDS as values are added to the dictionary, and regular merging of the PDS Dictionary with the PSA Dictionary is required to ensure compliance with both the PSA and PDS.

## PVV and PVS

Compliance with the conventions and requirements on each mission / instrument, and with the PDS Standards is verified using the PVV validation tool developed by the PSA and distributed to all data providers, allowing them to syntactically validate their data at all phases in development of their pipelines, and before each delivery to the PSA.

The PVV checks the data set structure, the syntactic requirements of all labels and catalog files, and a number for other PDS requirements such as line lengths for ASCII files, and date and time formats. The tool can also be used to create the standard INDEX and BROWSE_INDEX files that provide a list of all data and browse products in a data set. All instrument teams are required to run the PVV on their data sets before delivering to the PSA. The verification of a delivery can be completed on a complete standalone data set, on a ‘release’ constituting an incremental delivery of data to an existing data set, or on a ‘revision’ of any file(s) already existing in the archive. The PVV runs from the command line, and a screenshot from this is provided in Figure 3 (left).

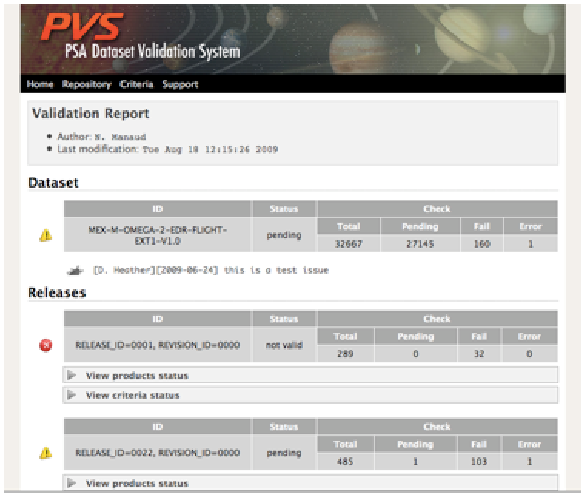


```
[dheather@ssols02 MEX-M-HRSC-5-REFDR-MAPPROJECTED-V2.0]$ pvv verify -Dwarn=off

                ------------------------------------------------
                            Verifying Entire DataSet
                ------------------------------------------------
**************************************************************************
  Checking Structure of Dataset
**************************************************************************
**************************************************************************
  Checking catalogs of Dataset
**************************************************************************
**************************************************************************
  Checking data products of Dataset
**************************************************************************
**************************************************************************
  Checking documents of Dataset
**************************************************************************
**************************************************************************
  Checking browse products of Dataset
**************************************************************************
**************************************************************************
  Checking geometry index of Dataset
**************************************************************************



EXECUTION SUCCESSFUL
Total time: 1 minute 36 seconds
[dheather@ssols02 MEX-M-HRSC-5-REFDR-MAPPROJECTED-V2.0]$ █
```



Validation Report

- Author: N. Manaud
- Last modification: Tue Aug 18 12:15:26 2009

Dataset

|  | ID | Status | Check |  |  |  |
|---|---|---|---|---|---|---|
|  |  |  | Total | Pending | Fail | Error |
| ⚠ | MEX-M-OMEGA-2-EDR-FLIGHT-EXT1-V1.0 | pending | 32667 | 27145 | 160 | 1 |

[D. Heather][2009-06-24] this is a test issue

Releases

|  | ID | Status | Check |  |  |  |
|---|---|---|---|---|---|---|
|  |  |  | Total | Pending | Fail | Error |
| ⊗ | RELEASE_ID=0001, REVISION_ID=0000 | not valid | 289 | 0 | 32 | 0 |

|  | ID | Status | Check |  |  |  |
|---|---|---|---|---|---|---|
|  |  |  | Total | Pending | Fail | Error |
| ⚠ | RELEASE_ID=0022, REVISION_ID=0000 | pending | 485 | 1 | 103 | 1 |

Figure 3: The PVV (left) and PVS (right) validation tools

A further more qualitative validation step is completed on many data sets at the PSA aiming to ensure correctness, completeness and cross correlation of all information, label and data content, within a data set. This is completed by the PSA Validation System (PVS), running within IDL and provided a report that can be visualized via a website (Figure 3, right). This is an internal PSA tool and there are no plans for distribution. The nature of the criteria being validated means that the output requires a degree of intelligence to interpret, and not all errors need fixing. In most cases, the critical aspects of the report can be included in an ERRATA.TXT file in a data set to flag issues that an end-user should be aware of when using the data.

## USER INTERFACES

After validation is complete, data are imported and ingested into the PSA and are provided to the end-user via three interfaces (Figure 4), all linked to from the PSA website (http://www.rssd.esa.int/PSA):

- *Advanced Search:* complex searches of various parameters are possible with this interface.
- *Map-based:* for Mars Express, data from two instruments (HRSC and OMEGA) are available via a map-based interface. Work is ongoing for map interfaces for other instruments and missions.
- *FTP Access:* all public data are available for direct ftp download from the website.

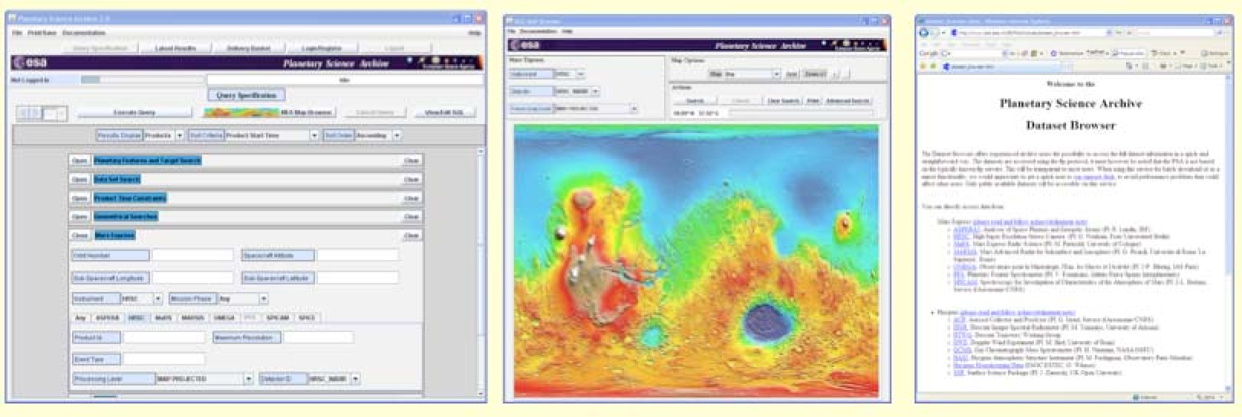

esa
Planetary Science Archive
Query Specification
Planetary Science Archive
Welcome to the
Planetary Science Archive
Dataset Browser

Figure 4: The Advanced Search (left), Map-based (centre) and FTP (right) user interfaces

## Linking the Dictionary to the Advanced UI

All parameters that can be queried in the Advanced Search Interface must be available for ingestion into the PSA database, either from a data product label or associated file in the data set. The Advanced User Interface definition for a given instrument and required queries must therefore be determined quite early in the data preparation in order to ensure that all parameters are included in the relevant labels defined within the PSA Dictionary.

## FUTURE INTERNATIONALISATION OF THE STANDARDS (IPDA)

The PSA work very closely with experts at NASA's PDS as the Standards continue to develop. In order to ensure cross-compatibility of data with other archiving authorities around the world, a group of archiving experts from all major countries involved in planetary exploration gathered together to form the International Planetary Data Alliance (IPDA). One of the main objectives of this group is to try to develop data and archives that are inter-operable. The validation of data to allow for this interoperability will be a future requirement in addition to the standard validations already incorporated in the set-up of the PSA and discussed in this paper.

## SUMMARY AND CONCLUSIONS

In order to ensure long-term usability, compatibility and to maintain the scientific integrity of the data in the long-term, effort must be put in at a very early stage in data preparation and be maintained throughout the entire mission lifetime. PSA personnel provide support to ESA's instrument teams as they develop their data, and validation of data and documentation is completed at many levels and in several stages.  Independent external reviews of data by scientists are of critical importance to ensure the scientific quality and usability of data.

A rigorous set of syntactic and qualitative validation steps are completed on all data delivered to the PSA.  This is done by the PVV and PVS tools and via a number of manual checks that vary depending upon the data being validated.  In order for these tools and checks to work a robust infrastructure is needed, with labels, keywords and keyword values carefully defined within a PSA Dictionary.

In close collaboration with NASA and the PDS, the Archiving Standards for planetary data are being internationalized and streamlined.  These activities form part of PSA's involvement in the International Planetary Data Alliance (IPDA).

It is hoped that the validation system set-up and used by the PSA will allow for all of ESA's planetary data to be available and useful to the community for many years after the mission has ended, and that they will be as compatible with as many other planetary archive data as possible.

## REFERENCES

[1] – Caltech, JPL, Planetary Data Standards Reference Version 3.8 JPL D-7669, Part 2, February 27, 2009

The authors would like to acknowledge all of ESA's instrument teams for the missions included in the PSA, who have helped shape the PSA and provided invaluable feedback on the various validation steps and techniques we use.