

Validation of ESO Phase 3 data submissions

N. Delmotte^{*a}, M. Arnaboldi^a, L. Mascetti^b, A. Micol^a, J. Retzlaff^a

^aEuropean Southern Observatory, Karl-Schwarzschild Strasse 2, 85748 Garching bei München, Germany; ^bTERMA GmbH, Europahaus, Europaplatz 5, 64293 Darmstadt, Germany

ABSTRACT

The data validation phase is an essential step of the Phase 3 process at ESO that is defining and providing an infrastructure to deal with interactions between the data producers and the archive. We are using a controlled process to systematically review all Phase 3 data submissions to ensure a homogeneous and consistent science archive with well traceable and characterised data products, to the benefits of archive users. How the Phase 3 data validation plan is defined and how its results are subsequently managed will be described in the presentation. For a description of its technical implementation, please refer to the contribution by L. Mascetti.

Keywords: archive operations, metadata curation, quality control, data validation, astronomical surveys, data management, ESO, Phase 3

1. INTRODUCTION

Phase 3¹ denotes the process of preparation, validation, and ingestion of science data products for storage in the ESO Science Archive Facility, and subsequent data publication to the scientific community. Science data products are produced by 1) principal investigators of ESO observing programmes, and 2) ESO pipelines as part of the quality control process or from specific, dedicated, re-processing projects for homogeneous raw data sets.

To ensure the successful integration of science data products into the archive, ESO supports the users in carrying out the Phase 3 process by defining data standards, by devising procedures and providing the infrastructure for the delivery of data, and by supplying tools for the data preparation. A general overview of the Phase 3 submission workflow is described in [1].

The mission of the archive being to preserve and re-distribute data, this requires some level of validation on the contributed data to build high-quality, trusted content to the benefits of archive users. In particular, the validation process has the goals to:

- Ensure that data meet the specifications of the submission policies².
- Ensure homogeneity and completeness of the published data
- Ensure well traceable and characterized data products
- Capture relevant metadata at submission time to satisfy archive user requirements (e.g. data discovery services). This is achieved by defining a data interface document, the so-called ESO Science Data Products standard [3]. It ensures a coherent and comprehensive high quality description of the data.
- Identify any errors before ingestion of the data in the archive and before usage of the data by the community.
- Ensure homogeneous, reliable user documentation of the data.

Those goals are achieved by carrying out audits of the data submissions, following a well defined validation plan embedded within the Phase 3 workflow.

*ndelmott@eso.org

¹<http://www.eso.org/sci/observing/phase3.html>

²http://www.eso.org/sci/observing/phase3/policies_eps.html

2. CONTEXT

As a service provided by ESO's archive science group, the validation of Phase 3 data submissions supports data producers when interfacing with the ESO science archive. Data producers are typically Public Survey [2] teams and PIs of Large Programmes. For external data deliveries, ESO's policies governing Phase 3 are specific to the type of observing programme. Phase 3 is mandatory for ESO Public Surveys and for ESO Large Programmes since period 75. For other ESO programmes there is no obligation but PIs are invited to take advantage of the Phase 3. Another major producer of data is ESO itself, contributing pipeline reduced products generated in-house.

Phase 3 deals with science-ready products only (no raw data). The types of incoming data are very heterogeneous. Currently supported are images, spectra, flux maps, cubes, source lists, and catalogues. To allow for unified archive queries across those heterogeneous products, a mapping of scientific parameters to common concepts has been introduced (e.g. PSF_FWHM, SKY_RES, SPAT_RES), hence also the need to check the correct implementation of their definitions by validating the contents of the corresponding quantities.

Phase 3 data validation takes place before the data ingestion phase in the archive and after the following steps are completed:

- Submission agreement with the data producer in place (which data and information is to be delivered, how the structure of the submission is organized, delivery schedule).
- Science Data Product standard [3] available.
- Data transfer to the Phase 3 FTP staging area done.

To review the data submissions and carry out their validation, several elements acting as reference points are available to the archive science group. Those are:

- The Phase 3 policies and submission agreement.
- The Science Data Products standard [3], that is the core Producer-Archive interface document defining the data format and structure, the content information (FITS header metadata) and the encoding of file associations.
- The data release description document delivered by the data producer and that acts as a user manual to the data release. It provides a short broad overview of the program, with an overview/layout of the observations. It is essential input for data content validation. It covers the following aspects: release content (extended listing for each sky position, filters, exposure times, seeing), release notes (reduction method used, calibration procedures, data quality, known issues), data format (description of files in the data release, associated files and naming conventions), relation to previous data releases (if any), acknowledgments (bibliographic reference to be included when using these data).
- Dependencies with other data already stored in the archive. Provenance information of reduced data are generally provided in terms of pointers to the original data (e.g. PROV keywords, pointing to previous data releases or raw data depending on the type of product). Updating releases must contain references to previously released Phase 3 data. At this point it is also taken into account if the data provider reports a problem related to the quality of previously released data.

3. VALIDATION PLAN

The Phase 3 data validation is split in a series of sequential steps. First run two instances of the automatic validation, namely the file-level verification and the batch-level verification. Both are described in Section 4. Finally, an in-depth content validation is carried out by the members of the archive science group responsible of Phase 3 operations. It is described in Section 5. The general flow chart of the entire ESO Phase 3 validation plan is depicted in Figure 1.

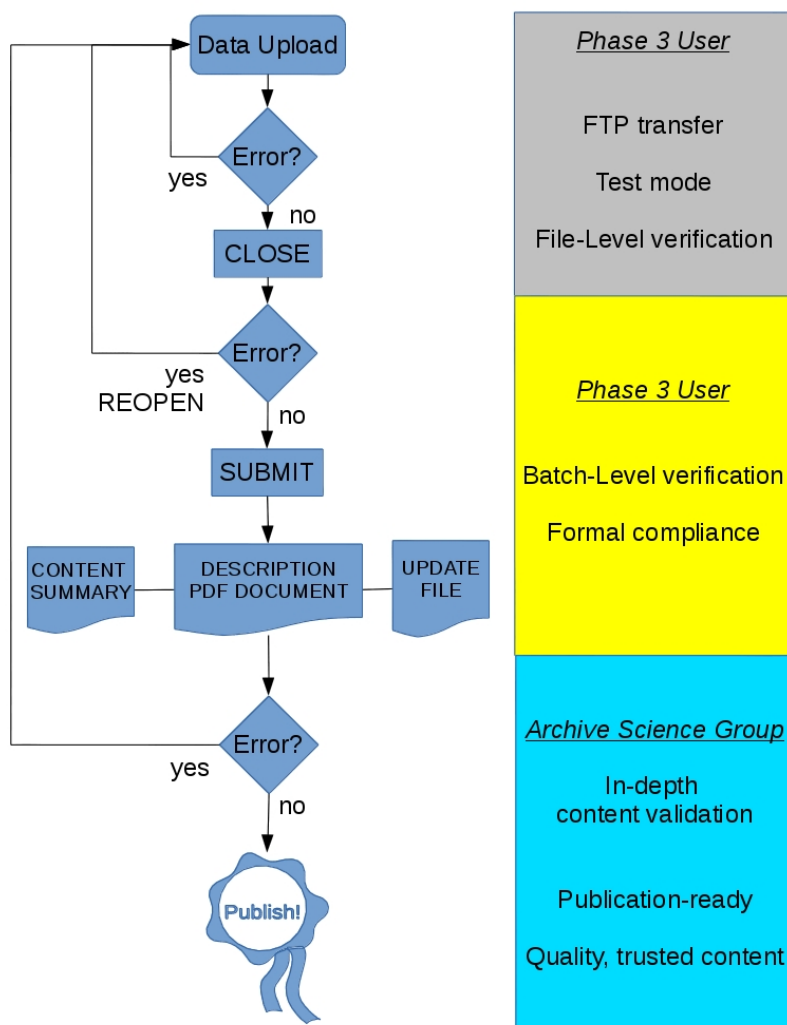


Figure 1. Flow chart of the ESO Phase 3 validation plan, including the different levels of verification.

Data validation comes in different flavours. Some checks are in-built (executed automatically by the P3 system) whereas others are manual, that is involving a particular assessment by a member of the archive science group. Some checks are complete (over the entire submission) whereas others are spot-checks.

Phase 3 also supports the concept of initial validation test. It gives the possibility to the data producer to verify the formal compliance of a test data set with data format requirements, and without actually submitting those data for archiving and publication, i.e. without going through the entire Phase 3 data submission work flow. In addition to the formal checks, a preliminary analysis may also be carried out by the archive science group, though by no means

complete because some checks require the entire data objects to be submitted in order to be executed. Still, experience shows that a preliminary analysis helps identifying early on any major errors or problems.

4. AUTOMATIC VALIDATION

The automatic validation is split in two parts and a description of its technical implementation is available in [4]. First the system does an automatic file by file verification that takes place at upload time to provide immediate feedback on the compliance of the data with the format requirements. The verification involves the following:

- Integrity checks to spot any corruption of data during transfer and checks against the FITS standard. In particular the HEASARC fitsverify utility is run.
- Data objects must map to the object categories defined by the science data products standard and their attributes must conform to the data dictionaries.
- Existence of mandatory metadata depending on the detailed format.
- Checks on metadata types.
- Checks on inapplicable keywords (partial).
- A few specific checks are carried out at this stage, like variable-length arrays not allowed for 1D spectra.

The data producer can view the results of the file-level verification from the Phase 3 release manager³ application, an online tool that is available to both the data producer and the Phase 3 operator to monitor and manage data submissions.. An example of the file-level verification report is shown in Figure 2.

Then the system will proceed with verifying the compliance of the data at the so-called batch level, that is taking into account all datasets and files uploaded. It takes place upon closure of the data submission with the Phase 3 release manager. The close signal triggers the Phase 3 format and provenance verification process on the ESO server, which requires time between a couple of minutes and several hours to complete depending on the amount of data. To help the producer assess and confirm the completeness and consistency of the uploaded data, the Phase 3 Release Manager application provides a detailed summary of the uploaded data in terms of content, e.g. total number of files per data types, date of observation, sky coverage, filter. This proved to be helpful to spot outliers already at this stage.


In the list below we give examples of checks carried out as part of the batch-level validation:

- Check that the associations of the science files with the ancillary files are correct and complete. For instance, are there any associated files declared in the FITS headers but missing from the disk or conversely, were files uploaded to the system but not being part of any association.
- Check the file provenance pointers to maintain referential integrity: do the files pointed to really exist, either in the archive or within the current submission area.

Any error at this stage is immediately reported and email communication is sent to the producer with instructions for revision of contents. This feedback is also available via the Phase 3 release manager, as illustrated in Figure 3.


The (successful) automatic verification signals that the data are formally compliant with the ESO science data products standard. Though a necessary condition, being formally compliant with the standard does not guarantee that the data content is consistently documented to the required level of detail that allows further scientific exploitation independent of the survey teams. The process that brings the science data products from the formally-compliant to the publication-ready state is the in-depth content validation, further described in the section 5 below.

³<http://www.eso.org/rm>



European
Southern
Observatory

ESO — Reaching New Heights in Astronomy



ESO Home

User Portal

Contact

Site Map

Nausicaa Delmotte-Ops (Admin) | Logout

Science User Information

ESO User Portal

Phase 3 Release Manager

Batch 979 files in error

Download

Data Collection: NDE_TEST

FTP directory: ftp://phase3ftp.eso.org/NDE_TEST/batch_979

File name	Category	Date	Size (Bytes)	Error message
science_ancillary.fits	-	-	472320	Missing PRODCATG
science_rbnspectra.fits	-	-	576000	Missing PRODCATG
science_spectrum_074.fits	-	-	262080	Found keyword ASSOC1 associated to an AS...
science_spectrum_008.fits	-	-	259200	Found keyword ASSOC1 associated to an AS...
science_spectrum_038.fits	-	-	262080	Found keyword ASSOC1 associated to an AS...
science_spectrum_01.fits	-	-	172800	Found keyword ASSOC1 associated to an AS...
science_spectrum_106.fits	-	-	262080	Found keyword ASSOC1 associated to an AS...
science_ancillary.fits	-	-	472320	Missing PRODCATG
science_spectrum_083.fits	-	-	259200	Found keyword ASSOC1 associated to an AS...
science_spectrum_059.fits	-	-	262080	Found keyword ASSOC1 associated to an AS...
science_spectrum_069.fits	-	-	262080	Found keyword ASSOC1 associated to an AS...
science_rbnspectra.fits	-	-	576000	Missing PRODCATG
science_spectrum_035.fits	-	-	259200	Found keyword ASSOC1 associated to an AS...
science_rbnspectra.fits	-	-	4112640	Missing PRODCATG

Figure 2. Automatic validation: example of a report resulting from the file-level validation.

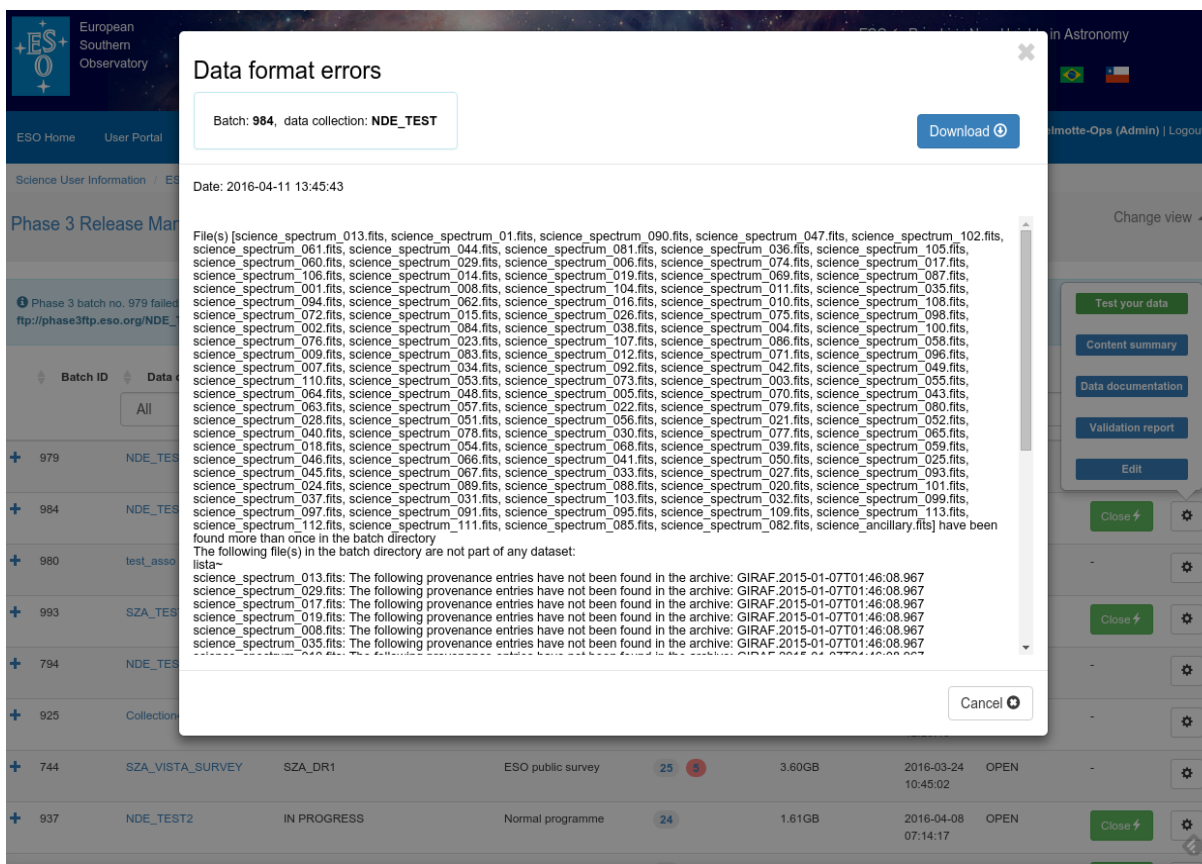


Figure 3. Automatic validation: example of a report resulting from the batch-level validation.

5. IN-DEPTH CONTENT VALIDATION

The in-depth validation takes place after 1) the initial automatic verification is declared successful by the Phase 3 system and 2) the producer has submitted the data and confirmed the completeness and consistency of the actually uploaded data with respect to the data release description. A confirmation of receipt of the submission is sent to the producer. The submitted data and documentation undergo a final check by ESO, before being published through the ESO Science Archive Facility.

At this stage the release description document plays a major role in the review process. It acts as the identity card of the data and will be later published via the archive website along with the data. Hence its contents must be complete and precise enough to understand the nature of the data submission, not only for review purposes, but also for the future archive users to assess the suitability of the release with respect to their own science needs. Hence the release description must be self-contained, clear, and should have a structure in accordance with the template to ensure homogeneity across data releases in the archive. In addition, the information reported in the document must match with the actual submitted products in terms of numbers and types of products, targets fields, passbands, time coverage, spatial coverage.

The other major aspect of the in-depth validation is the focus on detailed metadata verification. FITS header keywords statistics are computed (min/max values, cardinality, distribution of values) to assess the plausibility of metadata values, and spot outliers. Duplication of some metadata (e.g. across header extensions) are flagged as potential source of conflicting information. Blank metadata values are chased and it is verified that unspecified data values are properly encoded using TNULLi/NaN. Metadata items are sometimes related (e.g. total exposure time and Julian date of

observations), hence putting constraints on their values, which is also checked. The source of the products is traced back to the raw data to ensure consistency of the product metadata with the ones of the originating science files.

Here follow a few examples of checks that are carried out depending on the data product type. For images, the keywords RA/DEC must refer to the image center and not the telescope pointing. For catalogue data, the consistency of the UCD identifiers and the uniqueness of the ID of the source records is checked. For 1D spectra, checks are in place to ensure that the wavelength axis values are monotonic. The coordinate spread across all spectra sharing the same OBJECT value is analysed. Some basic visualisation of the data products is also done, to ensure the data were properly encoded and common astronomical tools are able to display them (e.g. fv, ds9).

At the end of the review process, the archive science group provides a detailed validation report to the data producer, notifying either acceptance of the submission or requesting to fix the identified issues. In the latter case, instructions on how to proceed are given. Where applicable, a classification of non-conformities is established, whether it is a data, format, metadata, or release description issue. The validation report includes review items: requests for fixes of the meta/data, or requests for clarification or complementary information. A procedure for the re-submission is defined, including answering each review item to keep track of the progress, organising a new data transfer and setting submission deadlines. Several iterations with the data producer are typical. Once the in-depth content validation is successful, the member of the archive science group triggers the ingestion of the data and the extraction of the metadata into the science database repository for publication through the Science Archive Facility.

6. CONCLUSIONS

A controlled process is used to systematically review and validate all Phase 3 submissions. Validation activities are essentially building quality assurance for the archive and its query services offered to the community. Phase 3 meta/data validation is an essential step towards the more general goal to provide a sustainable and trusted meta/data repository for the science archive.

Validation plans are of course not static and are evolving with time, whenever updates or additions to the science data products interface document are done, and when new functionalities in the validation software are available. For instance, with the Phase 3 upgrade in May 2016, it has become possible to move several validation checks from the in-depth validation to the initial built-in validation to catch errors as soon as possible and thus shorten the validation cycle and reduce the overall time to publication.

The level of validation has a strong impact on the quality of services offered to the archive users. Hence lots of efforts go in checking existence *and* content of the metadata. The higher the quality of the metadata, the better the quality of services to the archive users. Thanks to the controlled quality of their metadata and their integration into a common model, Phase 3 data products are ideal candidates to support the development of a new ESO data discovery portal.

The impact of quality archived science data can be measured by analysing the growth of the user community of the ESO science archive and the trends in ESO data-based refereed papers [5] [6].

REFERENCES

- [1] Retzlaff, J., et al., "Publication of science data products through the ESO archive: lessons learned and future evolution," Proc. SPIE 9910, (2016).
- [2] Arnaboldi, M., et al., "Public Surveys at ESO," Proc. SPIE 9910, (2016).
- [3] ESO Science Data Products standard, GEN-SPE-ESO-33000-5335, Issue 5 (2013) <http://www.eso.org/sci/observing/phase3.html>
- [4] Mascetti, L., et al., "ESO Phase 3 automatic data validation: groovy-based tool to assure the compliance of the reduced data with the science data product standard," Proc. SPIE 9910, (2016).
- [5] Romaniello, M., et al., The Messenger, 163, 5-9 (2016).
- [6] Romaniello, M. "Science data management at ESO," Proc. SPIE 9910, (2016).

Please verify that (1) all pages are present, (2) all figures are correct, (3) all fonts and special characters are correct, and (4) all text and figures fit within the red margin lines shown on this review document. Complete formatting information is available at <http://SPIE.org/manuscripts>

Return to the Manage Active Submissions page at <http://spie.org/submissions/tasks.aspx> and approve or disapprove this submission. Your manuscript will not be published without this approval. Please contact author_help@spie.org with any questions or concerns.

