Ensuring the scientific quality of science products from several data providers

Laura Mascetti

M. Arnaboldi, N. Delmotte, A. Micol, J. Retzlaff

ESO Archive Science Group

Terma GmbH

Phase 3

Channel to submit and publish science-ready data for

- Distribution

- Long-term preservation

http://www.eso.org/sci/observing/phase3

It includes

- Data preparation

- Validation

- Ingestion

- publication

Phase 3

17 October 2017

- LIGO-VIRGO Gravitational Wave: Public Release of ESO Follow-up Observations [Read More]

29 September 2017

- First data release of spectra and catalogue resulting from the deep VIMOS survey of the CANDELS UDS and CDFS fields (VANDELS) [Read More]

21 September 2017

- Third data release of images, source lists and aperture matched multi-band catalogue resulting from the VISTA Kilo-degree Infrared Galaxy Survey (VIKING) [Read More]

[More News...]

Data origin

- Pipeline-processed products (ESO)

- Public Surveys: VISTA (6+7), VST (3), SPS (4)

- Large Programmes

- Others

- Telescopes
  - VLT, VLTI, VISTA, VST, ESO-3.6, MPG/ESO-2.2, NTT, APEX, GTC

Data origin

■ Types of data (single and stacked)

> Images (monolithic, MEF)

> Spectra

> IFS (3D cubes)

> Source tables

> Catalogues

> Submm flux maps

> Interferometry

SDP Standard

- **Ensures integration into the Archive**

- **Defines data format, structure and metadata**

- **Defines how to encode**
  - level of calibration
  - scientific quality
  - original files
  - ancillary data
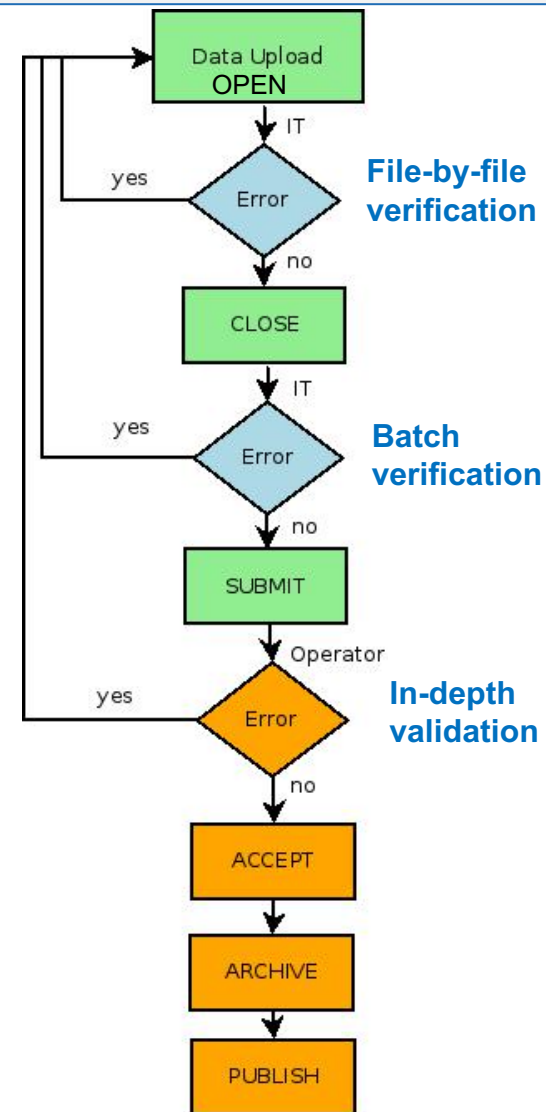
- **Growing document**

- **Allows access from a unique interface**

Audit process needed → **Data Validation**

User support

- Support in generating compliant products
  - Conversion tools
  - Maintain up to date documentation (FAQs, templates)
  - Direct user support
  - Testing data sub-sets

- ESO provides Pipelines that generate Phase3 compliant products

Data flow - Ingestion Tool

- Governs the P3 workflow

- Written in Groovy

- Runs as a demon

- Checks the batches status

- Performs several operations

File-by-file verification

Batch verification

In-depth validation

Data flow - Release Manager
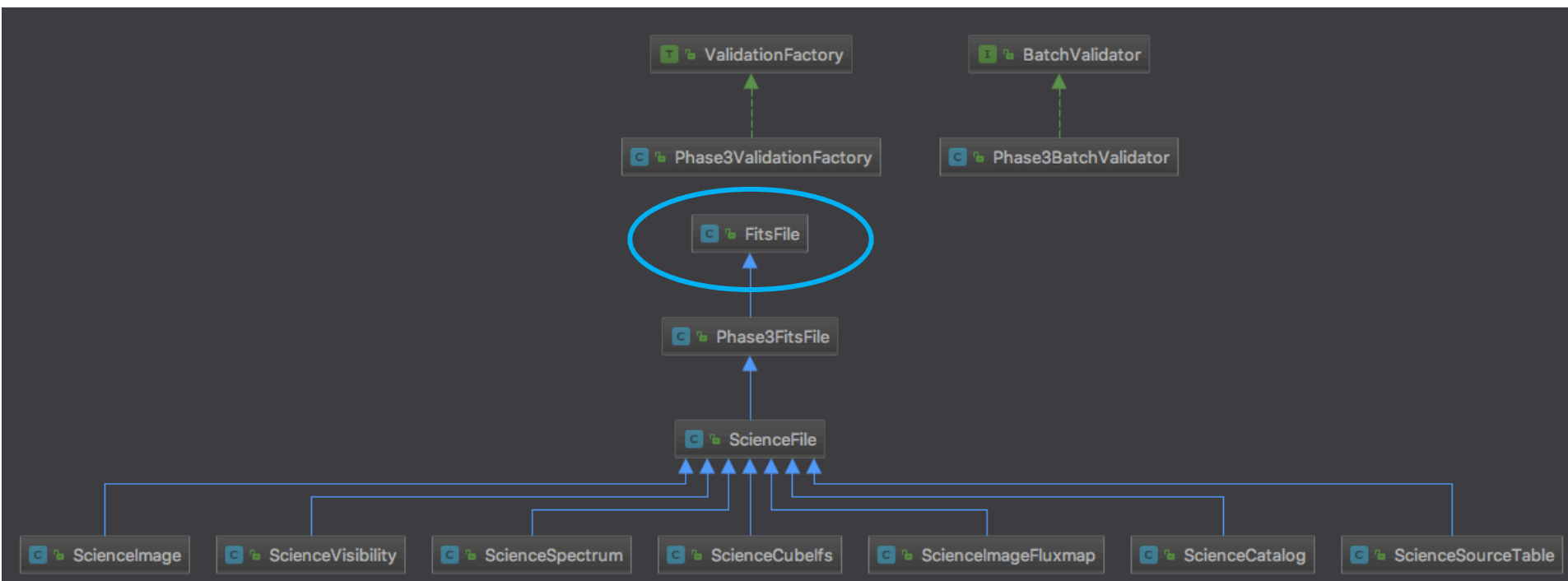
Validation

■ **File-by-file verification**

➢ Triggered upon files upload

➢ It checks:
- • *Fitsverify* / MD5 hash for non-FITS files
- • Mandatory keywords/keywords not allowed
- • Keyword type/values
- • Cross-header checks
- • Checks on the data, consistency data-metadata
- • File structure
- • Mandatory associations

Validation

- Groovy hierarchical validation classes

Validation

■ Batch verification

➢ Triggered by CLOSING the batch

➢ Verifies the batch structure

➢ It checks:
  • PROVENANCE

| Raw files | ← | Pawprints | ← | Tiles | ← | Source lists | ← | catalog |

  • Catalogue tiles structure

Validation

■ **In-depth content validation**

Checks that cannot be performed automatically:

- ➢ Statistics (plausibility, outliers, null values)
- ➢ Checks depending on the type (UCDs)
- ➢ Spot checks
- ➢ Documentation review
  - Consistency with the data
  - Completeness
  - Clearness

Detailed validation report sent to the provider:

- ➢ Request to fix issues → batch re-OPENED
- ➢ Acceptance   → batch ACCEPTED

Conclusion

Validation

- Ensures the science value of the data

- Guarantees quality of services offered

- Enables discoverability

- Enables scientific exploitation

  ➡ legacy value

- Costs resources

  ➡ minimize data curation

Thank you!

Improvements

- ■ Provide as much feedback as possible to the users at early stage

- ■ Consolidate batch validation

- ■ Validation report embedded in the Release Manager