ESO Archive Data and Metadata Model

Adam Dobrzycki, Cristiano Da Rocha, Ignacio Vera, Mỹ Hà Vuong,
Thomas Bierwirth, Vincenzo Forchì, Nathalie Fourniol, Christophe Moins
and Stefano Zampieri

European Southern Observatory
Karl-Schwarzschild-Str. 2
D-85748 Garching bei München, Germany

ABSTRACT

We present the data model utilised in maintaining the lifecycle of astronomical frames in the ESO Archive activities. The principal concept is that complete file metadata are managed separately from the data and merged only upon delivery of the data to the end user. This concept is now applied to all ESO Archive assets: raw observation frames originated in ESO telescopes in all Chilean sites, reduced frames generated intra-ESO using pipeline processing, as well as the processed data generated by the PIs and delivered to the ESO Archive through "Phase 3" infrastructure. We present the implementation details of the model and discuss future applications.

Keywords: databases, data management, observatory operations

1. INTRODUCTION

The ESO Archive is one of the biggest astronomical archives in the world. It combines two roles: operational and scientific. In the first role the Archive safeguards all data ever taken with ESO telescopes and serves as the principal – and, very soon, sole – point of data delivery to users of ESO telescopes. The Archive also tallies and stores all pipeline-processed data generated internally (the so-called Internal Data Products – IDPs) and processed data provided to ESO by the PIs (PIs of Large Programmes and Public Surveys are required to deliver processed data to the ESO Archive; other PIs are encouraged to do so as well). In the second role the Archive publishes and delivers data to external users, both PIs and – after the proprietary time has expired – to archive researchers.

Both roles underwent significant upgrades recently.

The Archive now employs a new infrastructure for handling the inflow of processed data generated by the PIs (known as "Phase 3" data).

On the data delivery end of the activities, the Archive now uses a new request-handling tool, allowing for delivery of data via a download manager. The request handler allows for associating raw frames from the Paranal Observatory instruments with calibration frames.

A data model in which frame metadata are separated from the data and stored in a metadata warehouse is used for all those diverse activities. In this model both data and metadata are considered archive assets. A separate infrastructure is employed for controlling access to the assets and for publishing them.

While the data themselves are never modified, the model allows to update/augment the metadata.

Upon data request, the metadata and data are merged again. The delivered frames reflect the most up to date content.

This paper is organised as follows. In Section 2 we describe the ESO Archive's role in the ESO Data Flow. In Section 3 we describe the principal concepts behind the model and how it is implemented. In Section 4 we describe the relevance of the model to future developments in ESO Archive.

Send correspondence to A.D., E-mail: adam.dobrzycki@eso.org, Telephone: +49 89 3200 6739

2. ARCHIVE AS PART OF THE ESO DATA FLOW

Among astronomical data centres, ESO Archive stands out in several ways. One of distinguishing features is the fact that in addition to being a traditional data centre, it is an operational archive for a ground-based observatory, the La Silla Paranal Observatory (LPO). The Archive has to accomodate not only run-of-the-mill observational data conforming to defined interface specifications, but it also must ingest, account for and deliver to users data frames generated during instrument and detector tests and technical maintenance, as well as ad-hoc, "shoot-from-the-hip" realtime observations of unanticipated transient phenomena. By their very nature these test and realtime frames sometimes push the limits of the interface specifications.

The Archive also ingests pipeline-processed data generated intra-ESO as part of Quality Control and the data delivered to ESO by the PIs of Large Programmes.

The Archive-related infrastructure for handling the data has undergone major evolution in recent years:

- The Data Transfer System[1,2] utilising the internet has become the channel for data delivery from the Chilean sites to the Archive. The data typically arrive in the Archive within minutes from acquisition. The previous system involved shipping data on disks and the typical time it took the data to reach Europe was 7-10 days.
- A warehouse database (Sybase IQ)[3–5] in Garching has gradually become the centerpoint for metadata handling. The metadata repository holds *all* metadata. The old system relied on extraction of selected metadata items and on Chile-Garching replication.
- Pipeline-processed data generated at ESO were incorporated into the aforementioned metadata repository handling scheme. Currently this is an intra-ESO operational activity, but it is mentioned here because it is a pathfinder for the planned evolution of Archive services to include delivery of science-grade products (see Section 4 below).
- A new infrastructure[6] was implemented for ingestion and delivery of PI-generated data products (known as "Phase 3"). This inflow of data has become an integral part of ESO operations. In the past ESO ingested and delivered processed data, but it was done on a case-by-case basis and it was a mostly manual process.
- A new ESO Archive Request Handler[7] has been implemented. The tool directs requests to the download manager, where the user can monitor downloads in realtime.
- The Request Handler can be called with the CalSelector,[8] a service that can associate raw science files with relevant calibration frames.

The common denominator for all those rather diverse activities is the underlying data/metadata model.

3. THE MODEL AND ITS IMPLEMENTATION

The principal concept of the data/metadata model in handling all FITS frames in the Archive is that the file's header is a separate entity from the data part.

3.1 Archive front end: ingesting and publishing

Application of the model to the Archive front end is shown schematically in Figure 1.

The file itself is stored in the Archive and remains there until the actual delivery to the end user. The data part is never modified in any way. In (extremely) rare situations when the data part is compromised/corrupted, a completely new version of the file is generated. The old version – both data and metadata – is flagged as unusable and not published. It can only be made available to the instrument teams to support troubleshooting.

The metadata (header) of each FITS frame entering the Archive are extracted during the archiving process and enter a separate path. For frames taken on Paranal and La Silla sites the headers are read in Chile and transferred to Garching separately (and with higher priority) from the data. For other data this happens in Garching, upon ingestion to the Archive following product verification.

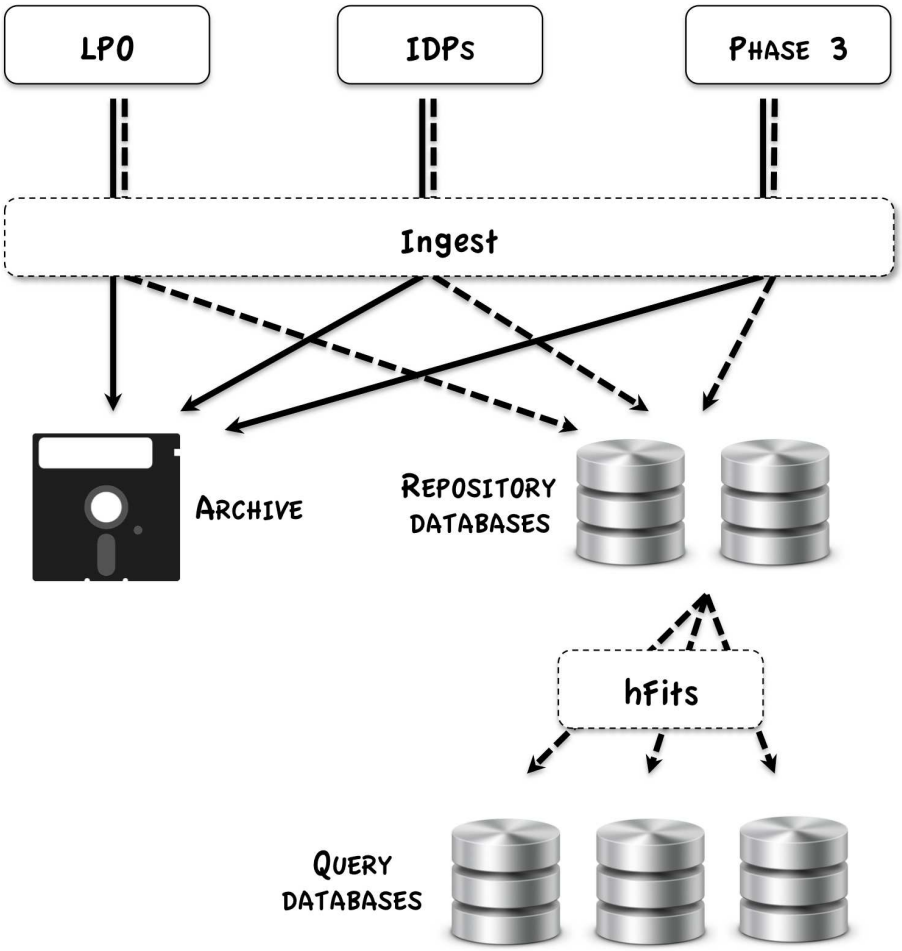

Figure 1. Schematic view of the flow of the data (solid lines) and metadata (dashed lines) in the ESO Archive front end. Upon ingest, metadata is extracted from FITS files and stored in repository databases. Asynchronous "hFits" process moves selected metadata to ESO Archive query database tables. See Secs. 3.1-3.2 for details. For figure clarity we omit the flow of data/metadata access information.

The headers are parsed and the FITS keyword records are extracted into relevant items (keyword, value, type, comment, etc.), which are then loaded into the Sybase IQ-based database called `keywords_repository` (the name should hopefully be self-explanatory). Each frame is accounted for in the accounting database, containing header properties, quality flags and update history.

All back-end Archive services utilising metadata are based on the information from these databases. The principal component are the ESO Archive query tables – database tables tailored for access by the ESO Archive query forms – containing selected metadata for released Archive assets. The tables are populated asynchronously, but with high frequency (currently twice per hour), by the so-called "hFits" tool.[9] The tool identifies newly arrived frames, extracts relevant information from the repository database and stores it in the query tables. The tools also calculates and stores "derived" items, such as ecliptic and galactic coordinates, footprint information, etc.

We mention that all released Archive assets, data and metadata, are assigned access rights according to ESO Data Access Policy. This is not directly relevant to the subject of this paper and is not discussed here in detail.

3.2 Metadata updates

At this moment one of the key features of the model – possibility of metadata update – comes to play. Several considerations made this necessary:

- ESO is on the forefront of astronomical research and its facilities are constantly pushing the limits of technology and science. This leads to frequent instrument upgrades and implementation of new observing modes. Both result in the evolution of interface specifications for ESO instruments and it is often necessary to augment/modify metadata in historical frames in order for the old data to be compliant to the new specifications.
- Importance of certain parameters for scientific analysis is sometimes realised *a posteriori*. A typical case is commissioning and science verification data. These metadata items are added to interface specifications for the subsequent observations, but, for consistency and to enable proper analysis, they also need to be added to the already archived frames.
- At LPO it is sometimes necessary to perform observations outside of prepared schedule (e.g. to observe unexpected transient phenomena). Those data may have incomplete or substandard metadata, which needs to be updated.
- Phase 3 data do have interface specifications which are verified prior to ingestion. However, earlier deliveries of processed files did not have such specifications and their metadata may be inconsistent with Phase 3.

One could argue that the above operations could be performed on the actual frames, new versions of which would be ingested to the Archive. The result would be that the most recent version of the actual files would reflect the current best content and structure specifications. This procedure would, however, be very inefficient. Any of the above updates may be quite simple, and it may have to be performed on a large number of files – and repeatedly. It is thus possible that a small change in the interface specs would result in a large increase in the archive volume.

Performing updates on the metadata only removes this problem completely. Metadata are a small fraction of size of the data, and are much easier to access. Updates on large numbers of frames take a tiny fraction of time and diskspace it would have to take to perform them on the actual files.

The updates are recorded in the accounting database. This serves as a signal to the hFits tool to identify the updates and modify the contents of the query tables accordingly. In that way the query tables – the entry point to all Archive contents inquiries – contain most up-to-date information. The fact that hFits is run several times per day means that the updates are typically reflected in the query tables within minutes of their introduction in the repository database.

3.3 Archive back end: queries and data delivery

Application of the model to the Archive's back end is shown schematically in Figure 2. Upon identifying the files through querying the Archive metadata in the query tables, the user issues an Archive request. The Request Handler (and more specifically the Download Manager) invokes the process in which the data from the Archive are merged with the metadata from the keywords repository database, delivering to the user the file with most up-to-date structure and content.

The scheme (referred to as Headers-On-The-Fly, or "HOTFly") in which the updates are incorporated into the headers of the delivered files has been implemented in the Archive already for some time,[10] but until recently it was based on a small subset of keywords and not on the full keywords repository.

The complete system, allowing for updates/manipulations on any keyword (see also Section 4 below) was incorporated into the release of the new ESO Request Handler in March 2011. We note that the implementation involves a call into the accounting database to verify whether the file in question was updated and only proceed with header modification if it was, thus avoiding the actual run in the vast majority of cases (updates affect only a small fraction of frames). This is in contrast to the old system, where the updates were performed in all cases.

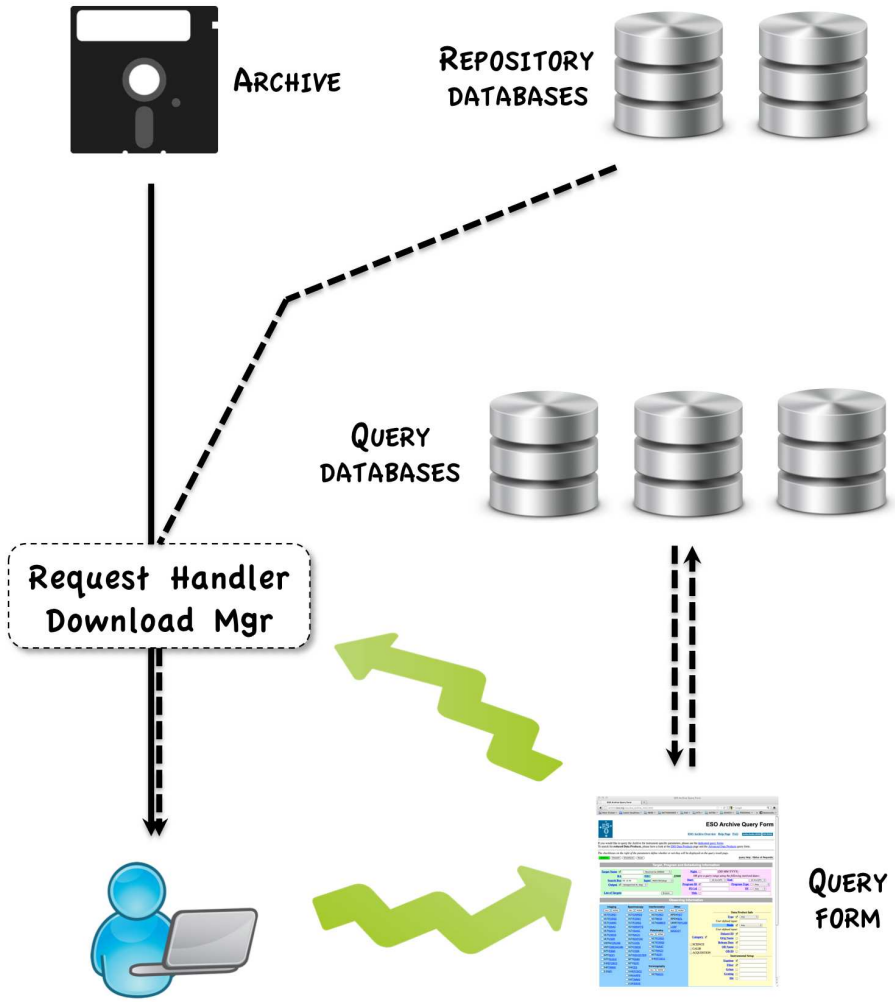

Figure 2. Schematic view of the flow of the data (solid lines) and metadata (dashed lines) in the ESO Archive back end. The users query the released data through a web interface and issue requests. The request handler merges the data and appropriate metadata into the delivered product. See Sec. 3.3 for details. For clarity the flow of data/metadata access information is omitted.

4. ARCHIVE'S EVOLUTION AND THE MODEL'S ROLE

ESO Archive is constantly evolving; a paper presented at this conference[11] describes the long term development plans. In here we will discuss the relevance and applications of the data/metadata model presented in this paper to these plans. An important part of the planned development is putting more emphasis on availability of science-oriented data, both in content and the file structure.

ESO is actively working on the development of pipelines that would enable in-house generation and delivery to users of science-grade data based on publicly available raw data and on definition of data processing environment enabling users to perform their own analysis. Also planned is redesigning of the Archive ingestion process in order to unify, as much as possible, various data inflows.

All planned developments utilise the model presented here. Header extraction, which currently is an offline process done independently for each of the inflows, will be integrated into the unified archiving process.

Generating science-friendly products may, for example, involve removal from headers of metadata related to ESO operations and so uninteresting – and possibly confusing – to the science user. In any case, this process boils down to a header manipulation and as such is just a generalisation of the HOTFly process, which is already

in place. This infrastructure is currently in the planning stages, and so only sketchy details can be presented here. It is not yet known whether this generalisation will be based on the current concept of having database containing the header as it should look in the end product and call a virtually unchanged current HOTFly tool, or whether it will involve extending HOTFly to allow it to provide configurable output. The former option utilises existing software, while the latter appears to be more flexible. Several options for the storage of science products are being considered: they may be archived, cached or generated on the fly for each request (or some combination of these options).

5. CONCLUSION

ESO developed and implemented a data/metadata model in which complete metadata are extracted from the data and handled as a separate asset, and merged again with data only upon delivery to the end user. It works. It will be used in the future.

ACKNOWLEDGMENTS

Important contributions to this project made throughout its history by C. Avelans, D. Brandt, A. Brion, P. Eglitis, J. Lockhart, N. Rainer, M. Romaniello, N. Rossat, J. Rodríguez and A. Wicenec are gratefully acknowledged.

REFERENCES

[1] Zampieri, S., Forchì, V., Gebbinck, M. K., Moins, C., and Padovan, M., "The ESO Data Transfer System," in [*Astronomical Data Analysis Software and Systems XVIII*], Bohlender, D. A., Durand, D., and Dowler, P., eds., *Astronomical Society of the Pacific Conference Series* **411**, 540 (Sept. 2009).

[2] Romaniello, M., Zampieri, S., Cerón, C., Wright, A., Hanuschik, R., Ledoux, C., and Comerón, F., "From Chile to Europe in minutes: handling the data stream from ESO's Paranal Observatory," in [*Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*], *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* **7737** (July 2010).

[3] Dobrzycki, A., Brandt, D., Giot, D., Lockhart, J., Rodríguez, J., Rossat, N., and Vuong, M. H., "Observations Metadata Database at the European Southern Observatory," in [*Astronomical Data Analysis Software and Systems XVI*], Shaw, R. A., Hill, F., and Bell, D. J., eds., *Astronomical Society of the Pacific Conference Series* **376**, 385 (Oct. 2007).

[4] Vuong, M. H., Brion, A., Dobrzycki, A., Malapert, J.-C., and Moins, C., "Applications of the ESO metadata database," in [*Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*], *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* **7016** (July 2008).

[5] Vera, I., Dobrzycki, A., Vuong, M.-H., and Brion, A., "Centralized FITS Metadata Handling Framework at ESO," in [*Astronomical Data Analysis Software and Systems XX*], Evans, I. N., Accomazzi, A., Mink, D. J., and Rots, A. H., eds., *Astronomical Society of the Pacific Conference Series* **442**, 33 (July 2011).

[6] Arnaboldi, M., Retzlaff, J., Slijkhuis, R., Forchì, V., Nunes, P., Sforna, D., Zampieri, S., Bierwirth, T., Comerón, F., Péron, M., Romaniello, M., and Suchar, D., "Phase 3 – Handling Data Products from ESO Public Surveys, Large Programmes and Other Contributions," *The Messenger* **144**, 17–19 (June 2011).

[7] Fourniol, N., Lockhart, J., Suchar, D., Tacconi-Garman, L., Moins, C., Eglitis, P., Bierwirth, T., Vuong, M. H., Micol, A., Delmotte, N., Vera, I., Dobrzycki, A., Forchì, V., Lange, U., and Sogni, F., "News from ESO Archive services: Next Generation Request Handler and Data Access Delegation," in [*Astronomical Data Analysis Software and Systems XXI*], Ballester, P. and Egret, D., eds., *Astronomical Society of the Pacific Conference Series*, TBD, ASP, San Francisco (2012).

[8] "Introduction of Calibration Selection via the ESO Science Archive Facility and Discontinuation of PI Packages," *The Messenger* **146**, 47 (Dec. 2011).

[9] Vera, I., Dobrzycki, A., Vuong, M.-H., and Da Rocha, C., "hFits: From Storing Metadata to Publishing ESO Data," in [*Astronomical Data Analysis Software and Systems XXI*], Ballester, P. and Egret, D., eds., *Astronomical Society of the Pacific Conference Series*, TBD, ASP, San Francisco (2012).

[10] Dobrzycki, A., Delmotte, N., Rossat, N., Pirenne, B., Avelans, C., and Rainer, N., "Data interface control at the European Southern Observatory," in [*Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*], Quinn, P. J. and Bridger, A., eds., *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* **5493**, 117–125 (Sept. 2004).

[11] Romaniello, M., Arnaboldi, M., Ballester, P., Dumas, C., Forchì, V., Freudling, W., Hanuschik, R., Retzlaff, J., Smette, A., and Zampieri, S., "Current status and future directions of the ESO science archive facility: content and services," in [*Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*], *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* (2012). This volume.