Astronomical Data Analysis Software and Systems X
ASP Conference Series, Vol. 238, 2001
F. R. Harnden Jr., F. A. Primini, and H. E. Payne, eds.

An Enhanced Data Flow Scheme to Boost Observatory Mine-ability and Archive Interoperability

Alberto Micol

Space Telescope European Coordinating Facility, European Space Agency

Paola Amico

European Southern Observatory

Abstract. The major astronomical observatories in the world, Gemini, HST, VLT, NGST, etc, invested or plan to invest large amount of human resources and money to build archive facilities that support their data flow. The classical Data Flow Scheme allows its users (calibration/quality control scientists, principal investigators, archive scientists) simple files retrieval and access to basic ambient and calibration data, leaving other valuable information totally unexplored.

Is there more to exploit in a large data archive/data flow? Is it possible to improve the Data Flow Scheme in order to foster the mine-ability of an archive, making, at the same time, the every day life of the quality control scientist easier? What are the common missing steps for an archive/observatory to be miner-ready? We will answer all these questions and suggest a newer approach for a data flow scheme, where the Data Quality and the Archive can be seen as two different clients of the same sub-system: the Observatory Data Warehouse.

1. Introduction

HST and VLT insiders have great familiarity with the concept of Data Flow System (DFS): it is a closed-loop software system, which incorporates various subsystems that track the flow of data all the way from the submission of proposals to storage of the acquired data in the Science Archive Facilities. Typical DFS components are: Program Handling, Observation Handling, Telescope Control System, Science Archive, Pipeline and Quality Control. All these components produce various sorts of data with different formats and "handling" rules. Therefore, the information flow among the subsystems suffers from "hiccups". Ultimately, some data may be lost in the process and the referential integrity within the DFS may be compromised.

ST-ECF and CADC have always been busy in trying to patch the current HST Data Flow but unfortunately, as we will see later, only a posteriori. The nitty-gritty details of on-the-fly calibration of science HST data, the jitter extraction pipeline, the WFPC2 associations, and the FOS associations are examples of patching the engineer-oriented HST Data Flow. Those systems have

introduced a previously missing, basic, scientifically-oriented description of the HST datasets.

It is thanks to this a posteriori effort of reconstructing what actually happened during the observations that higher-level, ready-for-science data products are now immediately available to scientists. Indeed, cosmic ray free co-added images, mosaics of dithered WFPC2 observations, and combinations (at least for a first visual inspection) of FOS spectra are generated on-the-fly upon demand.

In the case of the VLT, soon after the completion of the observing phase of a service mode programme, the PI receives a complete data package, which includes, among other, zeropoints and science frames free of instrument signatures. The archive, and therefore the future user of the science data, does not receive the same information and therefore the work of the quality control scientist, at the back end of the data flow, is partially lost.

2. The Classical Data Flow

After Phase 1 and 2 of the proposal preparation, the observations are scheduled and then executed. Both telemetry and science data are acquired and stored, while some reduction pipeline produces quick look products and a Quality Control team inspects the data, but usually only a few measurements are taken and passed to the PIs. For example, the PIs of VLT programmes receive a Quality Control report, which compares the user requirements in terms of seeing, fraction of lunar illumination, moon distance and airmass with the true values measured on site and stored in the ambient database.

These are certainly necessary steps. But are they sufficient? What is described by an investigator in his/her proposal (pointing information, S/N ratios, image quality etc) doesn't necessarily agree with what the observatory has been able to achieve at run time[1] [2]. Discovering what actually happened during the observations is usually left to the PIs. Furthermore, such effort is then lost since no feedback is given to the Observatory. Moreover, an archive scientist will later have to go through the same reduction for his/her own study. Again no feedback will be provided to the archive facility.

Any data-miner will have to go through those steps again and again. This observatory does not qualify as miner-ready.

---

[1] An HST example: the observed dither pattern can differ from what was actually requested, due to errors (usually at the sub-pixel level) in the positioning of some optical element (e.g., the lack of repeatability of the STIS Mode Selection Mechanism), or because of jitter, or some other effects (aberrations, deformations). Hence the offsets as from the proposal DB (and/or the WCS in the science data header) are not necessarily the correct ones to be used to combine the images. Some other more reliable source of pointing information (e.g., Observatory Monitoring System jitter files in the case of HST) must be used, or otherwise, direct measurements on the images (via cross-correlation techniques) are required.

[2] A VLT example: users do not receive any information on the wind speed and direction with respect to where the telescope points, an important piece of information when excellent image quality is required. Another example comes from the requirement, typical of infrared observations, to monitor the level and the variations of the sky background in bands over 2 microns and, eventually, to measure if it correlates with the mirror temperature and/or the external temperature.

3. A Better Data Flow Scheme

As highlighted in the previous paragraph, there are two main problems in todays' DFS implementations: (1) lack of interoperability within the various DF subsystems (2) insufficiently detailed description of the observations.

Two steps are necessary to overcome these limitations:

(a) The adoption of a Data Warehouse[3] to control the various DFS activities. The information flows among DF subsystems via the data warehouse. The advantages of having it at the centre of the DFS are multiple. Among them, it guarantees a homogeneous access to the information created by any DF subsystem; it may also be the place used to develop and integrate tools to check for referential integrity. **(b)** The introduction of a new DFS component, let's call it "Characterisation" step, responsible for any data manipulation/reduction to extract all the parameters, which are useful for a thorough description of what actually happened during the observations. It should consist of a set of reduction pipelines to measure (and compare) those parameters requested in phase 2 (e.g., offsets of dithered frames, S/N of spectra, image quality, etc).

These Characterisation tasks should be executed at a later time than the Data Quality ones since they require a better calibration, using improved reference files and software typically unavailable at the time of the observations. This special activity should be carried out some time (1 year ?) later[4].

In the end the data warehouse should not only contain Phase 2 and scheduling information, but also:

1. Parameters to be used for instrument trend analysis. (e.g., PSF, noise properties, bias level trends, etc.)
2. Parameters to be used to calibrate the data (e.g., zeropoints)
3. Parameters requested by the PI during phase 2 (e.g seeing, fraction of lunar illumination, etc). Comparison of requested and obtained values will be used for Data Quality assessment.
4. Parameters used to scientifically characterise the observations (Limiting magnitudes, background properties, object detections along with rough preliminary object properties measurements, density of point/extended sources for images, etc.)

---

[3]The term *data warehousing* generally refers to combine many different databases across an entire enterprise and its application is rather general and not at all confined to scientific databases. Data warehousing emphasizes the capture of data from diverse sources - the DF subsystems - for useful analysis and access, but does not generally start from the point-of-view of the end user or knowledge worker who may need access to specialized, sometimes local databases. The latter idea is known as the data mart. These data-marts (files, spread-sheet, DBMS, etc) are used in the day to day operations and allow insertions, updates, deletions or, in other words, all those operations that are not possible in the data warehouse itself. Once consolidated, the information is extracted and translated from the data mart and sent to the data warehouse for permanent storage.

[4]It could be combined with the production of Preview data (1992), which pipeline is run just after the data exits the proprietary period, to ensure the best quality (better calibration s/w, better calibration files) of the products, and to respect the data reservedness. The fact that the same data is processed twice -the first time to assess the data quality the second one to extract parameters useful for a scientific charactersation of the archive contents - helps in understanding better the data, their problems, in discovering and resolving possible inconsistencies.

5. Some preview products (imagettes, binned spectra, histograms, etc) could be generated.
6. A layer with pre-compiled and up-to-date statistics of some of the parameters listed above.

While certain parameters are already measured (mainly 1 and 2 above), others (some of 3, and mainly 4, 5 and 6) are not part of the current Data Flow Systems. Though 1 and 2 above are stored in the so-called calibration database, and though 3, 4, 5 and 6 above could end up into a "characterisation database", more is to be gained by integrating those two aspects within the observatory data warehouse. Having all this information on-line will greatly improve the way an instrument scientist or an archive scientist works.

The mine-ability of the system is greatly enhanced since engineers and scientists, both inside and outside the observatory, will have homogeneous access to information like: **(a)** ready-to-use measurements, **(b)** ready-to-view preview products, **(c)** a scientific view on the archive as opposed to the standard, sterile catalogue (observation log), **(d)** a quality control view on the archive, (trend analysis techniques and instruments/telescope health checks could benefit from monitoring parameters such as the noise levels of detectors, the measured resolution versus time and slit width, the image quality, etc.), **(e)** a superior level of abstraction, since at this level the underlying complexity of the various subsystems that collected the necessary information must have been removed.

Without this level of abstraction, it will be difficult to achieve effective interoperability among archives.

4. Conclusions

We highlighted the typical problems which HST and VLT Data Flow Systems are facing today. Dispersing the information into several subsystems that are not interoperating is the immediate cause of glitches and inconsistencies, which, for the intrinsic heterogeneous nature of the DFS, are then difficult to identify and repair. We claim that a central repository of the information produced by all the various DF subsystems will greatly help to reach smoother operations.

Industry is facing the same kind of problems; indeed data warehousing is one of the hottest industry trends. The astronomical community should try to benefit from that effort.

Up to now an archive user, being an external user or an instrument scientist, has been able to browse through an observation log representing basically Phase2 information. The aim of introducing a characterisation step is to provide not only better information on what actually happened during the observation, but also to provide a higher level interface to the archive: a miner-ready interface which doesn't need nor want to know the details of the particular DFS, but which can help the scientist in identifying the data s/he needs.

A good DFS must be able to remove its own signature. A good Observatory (not only the archive) must be miner-ready.