Chandra Data Archive Operations

Arnold H. Rots, Sherry L. Winkelman, Stéphane Paltani, Edward E. DeLuca

Smithsonian Astrophysical Observatory, 60 Garden Street, Cambridge, MA 02138 (USA)

ABSTRACT

The Chandra Data Archive plays a central role in the Chandra X-ray Center (CXC) that manages the operations of the Chandra X-ray Observatory. We shall give an overview of two salient aspects of the CDA's operations, as they are pertinent to the operation of any large observatory.

First, in the database design it was decided to have a single observation catalog database that controls the entire life cycle of Chandra observations (as opposed to separate databases for uplink and downlink, as is common for many scientific space missions). We will discuss the pros and cons of this design choice and present some lessons learnt.

Second, we shall review the complicated network that consists of Automated (pipeline) Processing, archive ingest, Verification & Validation, reprocessing, data distribution, and public release of observations. The CXC is required to deliver high-level products to its users. This is achieved through a sophisticated system of processing pipelines. However, occasional failures as well as the need to reprocess observations complicate this seemingly simple series of actions. In addition, we need to keep track of allotted and used observing time and of proprietary periods. Central to the solution is the Processing Status Database which is described in more detail in a related poster presentation.

Keywords: data archive, data processing, data distribution, observatory operations, X-ray

1. INTRODUCTION

The life cycle of astronomical observational data starts with an idea born in someone's brain or in a discussion; it ends with the last reading of a paper from the collection of articles that reference the data in some way or other. The involvement in this life cycle of the institution that is responsible for the operation of the telescope with which the observations are made and for the preservation and dissemination of the data products has a definite beginning, but an indefinite end: it starts with the receipt of a proposal but does not end until the data are absolutely obsolete - even for historians.

The tracking of an observation, its specification, status, and subsequent data products — all conceivably in multiple versions — is crucial for the integrity not only of the observatory's operations but also of the data depository that holds the products during and after the mission. There are (at least) two aspects of this process that are crucial to the design and implementation of the observatory's infrastructure: the observation catalog database and the quality control *cum* data release process. As such, their specification ought to be a central driver in the design of all other parts of the observatory's elements. Regrettably, they often end up as afterthoughts that need to be shoehorned into ill-fitting interfaces. On the other hand, we must concede that their own design is usually lagging a schedule that would be appropriate to their central role, generally due to a lack of clear insight early on in their requirements and in the requirements of the other components that they are expected to fit in with.

Send correspondence to AHR; E-mail: arots@head-cfa.harvard.edu

This paper was published in *Observatory Operations to Optimize Scientific Return III*, Peter J. Quinn, Editor, Proceedings of SPIE Vol. 4844, p. 172, and is made available as an electronic reprint with permission of SPIE.

This paper describes and analyzes the role and interfaces of the observation catalog and of the database that is tracking the conglomerate of Automated Processing (AP), Verification and Validation (V&V), and version control system as managed by the Chandra Data Archive (CDA) Operations Group within the Chandra X-ray Center's (CXC) Data Systems division. We hope that the lessons learnt will be useful to future missions and observatories.

2. THE CHANDRA X-RAY OBSERVATORY

The Chandra X-ray Observatory (CXO) is a spacecraft, launched in July 1999 that carries an X-ray telescope with two main instruments, the Advanced CCD Imaging Spectrometer (ACIS) and the High Resolution Camera (HRC), supplemented with optional transmission gratings. It provides imaging and spectrographic capabilities that are unprecedented in the X-ray band. It is the first — and will for a long time be the only — X-ray telescope that provides sub-arcsecond spatial resolution in the 0.2–10.0 keV band. The mission is operated by the Smithsonian Astrophysical Observatory at the Harvard-Smithsonian Center for Astrophysics in Cambridge, MA, under contract with NASA. This operation covers the entire institutional life cycle of the observations: editing the NASA Research Announcement, receiving observing proposals, managing the peer review, Mission Planning (MP), flight operations, receipt and processing of telemetry, V&V, data archiving, and data distribution.

3. THE CHANDRA DATA ARCHIVE (CDA)

The CDA comprises three major components: a database server, a data products warehouse, and interfaces for ingest, search, and retrieval. Functionally, we can distinguish the following components:

- Observation catalog; this actually is a conglomerate of several databases but may be considered one unit for the purposes of this paper.
- Automated Processing (AP) status database; a database that tracks the status of each observation in the automated processing pipelines, in reprocessing, V&V, and data distribution.
- Data products databases; track the location of data products in the data warehouse, hold metadata, and control versioning.
- Data warehouse; contains all versions of all data products.
- Ingest interface; for ingest of data products by production pipelines.
- Search and Retrieval (S&R) interface; for operations personnel and for users.

In this paper we shall describe some of the lessons learnt in connection with the observation catalog and concentrate on the function and working of the AP status database. The Archive Operations team bears responsibility for the maintenance and integrity of the CDA, and thus for these databases.

For completeness we list here the other tasks of the Chandra Data Archive Operations Team to provide the reader a more comprehensive view of archive operations activities and responsibilities.

- Maintain and update the observation catalog, following requests by the Director's Office and and User Support.
- Monitor ingest of data products and repair failures; we built a special database for this.
- Ensure the integrity of the databases and the data product holdings.
- Maintain consistency between various copies of the archive (such as mirrors and the ftp archive).
- Provide the formatting standards for all FITS files that are to be archived.
- Provide usage statistics.

- Distribute proprietary and public data to users.
- Control public release of observations.
- Provide specifications for archive hardware.
- Monitor the performance of archive hardware.
- Monitor database and archive servers.
- Provide specifications for and testing of user interfaces.
- Provide support for integration and testing of new software releases.
- Maintain ftp space for data downloads and various special products, such as calibration files.
- Provide user assistance, for CXC personnel as well as outside users.
- Manage archive user accounts.
- Maintain Chandra bibliography; this requires every month an extensive search
- Maintain the CDA webpages.
- Coordinate and collaborate with other data centers on policy issues and interoperability.

4. THE OBSERVATION CATALOG

Many past missions have strictly separated the management of pre-observation ("uplink") and post-observation ("downlink") databases. There are clear advantages to this approach since the uplink and downlink requirements are very different and since both require a fair amount of flexibility, in the face of many operational uncertainties and rushed schedules. Such a system is simpler, there are fewer complicated interrelations, and hence one can more readily respond to requests for modifications without fear of breaking something on the other side.

On the other hand, having all operational activities controlled by an observation catalog in a single database contributes in no small way to the integrity of the final data holdings. It does require, however, a solid, well thought-out design that is done very early on and takes into account all known requirements of all known interfaces, while also allowing for shifting requirements in all those interfaces. That may sound like a utopian impossibility, but it is actually possible, provided one does a careful system analysis early in the project.

The function of the observation catalog in the CXC is illustrated in Fig. 1. The user generates a proposal from which approved observations are ingested into the observation catalog. The entries contain all pertinent information, such as coordinates and approved time, as well as instrument configurations and observational constraints. This, in turn drives Mission Planning, resulting in weekly observation programs which are uplinked to the spacecraft and fed back into the observation catalog, so that the database knows what the status of each observation is. The downlink telemetry is fed into the AP pipelines which consult the observation catalog on what is expected and what needs to be done with the data, and which, in turn, update the observation catalog with relevant status information and metadata. The AP data products are cached and subsequently ingested into the data archive, at which point metadata are extracted and stored in related databases. Finally, the user may query the observation catalog on the status of his/her observation and browse and retrieve the data through use of the data archive databases.

The first lesson learnt is that the specification of the observation catalog needs to take place very early on, that it needs to be the result of a thorough and wide-ranging requirements analysis, and that it needs to be designed to be able to handle considerable changes in these requirements. On the AP and S&R side this is fairly straightforward, as long as one conducts a thorough analysis of the metadata that are needed for the various functions and one designs explicitly a mechanism that allows new elements to be added seamlessly later on. The same is true for the proposal interface.

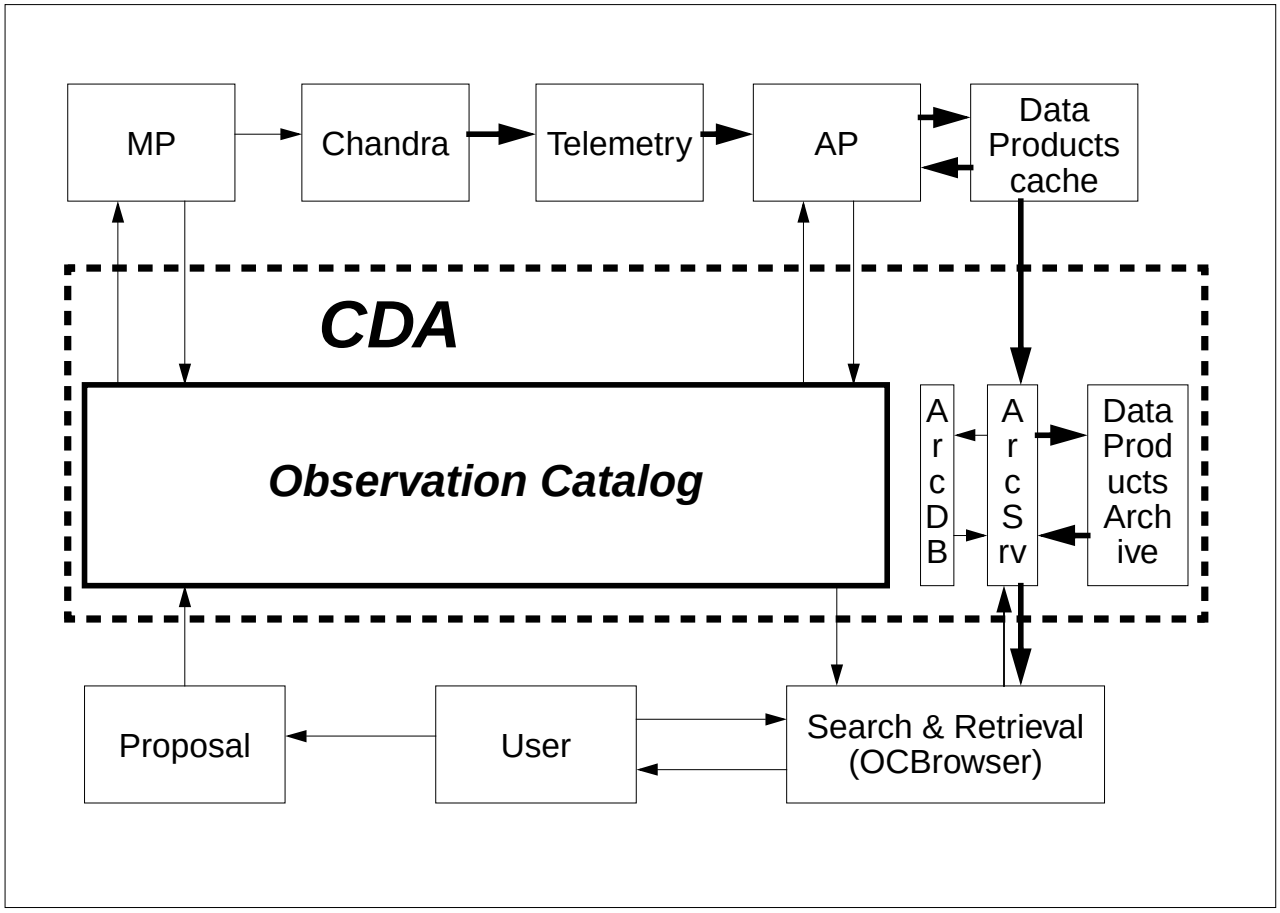

Figure 1. Function of the observation catalog in CXC. The sizes of the boxes in the CDA imply nothing about their relative data volume, but the thickness of the arrows indicates whether the data transfer volume is heavy or not. "Data Products Archive" is the data warehouse, "ArcDB" the data products databases, and "ArcSrv" the interface server.

The MP interface is far trickier and the requirements are much more stringent, in no small measure because MP is usually considered to be a "mission critical" operation. First, one needs to protect the MP process from external changes to the observation catalog that would affect the planning process that is in progress. Second, one needs to build in a mechanism that prevents multiple instances of the MP process to get into each other's way while still allowing (controlled) manual intervention. Third, all this needs to function in a very simple way (i.e., no complicated setting and resetting of locks by the user), and it needs to function without impeding the user in emergency situations, allowing almost any kind of override while preserving all pertinent information. Finally, one needs to be cognizant of the fact that most missions do not operate in the way foreseen prior to launch; hence, the system must allow for major changes in planning procedures. And one needs to realize that flight software and flight operations software do not always perform as assumed or as advertized.

The situation is not as hopeless as it may seem. If one conducts a solid system analysis, one should be able to come up with a good design. But there are four important dos-and-don'ts, based on our experience:

1. Do not assume to know exactly how things are going to work — such an assumption would lead to a design that is too rigid.

2. Try to think through as many alternative and what-if scenarios as possible — anticipation facilitates flexibility.

3. Start on the requirements and specifications of the observation catalog very early on, preferably before any operations group has fixed any interfaces — this encourages synergy.

4. Make sure all possible interfaces and requirements are covered, including, for instance, administrators who need to file monthly performance reports — this minimizes unpleasant surprises.

5. THE PROCESSING STATUS DATABASE

Contrary to what is customary in many other observatories, the CXC is required to deliver high-level products to its users; we expect this to become more common. It is achieved through a sophisticated system of processing pipelines (AP), necessitating a quality assurance component that is implemented in the Verification and Validation (V&V) system. Although this may seem a fairly simple series of consecutive actions, it makes in reality for a rather complicated system of interrelated components, further compounded by occasional failures (for instance, because of missing data), as well as the need to reprocess observations when improved calibration products and processing software become available. In addition, we need to keep track of allotted and used observing time and of proprietary periods. The proprietary period is, in most cases, one year from the shipping of the data, but there are exceptions to the one year period. In addition, the data shipped need to have been of good quality and the distribution needs to have been complete; this last requirement is especially relevant for split observations or observations that have lost time. The sequence of processing pipelines (some two dozen) take the data from raw telemetry, through Levels 0, 0.5, 1, and 1.5 to Level 2. Levels 1 and 2 are of most interest to the users of science data.

Aside from keeping track of observations in the processing system, the database allows users to monitor the progress of their observations in AP, and it provides valuable CXC performance statistics for our administrators. The AP Status Database and the Processing Status Tool are described in detail in paper 4844-47. Figure 2 presents the flow of observational data from telemetry to data distribution and public release. The boxes with rounded corners indicate activities and processes, the boxes with the solid heavy borders are databases. For all practical purposes, VVstat, SAPstat, and CPstat may be considered part of APstat, the AP status database; Obstat is the status component of Ocat, the observation catalog. The flow involves the following steps:

1. The telemetry data (TLM) enters the AP pipeline system which incorporates three other subsystems: Special Automated Processing (SAP; when AP has trouble with the data and needs manual nudging), Custom Processing (CP; when users request special manual processing or when AP/SAP would take too long for any reason — usually problematic data), and batch reprocessing (RAP; wholesale reprocessing of large numbers of observations). CP has a special place among these since its data products are not ingested into the archive.

2. AP updates APstat and ingests data products as it progresses.

3. A separate process monitors whether ingest into the CDA is complete and updates APstat.

4. When processing of an observation is finished, APstat notifies V&V.

5. The resulting V&V report updates VVstat.

6. If the V&V status indicates that review is required, it goes to the Chandra Director's Office (CDO) which reports back to V&V.

7. Otherwise, if the data products are deemed distributable, they are "promoted", i.e., made the current default products. The opposite is "demotion", but this is executed later (see 15).

8. The quality flags in APstat are updated.

9. If distribution is warranted, Data Distribution is activated.

10. Data Distribution dates are updated in APstat.

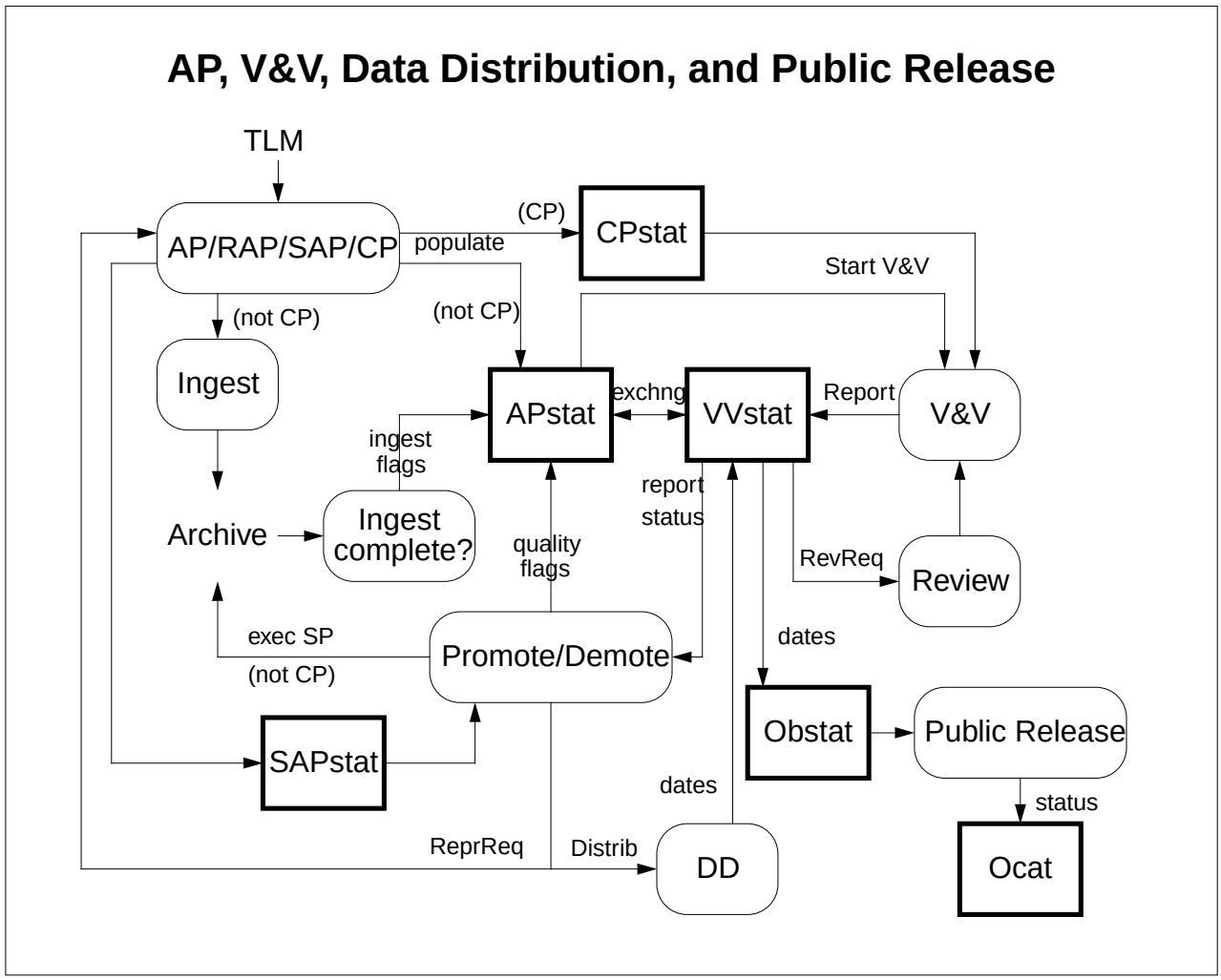

Figure 2. Data flow through Automated Processing, Verification & Validation, and Data Distribution

11. APstat updates Obstat, including the public release date.
12. The public release of observations operates autonomously on the basis of Obstat.
13. If the V&V report indicated that reprocessing is required, the observation goes back for SAP and/or CP.
14. SAP follows the same path as AP but additional information is kept in a separate database table, SAPstat, as well.
15. If the V&V report prohibited distribution, demotion of products occurs under the control of SAPstat, since it needs to be determined which products exactly are defective.
16. CP is kept track of in yet another table, CPstat; it also triggers V&V upon completion, VVstat and Obstat operate similarly to AP, but no data product ingest occurs and data distribution is taken from a separate directory.

17. In a related activity not included in the figure, the primary (scientific) data products of an observation are written to an anonymous ftp site when promotion occurs **and** the observation has been publicly released in the past, or when an observation is released for the first time.

We admit that this rather complicated system is in many ways very Chandra-specific. However, we want to convey four important notions. The first is that such a database is of vital importance to the processing operations. The second is that keeping careful track of all actions that pertain to data products and their disposition is immensely helpful for the integrity of the data archive, for providing operational personnel as well as users instant access to the status of all observations, and for providing administrators access to data for institutional performance statistics. The third is that when one does a careful analysis of the data flow through the various divisions, it is easy to spot deficiencies and see how these should be remedied; it certainly made us realize the pivotal role played by V&V. The fourth is that a processing status database should be kept from the start of operations. Even if the database structure needs to be revised at a later date, it is still easier to populate the new database from the contents of the old one. We started design and implementation 21 months after launch and full population took four months, even though we had kept careful records; we strongly recommend to start much earlier, preferably before launch or commission of the observatory.

As it stands, the processing status database is immensely helpful in aiding processing operations. In addition we have several cronjobs running that query this database and will sound warnings in a timely manner about developing problems, allowing the various groups involved to take pro-active action.

Finally, a word about time intervals. One should take note of the fact that different contexts require different definitions of observing time intervals. At this point, six different time intervals are associated with each Chandra observation:

1. Approved exposure time: the time allocated by the peer review (or, strictly speaking, by NASA HQ).
2. Scheduled time: the planned duration of the observation, in the mission schedule.
3. Administrative time: the length of period from end-of-slew till beginning-of-slew; this is used by NASA MSFC administration as a performance measure.
4. Observation interval: from mid-slew before the observation till mid-slew after the observation.
5. Exposure time: The amount of time during which useful science data were collected, i.e., the usual meaning of the word "exposure".
6. Charge time: the amount of time charged to the approved exposure time, in order to determine whether the observation is complete and the user has received his/her due. This is nominally 80% of the approved time, except in the case of short observations: pieces with less than 1000 s are dropped and observations less than 3000 s only get one shot. Normally, the charge time is equal to the exposure time, but there are exceptions, for instance when different CCD chips have wildly different exposure times due to dead time or telemetry saturation, or when a considerable amount of dead time was taken into consideration already in the allocation of approved time. As a result, the decision tree for charge time is fairly complicated.

All this information is, and needs to be, included in the Processing Status Database.

6. CONCLUSION

We believe that the CXC has built a streamlined and well-running data processing system and part of the quality and success of this system may be attributed to the two major databases that control these operations — the observation catalog and the automated processing status database. Having said that, we also are aware that we could have had a better system, and especially a better system earlier in the mission, if four years ago we had had our current experience. This is not just a 20/20 hindsight statement: this experience is transferable and we hope that this description may aid future missions and observatory projects to perform even better. We may be reached at arcops@head-cfa.harvard.edu.

ACKNOWLEDGMENTS

The authors would like to thank CXC Data Systems, and in particular Panagoula Zografou, Kimberly DuPrie, Alesha Estes, Padmanabhan Ramadurai, Edward Mattison, Adam Dobrzycki, Joy Nichols, Eric Schlegel, Dong-Woo Kim, Ian Evans, and Janet DePonte for numerous discussions related to this project; we also would like to thank Michael Preciado and Emily Blecksmith for their part in the project of backfilling the database, and Diane Hall for writing the Processing Status Tool. This work is supported by NASA contract NAS 8-39073 (CXC).