Tracking the processing status of Chandra observations

Sherry L. Winkelman[a],
Arnold Rots[a],
Stephane Paltani[a],
Diane Hall[b]

[a]Harvard-Smithsonian Center for Astrophysics, 60 Garden St., Cambridge, MA (USA)
[b]TRW, Cambridge, MA (USA)

ABSTRACT

The Chandra Data Archive (CDA) has been archiving and distributing data for the Chandra X-ray Observatory (CXO) and keeping observers informed of the status of their observations since shortly after launch in July 1999. Due to the complicated processing history of Chandra data, it became apparent that a database was needed to track this history on an observation by observation basis. The result is the Processing Status Database and the Chandra Observations Processing Status tool (Status Tool). In this poster, a description of the database design is given, followed by details of the tools which populate and display the database.

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

1. INTRODUCTION

Data from the Chandra X-ray Observatory can have a very complicated processing history. Keeping track of this history is essential for delivering complete data sets to the end users as well as providing up to the minute information on the status of an observation. The Processing Status Database and tools developed by the Chandra Data Archive Operations team track the processing history and current status of all data received from the CXO. The database and associated tools provide essential links between telemetry receipt, processing and reprocessing of data, verification and validation (V&V) of data, data distribution, and the release of public data.

Data are processed through a series of pipelines with each pipeline generally requiring the products from the preceding level. Each run of a pipeline results in a new version of the data products for that level. Furthermore, data can be reprocessed starting from any level. This can result in a final data set created from a complex history of pipelines run and data product versions at various processing levels. The Processing Status Database provides the means of tying together the pipelines and data versions at each level of processing that comprise a complete set of data products. Information from the database is displayed through a web interface which presents processing status information on an observation basis and allows the user to select standardized formats of the data or customize the output. The result is a system which provides the user a complete view of the status and entire processing history of each observation, with access to V&V reports and the processing issues database.

Send correspondence to SW.; E-mail: swinkelman@cfa.harvard.edu

2. BACKGROUND

The need for information regarding telemetry receipt and processing status of data developed early in the mission. Not only was this information necessary for planning/replanning observations, but the CXO felt it was important that observers be kept informed of what was happening with their data in a timely matter and that they receive their data as soon as possible after an observation was taken. To facilitate this, an RDB table was generated daily of all data received through telemetry receipt. Further information about the processing of the data was then added to the table through scripts. This table was then processed with yet another script which added additional columns from CDA database tables. Finally, the newly formed table was dumped into a database table which could be accessed through a rudimentary web tool. In addition, new lines added to the table triggered notification e-mails to the observer, letting them know that data had been received on their observation. This was a cumbersome process which required scripts to be run by several different users in sequence on different LANS and then produced only a daily snap-shot, not up-to-the-minute information.

This system worked reasonably well until the CXO started reprocessing the data. The CDA was designed to include all versions of the processed data. It was expected that at various times of the mission, the data would be reprocessed using refined pipelines and better calibration data. Since all data is stored in the CDA, some method of tracking the latest and greatest data was needed. The first attempt was to create a table, similar to the automatic processing table described above, but this new table would track the reprocessed versions of the data. This system had major short comings and it became apparent that a new tracking system was needed.

The new system needed to track data from telemetry receipt, through all forms of processing and tie together complete sets of data. It also needed to tie-in with the V&V tracking to aid in data distribution and special reprocessing of data. The result is the Processing Status Database, its associated tools, and the Status Tool.

3. THE DATABASE

This new database, *axafapstat*, uses the processing pipelines as its source of meta-data and tracks data versions by obsid, observation interval (obi), and alternating exposure mode for ACIS observations. It was designed to map the pipelines as they are run in succession, starting with telemetry and then jumping to the level 0.5 science* products and beyond. The intermediate and engineering products are not tracked, as they are not created on an obsid/obi basis and are not distributed to users due to proprietary issues with the data. This design optimizes the population of the tables, but not necessisarily the querying of the database. To improve the performance of the Status Tool, an additional table which points to the latest pipeline run for an obsid/obi was implemented. In addition to the *axafapstat* tables, there are a number of independent tables which have been linked through the Status Tool to provide a complete history of the disposition of Chandra data.

3.1. *axafapstat*

For purposes of data packaging, a given set of data begins with the level 2 version of the data products. From this, a complete set of lower products can be determined from the database. In addition to the tables which track the data for packaging purposes, there are tables which reference the pipeline logs for each pipeline run. These are used by the data operations group for reprocessing data. Those tables are: `science_2_log`; `science_1_5_log`; `science_1_log`; `science_0_5_log`; `aspect_1_log`; and `obidet_0_5_log`. There is a one-to-one mapping between these pipelog tables and their corresponding data tables.

The top level table is `science_2`, which is unique by obsid and revision†. Because level 2 data is the result of merging all obi's of an obsid into a single event list, an additional table, `science_2_obi` is necessary to link all lower products based on obi and alternating exposure mode‡ to the level 2 data. Each entry in `science_2_obi` is linked to one entry in `science_1`. If the obsid is a grating observation, each entry for that obsid is also linked to one entry in `science_1_5`.

All observations that are processed to level 2 have a record in `science_1` which is unique by obsid, obi, alternating exposure mode and revision. For grating observations the data is processed through level 1.5 before

*The term science is used to generically refer to data from ACIS or HRC.

†Data version is represented by revision in the tables

‡Alternating exposure mode is represented by `alt_exp_mode` in the tables

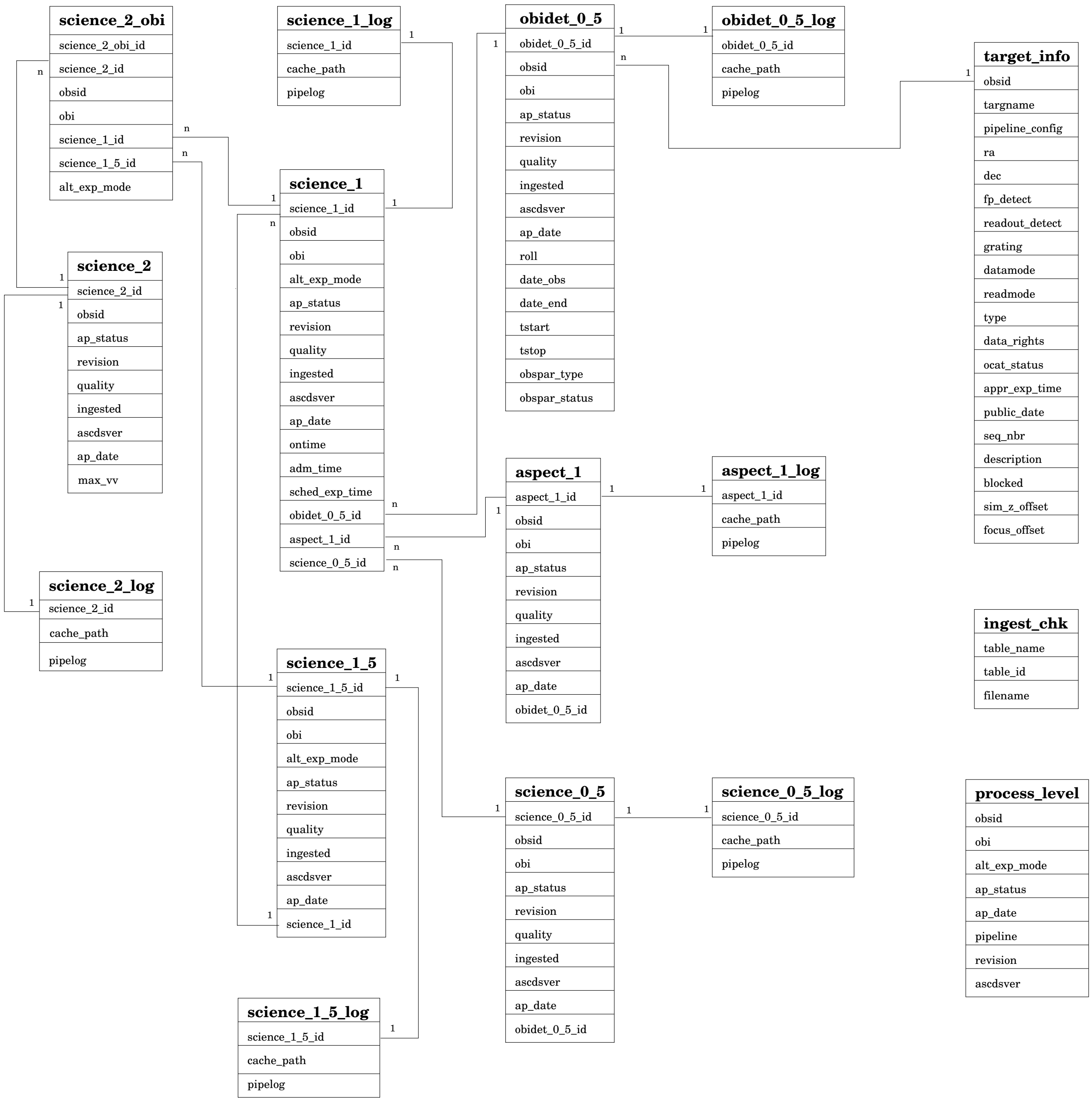

Figure 1. Layout for the axafapstat database.

getting to level 2. This data is stored in the `science_1_5` table, unique by obsid, obi, alternating exposure mode, and revision. Each record in science_1_5 points to a record in science_1. Each record in `science_1` points to a record in `obidet_0_5` and possibly `science_0_5` (for HRC and ACIS interleaved mode only) or `aspect_1` .

All observations that are processed to level 1 have a record in `obidet_0_5`, unique by obsid, obi, revision. For HRC and ACIS interleaved mode observations the data is processed through science level 0.5 pipelines before being processed through level 1. This data is recorded in `science_0_5`, unique by obsid, obi, revision. Each record in `science_0_5` points to a record in `obidet_0_5`. Also, many observations are processed through the aspect level 1 pipeline before proceeding to level 1. That data is recorded in the `aspect_1` table, unique by obsid, obi, revision. Each record in `aspect_1` points to a record in `obidet_0_5`.

Every obsid for which telemetry is received is in `target_info`, unique by obsid. This table contains the parameters that remain the same regardless of the obi or alternating exposure mode.

Then there is `process_level`. This table is populated by triggers on `obidet_0_5`, `aspect_1`, `science_0_5`, `science_1`, and `science_2`. It is unique by obsid, obi, and alternating exposure mode. It is used by the Status Tool to determine the latest, highest level of processing.

Finally, there is `ingest_chk`. This table is used as a temporary table for determining whether all of the data products associated with an entry in `science_2`, `science_1_5`, `science_1`, `science_0_5`, or `obidet_0_5` have been ingested.

3.2. Related Tables

As mentioned earlier, there are a number of tables which are linked to *axafapstat* through the Status Tool. These tables round out tracking by including information about V&V, custom processing (CP), and processing issues.

The V&V tables track the validation and verification of each observation that has been processed to level 2. The V&V tables are linked directly to `science_2` through the `max_vv` column. This column points to the latest V&V entry for that set of data.

Generally, only data that has been processed through automated processing is available to outside users as this is the only data that is sent to the CDA. On occasion, however, the data needs to be processed manually before it can finish through the automated processing. This data is V&V'ed and recorded in the V&V tables as well. CP data is not sent to the archive, so the V&V tables have a link to `cpstat`, a table which contains relevant information about the manual processing used to produce the data set.

The other set of tables that the Status Tool links to *axafapstat* are the issues tables. These tables are used to track processing and data issues that arise with various observations. Obsid is the link between the issues tables and *axafapstat*.

4. POPULATION OF THE DATABASE

The population of *axafapstat* begins with the automated processing pipelines. As the various pipelines run, the finished products are put into a data cache and the filenames are put into an ingest[§] queue which is used to submit the files to the archive. Similarly, a queue is created which provides the information necessary to populate the tables in *axafapstat*. The queues for *axfapstat* are placed in predetermined directories and are picked up by Perl scripts which connect to the database and insert or update entries as directed. Several other scripts are run which update some of the columns after certain events have occurred.

4.1. Population from queues by `darch2apstat`

Each pipeline sends a line of data to the queue which contains information about the inputs to, as well as the products from, that pipeline. The various pieces of data are written in a `keyword=value` format and are separated by semi-colons; each line ends in a `|`. The information from a typical science 1 pipeline is shown in Table **??**.

[§] Ingest is the term used for putting data into the archive.

Table 1. Keyword/value pairs from a typical science 1 pipeline

Keyword	Value
operation	apstatus
pipe	SI1
pipelog	/dsops/ap/sdp.8/cache/2002_07_22/acis/acis_f_l1_f03488_000N001.log
obsid	03488
version	001
obi	000
apstatus	DONE
release	6.8.0
apdate	2002-07-23T22:27:08
ontime	7.1720001068711E+03
admintime	7.1760252949893E+03
versionOD0.5	001
versionAS1	001
archfiles	acis_f_l1_f03488_000N001.log, acisf03488_000N001_aoff1.fits, acisf03488_000N001_bpix1.fits, acisf03488_000N001_evt1.fits, acisf03488_000N001_flt1.fits, acisf03488_000N001_msk1.fits, acisf03488_000N001_mtl1.fits, acisf03488_000N001_soff1.fits, acisf03488_000N001_stat1.fits

The work horse of the *axafapstat* database is the script `darch2apstat`[¶]. The script parses the darch queue, performs some consistency checks, and either inserts a new row or updates an existing row if all checks are passed. If a consistency check fails, the offending line from the darch queue is saved to an error file. All sessions are logged, so once an issue is fixed, the file can be re-submitted to `darch2apstat` and the insert or edit can proceed.

The most important consistency check is that the tables be populated in the proper sequence for a given set of data. Taking our sample data in Table ??, if there were no entry in `obidet_0_5` for obsid 3488, obi 0, revision 1, the insert into `science_1` would fail. That line of the queue would be copied to an error file and the other entries in the queue would be processed. The production of an error file results in a notification to the archive operations team so the source of the inconsistency can be determined and resolved quickly. Once a proper entry is made in `obidet_0_5`, `darch2apstat` can make the insert to `science_1` using the error file.

In addition to these consistency checks, `darch2apstat` also gets some information from other locations. This information is not available directly from the pipelines, but is useful for users to have and is not readily available from other sources. When entries are made to the `target_info` or `science_1` tables, some of the columns are filled in by querying tables in other databases. In the case of the obidet pipeline, certain values are retrieved from the observation parameter file which is given in the `archfiles` list.

4.2. Updates by other scripts

Several columns can only be populated after certain events occur and additional processing is performed. For example, the `ingested` column in the tables is a flag which is set to "Y" when all files from that pipe have been ingested into the archive. Ingestion of data is done independently from population of *axafapstat*. Each line in a darch queue has an `archfiles` keyword and a list of files produced by that pipe. A row is inserted into the

[¶] Darch is the system which produces the queues from the pipelines.

ingest_chk table for each file in the list. A separate script then checks to see if the files have been archived. If a file is found in the archive, it is removed from ingest_chk. Then, for every entry in a pipeline table where ingested is "N", ingest_chk is checked to see if any entries remain. If no entries remain, the flag is set to "Y".

Similarly, the quality column is updated whenever a V&V reported is submitted for a set of data. When an entry is made into a table, the quality is set to "pending". Once the data have been V&V'ed, the quality is set to "default" if it passed and "rejected" if it failed. There can be only one default version of the data at a time, so if a later version is updated to "default", the previous default version is changed to "superseded". These updates are done by a script which is called when a V&V report is entered into the V&V tables mentioned in Section ??. The max_vv column in science_2 is also updated when an entry is made into the V&V tables.

Finally, there are a number of columns in target_info which are copied from tables in other databases. They are duplicated here to improve performance of the Status Tool. Since those other columns may be updated, a script is run periodically to synch the target_info table with those other tables.

5. THE STATUS TOOL

The Processing Status Tool is a web-based interface to *axafapstat* and several related tables. It enables users to create customized reports on the processing status of Chandra observations. The physical format of the reports may be regular HTML, tab-delimited ASCII, or RDB. The content format of the reports may be chosen from a set of standard reports or may be arbitrarily customized. Where appropriate, links are made to V&V reports, if they exist. The observations to be included in the reports may be selected by specifying selection criteria for up to nine particular fields. The user may also order (sort) observations in the report by particular fields. One can use up to three ordering criteria.

Table 2. Descriptions of the various standard reports available through the Processing Status Tool

Report Type	Description
Short	See Table ?? for listing of columns
Long	See Table ?? for listing of columns
Dates	See Table ?? for listing of columns
Summary	See Table ?? for listing of columns
Scheduled	Observations with status scheduled but no processing information received
AP/no L2	Observations with status scheduled by no level 2 product has been processed yet
Awaiting Ingest	Level 2 products that have incomplete archive ingest
Awaiting V&V	Level 2 product has been ingested in archive but has no corresponding V&V report. If a later level 2 product has a V&V report, then the observation is not considered awaiting V&V.
Awaiting CDO Review	Level 2 products that have the CDO flag set in the V&V report
Awaiting Distribution	Level 2 product with V&V status of *OK* or *ReprReq+Dist*, in not awaiting CDO review, but has no distribution date
SAP	Observations awaiting, or in, Special Automated Processing (SAP); highest revision of a Level 2 product and V&V status of *ReprReq+Dist* or *ReprReqNoDist*

There are four standard reports that a user may choose from: short, long, summary, or dates report. There seven are additional reports available to CXC internal users to aid in the processing and dissemination of data. The report types are summarized in Table ??. Table ?? shows the columns that are presented in the four standard reports available through the Status tool.

ACKNOWLEDGMENTS

The authors would like to thank CXC Data Systems for making Diane Hall available for this project; Dave Plummer for creating the darch queue tools; and Ed Mattison, Janet Deponte, and Ian Evans for numerous discussions related to this project. This work is supported by NASA contract NAS 8-39073 (CXC).

Table 3. Fields available in the four standard reports

Field name	Description	Long	Short	Dates	Summary
PI First	First Name				X
Last Name	Last Name				X
Proposal	Proposal Id number				X
ObsId	Observation Id number (assigned by CXCDS)	X	X	X	X
SeqNum	Observation sequence number (assigned by CXCDS)	X	X		X
Object	Object name	X	X	X	X
RA	Right Ascension	X	X		
Dec	Declination	X	X		
Roll	Roll angle	X			
Instrument	Readout instrument/detector	X	X		
Grating	Transmission grating inserted	X	X		
DataMode	Data mode ACIS: RAW; FAINT; FAINT_BIAS; VFAINT; GRADED; CC33_FAINT; CC33_GRADED HRC: OBSERVING; NEXT_IN_LINE	X			
ReadMode	Read-out mode (ACIS only)	X			
Type	TIMED or CONTINUOUS Type of observation GO: Guest Observer GTO: Guaranteed Time Observation TOO: Target of Opportunity (approved by peer review) DDT: Director's Discretionary Time CAL: Calibration observation ER: Engineering Request	X	X	X	X
Data Rights	Data rights: S: Standard proprietary (12 months) D: Discretionary (3 months) N: None (public immediately) 1 month 2 months 3 months 4 months 6 months 9 months	X			
Obs Date	Start date and time of observation	X	X	X	X
OCat Status	Status of the observation in the archive unobserved, scheduled, partially observed, observed, archived, discarded, canceled	X			
Public Rel Date	Date when data products become (became) publicly available; NULL means unknown	X	X		X
Apprvd Time	Approved exposure time (ks)	X		X	X
Sched Time	Scheduled time (ks)			X	

Field name	Description	Long	Short	Dates	Summary
On Target Time	Total time spent on-target (ks)			X	
Good Time	Good data exposure time (ks)	X		X	X
Charge Time	Time charged against allocation (ks)	X			
Pipe. Config.	Pipeline configuration used for processing	X			
	primary: normal configuration for observations				
	secondary: limited configuration for observation using secondary instrument (i.e., instrument not in focal plane)				
	primary_obc: limited configuration using on board aspect				
	standard: limited calibration configuration (processing only through L0)				
AP/CP	Automated processing or custom processing	X	X	X	
Proc Version	Version of processing of observation	X		X	
Proc Status	Processing status	X	X	X	
Proc Date	Date of processing operation	X		X	
Release	version of CXCDS software used for processing	X			
Issue(s)	Issues that apply to processing of observation. Issues shown in prentheses () are closed.	X	X	X	X
Ingested	Ingested status of data products	X			
V&V Version	Version of Verification & Validation of processing	X		X	
V&V Status	Verification & Validation status of Automated Processing data products	X	X	X	
	NULL: No V&V Report available yet				
	OK: Approved for distribution				
	Hold: Internal DS hold				
	Problem: Referred to CDO				
	Repro+Dist: Reprocessing required but distribution approved				
	ReprNoDist: Reprocessing required; do not distribute				
CDO Review	Review by Chandra Director's Office required	X	X	X	
V&V Date	Date of Verification & Validation	X	X	X	
Data Distr Date	Date PI was notified of data being available	X	X	X	X
Data Mail Date	Date hard medium was mailed	X		X	