Astronomical Data Analysis Software and Systems XVII **P4.4**
ASP Conference Series, Vol. XXX, 2008
J. Lewis, R. Argyle, P. Bunclarck, D. Evans, and E. Gonzales-Solares, eds.

Chandra Source Catalog Quality Assurance

Ian N. Evans, Janet D. Evans, Kenny J. Glotfelty, Diane M. Hall, David Plummer, and Panagoula Zografou

Smithsonian Astrophysical Observatory, 60 Garden Street, MS-81, Cambridge, MA 02138, USA

Abstract. The Chandra Source Catalog will be the definitive catalog of all X-ray sources detected by the Chandra X-ray Observatory. Automated detection of data processing anomalies and assurance of data product quality is essential because of the large data volume that will be generated over a period of a few months during each catalog production run. Quality assurance analysis is performed at the completion of each stage in catalog pipeline processing, so that any issues can be identified and corrected before they can affect downstream processing. Pipeline processing errors are detected automatically, and automated comparison of key diagnostic outputs with predefined validity criteria identifies potential data quality issues. The most common data quality issues are resolved without any human intervention. Cases where interactive review may be required are identified, assessed, and bundled together for efficient analysis and repair via a web-based graphical user interface.

1. Introduction

The Chandra Source Catalog (CSC) will be the definitive catalog of all X-ray sources detected by the Chandra X-ray Observatory. The catalog will include fields of all Galactic latitudes, and sources from the entire detector field of view. For each detected X-ray source, the catalog will list the source position and a detailed set of source properties that will include aperture and model fluxes in multiple bands to construct X-ray colors, source extent estimates, and spectral fits for bright sources. In addition to these traditional catalog elements, file-based data products, including images, photon event lists, light curves, and spectra, will be included for each source individually.

The scientific integrity of the catalog depends in part on a set of quality assurance (QA) analyses that are performed at the completion of each stage of catalog pipeline processing, so that any issues can be identified and corrected before they can affect downstream pipelines. Data processing anomalies, data issues, and data product quality are verified. The data volume generated during each catalog production run is significant. The first catalog release, incorporating approximately 7 years of Chandra observations, is expected to include of order 180,000 distinct X-ray sources, $\sim$6 million file-based data products, and $\sim$40 million databased quantities.

The QA mechanisms have been designed to detect pipeline processing errors automatically, and identify potential data quality issues by comparing key diagnostic outputs with predefined standards. The most common data quality issues can be resolved without any human intervention.

2. Pipeline Quality Assurance

Every catalog processing pipeline that is run is subject to pipeline QA, which compares a subset of the output products (e.g., scalar data values, file-based data products, and processing log files) from the pipeline with predefined standards. Each standard that the comparison indicates is violated is flagged, and will trigger one or more predefined actions at the completion of QA processing for the pipeline. Multiple standards may be violated simultaneously, so the predefined actions are not executed until all of the evaluations are completed.

In some cases, interactive human review is necessary to assess whether a violation of the standards actually constitutes a problem, and whether the corresponding processing thread should be terminated. For example, if too many source regions are detected within a small spatial area, then human review is required to evaluate whether all of the detected source regions are real.

Unexpected pipeline warnings or errors (which are evaluated by a parser that scans the pipeline log files) always require human review to determine what happened and how to proceed. In such cases, the typical response is to terminate the current processing thread, perform any needed repairs, and initiate reprocessing of the thread.

If none of the violated standards requires human review, then the predefined automated actions are triggered. These actions typically result in termination of the processing thread for a subset of the input data. For example, if a detected source region is significantly smaller than the point spread function at the location of the source, then that detection is deemed to be an artifact and the processing thread for the source region is terminated.

3. Snapshot Quality Assurance

Snapshot QA creates comprehensive sets of products for human review from a randomly selected subset of the catalog processing pipelines. These reviews supports statistical monitoring of catalog processing and consider the very real possibility that unexpected conditions may exist that are not evaluated adequately by pipeline QA. Since the outcome of snapshot QA has no impact on any current processing threads, completion of the human review is not required for those thread to continue. Snapshot QA results in procedural and operational feedback. If the reviews identify a significant issue, then modifications to the processing pipelines and reprocessing may be required to correct the problem. Snapshot and pipeline QA are independent, and each may be performed for a pipeline irrespective of whether the other was performed for that pipeline or not.

4. Graphical User Interface

All human reviews are performed via a graphical user interface (GUI) that is modeled on the existing Chandra Validation and Verification GUI (Evans et al. 2007). The catalog QA GUI presents the items to be reviewed as a group of linked pages that includes images, plots, and tabular data appropriate for assessing the flagged conditions. The QA scientist is led page by page to review the relevant data associated with each of the standards violations. Depending

on the results of the review, the scientist may choose to accept the bad status identified by pipeline QA, or may reset the status to be good.

The GUI allows the scientist to modify directly a limited subset of key data products output from the processing pipelines that detect X-ray sources and that merge multiple observations that include the same source. Experimentation indicates that manual intervention may sometimes be required to evaluate poor source detections or source cross-matching from these pipelines at a frequency that, while still low, is high enough to warrant developing a flexible user interface to minimize the workload imposed on the QA scientist.

4.1. Detected Sources Interactive Live Display

If certain source detection pipeline QA standards (mostly related to local or global source density) are violated, then the set of detected source regions is presented to the QA scientist for review.

The calibrated full field image is displayed using the ds9 imager, with the source regions overlayed. Also displayed is a catalog listing key properties and detection information for each source region, as well as the results of pipeline QA. The QA scientist can select one or more detected source regions, and manually flag them for inclusion or exclusion from further processing, using a custom TCL interface to ds9. The source region position and ellipse parameters can be modified by interactively manipulating the source region in ds9, or by editing the numerical values directly in the catalog display. Immediate visual feedback of any changes is provided. Once the QA scientist is satisfied with the revisions, they can be submitted, and subsequent pipeline processing with the updated list of source regions is initiated automatically. Any modified source regions will automatically be flagged as such in the catalog.

4.2. Master Sources Interactive Live Display

The master pipeline attempts to crossmatch source regions detected in multiple observations that cover the same region of the sky, to ensure that the catalog includes only a single entry for each physical source. This is complicated because the shape of the Chandra point spread function varies significantly over the field of view and with X-ray energy. Under some circumstances the algorithms used cannot reliably resolve the matches, and human judgement must be applied.

The detected source regions for all of the observations are presented overlayed on the calibrated images for each of the observations separately, together with crossmatches identified by the master pipeline. The QA scientist can then interactively identify both unambiguous and ambiguous matches between source regions detected in different observations (the latter may result because a single source that is detected off-axis in one observation may be resolved into two or more distinct sources detected on-axis in another observation).

5. Quality Assurance Data Flow

Every catalog processing pipeline executes the QA comparisons as its last steps, and populates corresponding QA flags that record any standards violations or other issues (see Fig. 1). After each pipeline completes processing, the CSC

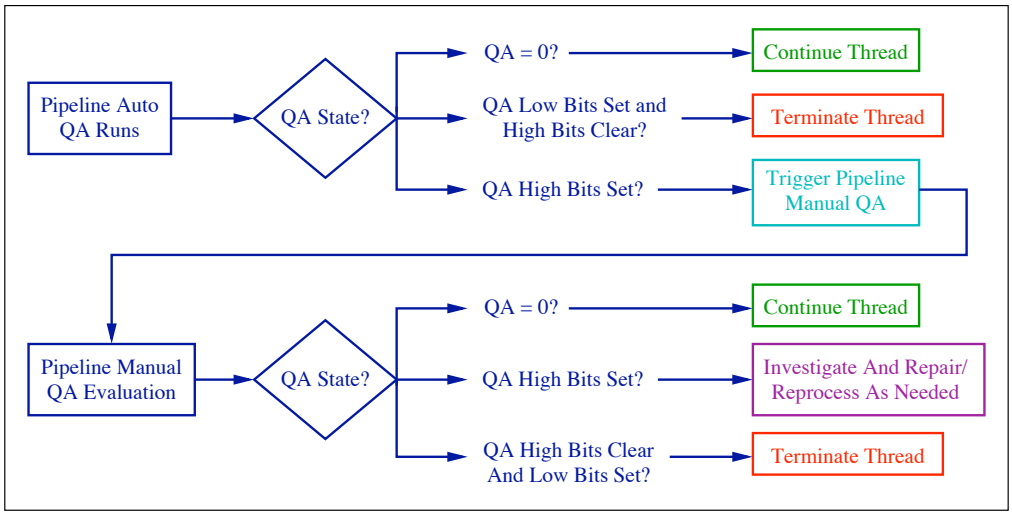

Figure 1. Quality Assurance Data Flow

automatic data processing (AP) infrastructure (Grier Jr. et al. 2007) checks the status of the flags. If no violations were detected, then AP continues the processing thread as normal. If human review is required, then AP passes the pipeline QA output to the GUI manager and blocks further processing of the thread until the review is completed. Otherwise AP terminates the processing thread and cleans up. The GUI manager is an autonomous process that notifies the QA scientist when a human review is required. After the review is complete, the GUI manager notifies AP, which again checks the status of the QA flags and then either continues or terminates the thread as appropriate.

One complication that is accounted for as part of the data flow is that although the source detection pipeline operates on a complete observation at a time, the QA step validates each detected source region separately, since some source regions may violate the predefined standards while others do not. If any of the detected source regions require human review, then the review is triggered for the pipeline and all source regions as a unit. This approach allows the reviewer to modify the QA flags for each detected source region as needed, considering the entire set of regions as an ensemble.

Acknowledgments. Support for development of the Chandra Source Catalog is provided by the National Aeronautics and Space Administration through the Chandra X-ray Center, which is operated by the Smithsonian Astrophysical Observatory for and on behalf of the National Aeronautics and Space Administration under contract NAS 8-03060.

References

Evans, J., et al. 2007, in ASP Conf. Ser. 145, ADASS VII, ed. R. Albrecht, R. N. Hook, & H. A. Bushouse (San Francisco: ASP), 555

Grier Jr., J., et al. 2007, in ASP Conf. Ser. 145, ADASS VII, ed. R. Albrecht, R. N. Hook, & H. A. Bushouse (San Francisco: ASP), 81