Astronomical Data Analysis Software and Systems XVI
ASP Conference Series, Vol. 376, 2007
R. A. Shaw, F. Hill and D. J. Bell, eds.

The *Chandra* Source Catalog

G. Fabbiano, I. Evans, J. Evans, K. Glotfelty, R. Hain, M. McCollough, F. Primini, A. Rots, T. Calderwood, S. Doe, J. Grier, P. Harbo, M. Karovska, J. McDowell, D. Plummer, L. Paton, M. Tibbetts, D. Van Stone, P. Zografou

Chandra X-ray Center, Smithsonian Astrophysical Observatory, Center for Astrophysics, Cambridge, MA, USA

Abstract. The *Chandra* X-ray Center (CXC) has undertaken the production of the *Chandra* Source Catalog, a data-mining project that makes use of the continuously growing *Chandra* public archive. The distinguishing characteristic of *Chandra*, the NASA Great Observatory for exploring the universe in the X-rays, is its sub-arcsecond resolution, which provides the most sensitive and detailed view of the ∼(0.3–7) keV sky presently attainable. The Catalog will characterize the X-ray sky at high resolution and with accurate astrometry, making use of all the imaging *Chandra* data. It will provide a uniform reduction of the *Chandra* archive, that will be a major interface for the Virtual Observatory; will be continuously updated as more data enters the public domain during the on-going *Chandra* mission; and will enable a dynamic interaction to produce user-configured views with on-the-fly analysis work-flows. In this paper we describe the Catalog, and the software and hardware needed for its realization.

1. Introduction

The *Chandra* Source Catalog is a data mining project, based on the uniform (Level 3) processing of the ever increasing *Chandra* public archive data. *Chandra*, the NASA great observatory for X-ray astronomy (Weisskopf et al. 2000), was launched on 1999 July 23, and its mission is on-going. The distinguishing characteristic of *Chandra* is its sub-arcsecond resolution, which provides the most sensitive and detailed view of the ∼(0.3-7) keV sky presently attainable. There are two imaging detectors on board *Chandra*: the High Resolution Camera (HRC), which provides the highest resolution images but does not have energy resolution; and the Advanced CCD Imaging Spectrometer (ACIS), which has pixels slightly larger than the mirror angular response, but is more sensitive than the HRC and provides information on the energy of the incoming photons. Two systems of gratings, used in conjunction with either the HRC or ACIS, provide high resolution X-ray spectra. An optical CCD camera is used to image the star-field for each *Chandra* pointing, so that accurate coordinates (of order $0\farcs4$ (1σ) absolute astrometry[1]) can be recovered in post-observation processing of the data. Most of the *Chandra* data so far have been obtained with ACIS imaging. As a default all the *Chandra* data are publicly available after a one-

[1]http://cxc.harvard.edu/cal/ASPECT/

year proprietary period for the original observer; some data are immediately released to the public.

In this paper we present a new project, aimed at producing a catalog of all the sources detected in *Chandra* public imaging data. We first describe the characteristics of the *Chandra* data, we then discuss the characteristics of the Catalog, the software that is being developed to build this catalog, the hardware set-up for the data processing, and our projected schedule. This project has been reviewed and approved by an ad-hoc board, composed by research and data scientists, convened by the CXC Director.

2. *Chandra* Data

The *Chandra* data are 4-dimensional: each individual detected photon is tagged with coordinates in the detector plane, energy, and time. These data are saved in FITS format event lists. The *Chandra* telemetry is processed at the *Chandra* X-ray Center (CXC), by a system of software pipelines, resulting in well-defined data products for distribution to observers and archiving (see Evans et al. 2006). The processing pipelines and the resulting data products are presently organized in well defined *Levels*, ranging from Level 0 (processing of telemetry), to Level 1 (derivation of coordinate correction, application of calibrations to each individual data pointing), and Level 2 (screened and corrected observation-based data, which may include several pointings). These products are all screened before release by the operations team, using the Verification and Validation software. The data products preserve their processing history in their headers, and include a "version" value that is used for archival and reprocessing purposes. Typically processed *Chandra* data products are distributed to the observers (and archived) within a couple of days from the receipt of telemetry on the ground. Periodically, when a significant number of calibration or software upgrades have become available, the CXC undertakes a reprocessing of the entire *Chandra* archive. We are presently undertaking the third such reprocessing.

The *Chandra* Data Archive (CDA) is a system of relational databases and a data warehouse, where all the public and proprietary data products are saved and made available to data operations and users. To date, the archive contains approximately 5 TB of data.

3. Catalog Characteristics

The aim of the Catalog project is to perform a uniform Level 3 processing of the *Chandra* data. This processing will consist of the detection of sources in all the public imaging data, the extraction of source-based properties, the estimation of detection thresholds and limits in any part of the sky observed with *Chandra*, and the derivation of new source-based data products, for both astronomical data and calibrations. To achieve this, the CXC has been faced with solving difficult technical and scientific problems, such as modeling the wide-field background to remove instrumental effects, the choice of the most effective algorithms for measuring source variability and extent, the approach to merging observations with different off-axis angles (and therefore different angular resolution), and the determination of scientific analysis paths to characterize the

catalog; moreover, science user cases have been assembled and are being analyzed to set scientific-use requirements on the user interface to the catalog. From a software point of view, the CXC must extend its pipeline processing software and archive infrastructure to reflect the new catalog scientific requirements, develop an archive interface to access catalog information and data products for user analysis and workflows, and extend the CIAO[2] data analysis software to provide tools for archival analysis.

The scientific goals of this project include:

- Characterize the X-ray sky at high angular resolution, including all the areas observed with the *Chandra* HRC and ACIS in imaging mode; these include both observations of the Galactic plane and observations of extragalactic targets at high galactic latitude. Produce source detections or upper limits for all these areas, and for each detected source determine accurate positions, X-ray colors and spectra (if enough photons have been detected), an estimate of time variability, and an estimate of the spatial extent of the source.
- Provide a uniform reduction of the *Chandra* archive, as a standard way for accessing the source information and cross-matching with other catalogs (e.g. in the NED, SIMBAD, and SDSS catalogs). We anticipate that this will be a major interface with the Virtual Observatory[3] (VO).
- Create a dynamic catalog that will be continuously updated as more and more data are entered in the public archive during the lifetime of the *Chandra* mission.
- Enable a dynamic interaction to produce user-configured views of the catalog with on-the-fly analysis work-flows.

The measure of success for this catalog will be the resulting science projects. This is not a catalog designed with a particular limited science objective in mind, but rather to give access to a wide range of user-defined projects. A number of use cases have been developed by CXC scientists to guide the catalog design and development. We want the catalog to stimulate many more unanticipated investigations.

We have estimated the sky coverage and expected source number, using the *Chandra* fields observed between 1999 and 2005. During these six years, 3,200 fields were observed, corresponding to 160 deg^2. If we extrapolate this estimate to 2015, we anticipate that the final catalog will contain of order 10,000 fields and cover 400-500 deg^2, or 1% of the entire sky. Scaling from a representative sample of ACIS fields observed to date, we expect to detect of order 400,000 sources by 2015. If we restrict ourselves to the high galactic latitudes (above or below 20°), the sensitivity of this catalog will reach 1×10^{-14}erg cm^{-2} s^{-1} in 34 deg^2 of the sky, and 2×10^{-15}erg cm^{-2} s^{-1} in 7 deg^2. The unique aspect of this catalog is given by the ability of *Chandra* to resolve confused regions and to detect very faint sources, because of the virtual lack of background contamination, resulting from the small source detection areas.

---

[2]http://cxc.harvard.edu/ciao/

[3]http://www.us-vo.org/

4. Software Approach

4.1. Pipelines

The *Chandra* Level 3 pipelines are based on the existing Level 0–2 pipeline infrastructure and CIAO tools. There are three main Level 3 pipelines: the *Detect* pipeline, resulting in a list of source detections; the *Per-Source* pipeline, where source properties are calculated; and the *Merge* pipeline, where information on a given source from different observations is merged. Each of these pipelines consists of different steps (or sub-pipelines, running CIAO tools). The *Detect* and *Per-Source* pipelines exist in prototype form; the *Merge* pipeline is being designed.

Detect consists of four main steps:

- *Calibrate and Clean*, where the data are cleaned of bad pixels, calibrations are applied, and high background events screened out.
- *Make Observations Products*, where the footprints of the observations are calculated, as are the exposure and background maps and the aspect histograms to correct for the motion of the spacecraft.
- *Detect Sources*, where exposure and background maps are tailored for the field in pre-determined energy bands, and sources are detected in these energy bands, using the CIAO *wavdetect* program.
- *Combine Detections*, where detections in different bands are associated with the same source.

Per-Source consists of five main steps for processing each detected source:

- *Extract Source and Background Events* for each detected source.
- *Run Spatial Analysis*, to compare the detected source count distribution with that expected from a point source at the source location, and estimate the angular extent of the emission.
- *Do Timing Analysis*, to extract the light curve of the source and estimate the time variability characteristics.
- *Extract Spectrum and Fit*, to constrain the spectral properties of the source.
- *Compute Source Properties*, to extract the source properties in pre-defined energy bands.

Merge consists of three main steps:

- Identify detections in different observations that refer to the same physical source. This is not a trivial task, since the pattern of sources may vary between observations because of time variability; moreover, a single detection off-axis (where the PSF is much larger), may be resolved into a number of sources on-axis.
- Determine the best parameter values for these sources.
- Create references between merged sources and per-observation sources.

5. Level 3 Archive

5.1. Databases and Data Objects

The *Chandra* archive architecture and infrastructure are being extended to support catalog activities (see Zografou et al. 2007). The new additions to the

archive include: the *Observation Database*, the *Source-by-Observation Database*, and the *Master Source Database*, where information about the observations, the results of observation processing, and the results of merge processing are tabulated, together with versioning information and links to the data products, stored in the new L3 Data Object Archive. Moreover, a new *Source Catalog User Interface* is being developed to support catalog browsing, retrieval and analysis.

Sources in the *Master Source Database* will be named using the acronym CXO, registered with the IAU Nomenclature Clearinghouse. References between the *Source-by-Observation Database*, and the *Master Source Database* will need to be retained in both directions, requiring multi-level referencing. This is needed to account for different detection patterns for a given region of the sky when at the telescope axis vs. at large off-axis angles, where the angular response of *Chandra* is significantly degraded.

The data products in the L3 Archive include the observation event files resulting from the *Detect* pipeline, and source-based *Data Objects*. These data objects, stored in FITS files, are the results of *Per-Source* processing (postage stamp, source based events, images, light curves, spectra and region information). Retaining these data objects in the archive will allow alternative parameter extraction by users, providing flexibility in archival analysis.

5.2. Versioning

Versioning is an important aspect of archive management because *Chandra* is an active mission, and we anticipate it will be on-going for many years while the catalog is being generated. As a consequence, the contents of the catalog tables and data objects will improve daily with the addition of new public data, which will increase the observed fraction of the sky or the depth of observation in some regions. The databases will be continuously updated as observations are added or reprocessed. To manage this evolving data archive, we will keep a complete history of catalog updates, including each version of each record, and the dates when each record was created or superseded. Users will have access to the current database, but also will be able to download past versions of the catalog. To facilitate use of the catalog, catalog releases will be carefully controlled and well characterized; certain significant dates will be aliased as virtual releases (or snapshots).

5.3. User Interface to Catalog and Data Objects

We envisage a web-based primary *Catalog* user interface, supporting Web-GUI queries equivalent of SQL or ADQL queries. This interface will allow selection from up-loaded files and equivalent command-line SQL-like querying, as well as full logging of submitted queries for traceability and usage analysis. The interface will reflect the evolving GUI technology and VO standards. All *Data Objects* access will be navigated through the databases. We will provide two methods of access to data objects: through the Catalog UI, as virtual column in the *Select* part of the query for either bulk retrieval or links in the results table; and through a VO-compliant API to tools and work-flows, allowing users to construct their own catalogs, do bulk analysis processing on sources (using either CIAO or user-provided tools), and construct sophisticated queries.

6. Processing Hardware: a Dedicated Beowulf Cluster

Catalog pipeline processing will be significantly more compute-intensive than the on-going *Chandra* standard data processing, requiring approximately 9 hr per observation per CPU, and 10 min per source per CPU on our processing hardware. Parallelization is essential for fast turnaround, and it is easily achievable within our pipeline paradigm. We have acquired a Beowulf cluster with 14 dual-CPU compute nodes to support this processing. With this hardware, on which we are currently testing our Level 3 pipelines, we anticipate being able to process all the data in-house in about four months.

7. Future Developments and Tools

Future releases of the Catalog will address the co-adding and mosaic-ing of the data from surveys of the same and adjacent regions of the sky, and the detection of sources from these mosaiced data sets. Tools for calculating the sensitivity far a given location and to mosaic the data off-line will be added to the CIAO toolset.

8. Summary and Outlook

In summary, the CXC has undertaken the production of the *Chandra* source catalog. Characteristics of this catalog include:

- *Characterization*, including determination of uncertainties for the catalog entries, and sensitivity maps for the entire observed portion of the sky.
- *Astrometry*: This is going to be the X-ray astrometric catalog for years to come, and will provide the best possible resolution for crowded fields.
- *Traditional catalog parameters and data objects*: both accessible in catalog's column and retrievable through the same interface; data objects will allow novel user interaction and data analysis.
- *Dynamic catalog with full version traceability*, to take advantage of the increasing observation base during the ongoing *Chandra* mission, while allowing users a steady reference base for their work.

Our goal is to present a pre-release catalog at the next ADASS, followed by a first public release at the AAS meeting in 2008 January.

Acknowledgments. We thank the following individuals, who have contributed to the catalog project at some time in the past few years: A. Dobrzycki, M. Elvis, P. Freeman, D. Harris, S. Paltani, J. Slavin, M. Wise. This work is supported by NASA contract NAS 8-03060 (*Chandra* X-ray Center).

References

Evans, I., et al. 2006, SPIE, 6270, 59

Weisskopf, M. C., Tananbaum, H. D., Van Speybroeck, L. P., O'Dell, S. L. 2000, Proc. SPIE, 4012, 2

Zografou, P., Harbo, P., Tibbetts, M., Van Stone, D. 2007, in ASP Conf. Ser. 376, ADASS XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell (San Francisco: ASP), 327