CXC Response to Source Catalog Review Panel

2006 Apr 2

1 Introduction

On the suggestion of the CUC (Chandra Users Committee), the CXC convened a panel to review the CSC (Chandra Source Catalog) project and its associated L3 (Level 3) pipeline. The panel met on 2006 Feb 8-9; preliminary feedback was provided during the meeting, with a fuller draft report sent on Feb 22; a final set of recommendations is awaited. The CXC was directed to provide its response both to the review panel and the CUC by the time of the CUC meeting in 2006 April. This document, together with documents made available to the committees, constitutes that formal response, which is based on the Feb 22 draft.

The CXC is grateful to the panel for its hard work and its recommendations. What follows is our initial response to those recommendations; we will continue to keep the community informed on our evolving plans, principally via the CUC.

R1: A Requirements Document should be produced

The CXC has produced a first draft of the requirements document (RD). We have attempted to capture a detailed description of the catalog contents and goals. We have not yet incorporated the definition descriptions of our implementation algorithms which are currently documented within our software system; these will be provided in a later draft. The RD is accompanied by other documents discussing design trades and auxiliary notes, which will eventually be defined as appendices.

R2: Descope first release to support quick completion

The panel suggested some descope possibilities: e.g. exclude certain difficult kinds of region, omit HRC, don't do fancy variability, make the UI crude.

Our initial review makes it clear that the driving factor for a quick release is the scientist time needed for catalog characterization, and not coding time, processing time, etc. We therefore expect to run the pipeline more or less as presently planned, but some outputs may not be included in the catalog. In parallel, we are developing a detailed catalog characterization plan, with particular emphasis on identifying those tasks which can be either deferred to later releases or substituted by simpler (albeit in some cases less accurate) analyses for the initial release.

At the same time we acknowledge the value of some of the descope possibilities suggested by the panel. We do feel, however, that the basic outputs of source existence, position and flux are critical, and getting them right implies getting a lot of other things right too (e.g. exposure). We need to perform an interdependency analysis to determine which catalog entries and data objects may be descoped without impacting these key results.

R3: Aim at multi-wavelength astronomer

The panel emphasized that the catalog should be targeted at the general multi-wavelength astronomer as the 'most important customer' rather than the X-ray expert astronomer.

Some examples of implications:

- the basic source catalog tables are the most important products (vs. the data objects)
- worth providing the energy flux (erg cm^{-1} s^{-1} keV^{-1}) and not just photon number flux, even though the latter is better constrained, since the energy flux is needed for multiwavelength comparisons.

We have generated a new set of use cases focusing on general astronomer use of the catalog, and are deriving requirements on the catalog and UI from these use cases.

R4: Distinguish between database and catalog

The panel directed us to add an extra stage of catalog definition involving filtering, merging and quality assurance. There will then be a 'database', containing all the latest pipeline results and a 'catalog'. The catalog is both a subset and a snapshot in time, containing a well characterized product. Both database and catalog consist of a master source table, a per-observation source table, and data objects such as PHA files. The difference is that the catalog has controlled (versioned) releases; has a subset of the sources; a subset of the table columns and data products, whose characterization we have a higher level of confidence in.

R5: Run pipeline to faint limit

We concur with the committee that the distinction between database and catalog allows us to run the pipeline to a deep threshhold while using a more conservative threshhold for the catalog.

The implementation consequences include minor extra bookkeeping requirements on the database and on the merge pipeline. For database sources with multiple detections, we must keep track of which detections are included in the

catalog while retaining the ability to retrieve links to non-included detections of the same source.

R6 Scope the UI soon

The panel felt that our UI plans were both vague and overambitious; the UI can drive some aspects of the back end functionality; and that we should "complete a very simple first UI design as soon as possible".

We have begun the definition of the UI by identifying outline requirements at minimum, highly desirable and longer-term levels; these are summarized below.

Minimum UI requirements
Web-based interface (no download of application required)
Access to all fields in master and per-observation source catalogs
Support cone-search type (location cross-match) interface
Support SQL-based interface implementing a subset of ADQL
Ability to upload lists of target positions/errors for crossmatch
Links to L3 data objects
User selection of fields (columns) and number of return rows
Return sorted sources with top N values of querys
Return results in plain text or HTML
Highly desirable UI requirements
Access to upper limit/sensitivity data
Link between sources and full field images
Name resolver in query interface
Support VOTABLE output
Virtual column definitions (query on functions on columns)
Longer term UI requirements
Full ADQL implementation
ADQL equation scripting
Integrate functionality with NED, SIMBAD, DataScope
Integrated link to Vizier and USNO-B, etc.
Link to proposal information in Chandra OCAT
Ability to query previous editions of catalog
Ability to query underlying database directly
Return flux in user defined band via event or pha data
User API (e.g. web service) access
Links to VOPLOT and other VO apps.

R7: Investigating External Solutions

The panel drew attention to ACIS Extract, 1XMM and XASSIST, and felt that we had not sufficiently described how we had looked at these solutions, and why we had or had not adopted their approaches.

In fact, the panel's phrasing was stronger: "team were often not aware of, or seemed to have ignored, existing solution ...". We believe this criticism is unfair, as we have indeed reviewed the three main solutions cited (and were aware of the majority of the other references listed). These approaches have indeed influenced our design, and the attached documents include comparative notes.

Nevertheless, we will revisit our rationale for not adopting some of the approaches in these other solutions.

R8: Quality Assurance Plan

The panel believes that fully automated quality assurance is not workable, and that we should plan manual spot checks. They also recommended that we clearly separate a catalog production and quality assurance step from the pipeline production of the database. This production/QA step can then be rerun as needed, independently of the database pipeline.

We did not describe our plans in sufficent detail in our presentation to the panel; we agree with their recommendation and it is essentially in line with our existing plans. The separation of the merge/filter/QA step has only a minor impact our our development and is consistent with our architecture.

As we commented at the review, we don't have the resources to perform manual inspection of all observations/sources, but we agree that manual spot checks are necessary.

R9: Extended emission

The panel agreed that accurate treatment of extended emission can wait until a later release; they emphasized that R&D work is needed on this in the shorter term. The Science Data Systems team has an ongoing effort on this problem in the CIAO context.

R10: Merging observations

The panel emphasized that for later releases it is important to run detect on merged observations of the same field, rather than just merge source lists from separate detect runs, which would miss faint sources.

We still have work to do on qualifying the detect algorithm and characterizing its results for this case. However, CXC needs to figure this out for normal CIAO users anyway, so we anticipate that this will be part of ongoing CIAO development in the near future.

The panel also requested full-field, background-corrected smoothed images and we will fold this into our plan.

R11: Avoiding low priority issues

For example, the panel noted that we shouldn't spend time worrying about pileup since it only affects a small fraction of sources. We note this recommendation, and comment that in many cases we are merely adopting knowledge obtained from supporting general user data analysis, rather than spending any catalog-project-specific time on the issues.

Other recommendations

- The panel felt the use cases presented at the review were too complicated and too VO-oriented. We have developed a draft of a new use case list which consists of much simpler use cases oriented to the multiwavelength astronomer.

- Choice of energy bands. We accept the suggestion to separate source-detection band from color-measurement bands. We have begun new simulations to optimize the detect energy bands; preliminary results suggest that in the soft band, the detection efficiency for supersoft sources is insensitive to the exact energy boundaries used. We have not yet selected the color-measurement bands to be used.

- Spectral fitting options. We will review the spectral fitting plans as part of the descope study.

- VO compatibility. We agree with the panel's comments; ADQL query and VOTABLE export constitute our baseline VO interoperability, and in later releases we as CXC will follow rather than lead the stabilization of VO standards; our separately funded VO team will of course remain in the forefront of establishing these standards and will be influenced by the needs of the catalog project.

- Variability characterization. We hope to include both KS, Kuiper and Gregory-Loredo estimates. The important thing is to provide some simple guidance to the user about whether the source is definitely, probably, probably not or definitely not variable; any high precision question about the variability or its nature will require analysis of the light curve anyway.

- Flux characterization. We will pursue approaches to model-independent flux estimation for the catalog.

- Links to observation database. We note the panel's recommendation. We expect that these links may not make it into the first release; we further note that the relevant metadata in the observation database is heterogeneous and incomplete, limiting the usefulness of the queries the panel contemplates. However, we will baseline these links for inclusion in a later release.

- Early Science Project. The catalog team will consider adopting such a project as part of its effort to develop further use cases.
- Incorporation delay. At this time, the CXC team does not feel that a specific incorporation delay beyond the normal proprietary period is appropriate; we will adopt the same (brief) verification process which is independent of the proprietary nature of the data.
- Verification against other X-ray surveys. As part of our test sequences, we have included and will include sequences in common with ChaMP, ChaMPlane, and COUP. We agree that comparison with other surveys is an important aspect of characterizing our final catalog.
- Photon arrival diagram. We have no current plans to include a photon arrival diagram in the first release of the catalog, but will consider it for later releases.