CHANDRA SOURCE CATALOG PROJECT REVIEW

REPORT OF VISITING PANEL (Draft 2006-02-27)

Summary of Key Recommendations

Urgent key recommendations

R1 A **Requirements Document** should be produced
R2 The scope of the first release should be firmly limited
R3 The design should be aimed at general multi-wavelength astronomers
R4 A distinction should be drawn between *database* and *catalog*
R5 The pipeline should be run to the faint limit
R6 The User Interface design should be scoped as soon as possible.

Medium to long term key recommendations

R7 External solutions to a wider range of problems should be investigated.
R8 A **Quality Assurance Plan** should be produced
R9 Better techniques should be developed for dealing with extended emission
R10 The following should be future priorities
- R10a Merging of observations
- R10b Dealing with larger sources
- R10c Provision of background subtracted exposure corrected full images

R11 The team should carefully avoid interesting but low priority technical concerns

Report contents

(1) INTRODUCTION

Following suggestions from the Chandra Users Committee (CUC), CXC management commissioned a review of their plans for completion of a Chandra Source Catalog (CSC), requesting recommendations and a written report. The membership of the panel is listed at Annex-A. The panel met at CfA during Feb 8-9 2006, and were given a series of presentations by CSC team members on progress and plans. These presentations, and other material provided in advance of the review, are available on-line at ftp://cda.harvard.edu/pub/arcftp/L3review/ . The following morning (Feb 10) the panel met privately, discussed their findings, and then presented recommendations to the CSC team. At this feedback meeting, the panel stated its intention to complete a written report in approximately two weeks, suggested that the CSC should produce a written response, and suggested that both of these should be provided to the next CUC meeting in April.

(2) GENERAL POINTS

The most general finding of the panel is that compared to many similar survey and/or catalog projects at a analogous stage of development, the CSC is actually in a good general state of preparedness; however given the short timescale for completion, the danger of failure is high. Furthermore, the general preparedness is to be expected given the maturity of the Chandra data pipeline upon which the CSC project is building. There is therefore no room for complacency. However, by performing some fairly brutal descoping, the panel believes that the CSC could in fact deliver an excellent product on time, followed by improved products in due course.

The panel recognised that the short timescale is a self imposed one. However, rather than suggesting that delivery be delayed, the panel strongly believes that rapid delivery of a simplified first product is the correct thing to do.

(3) COMMENDATIONS

The panel unanimously felt that the CSC is an important, exciting, and timely project, and strongly supports its completion. It is felt to be blazing a path which other facilities (such as the Hubble Space Telescope and the Spitzer Space Telescope) are likely to follow. It was agreed that an excellent start has been made, that a very talented team is in place, and that a number of key issues have been identified by the CSC team. These key issues include :

- the heterogeneity of the archive contents and the output product
- the non-uniform sensitivity on the sky
- the wise approach of building on the use of existing L1 and L2 software
- the importance of preparing for the Virtual Observatory (VO) context
- the importance of phased delivery of the intended products

Our most important recommendation is essentially about taking the last key issue (phased delivery) even more seriously.

(4) URGENT RECOMMENDATIONS

The panel agreed six recommendations that are considered both important and urgent

R1 A **Requirements Document** should be produced
R2 The scope of the first release should be firmly limited
R3 The design should be aimed at general multi-wavelength astronomers
R4 A distinction should be drawn between *database* and *catalog*
R5 The pipeline should be run to the faint limit
R6 The User Interface design should be scoped as soon as possible.

Some explanatory points are in order.

(R1) **Requirements Document**. A short and simple requirements document is urgently needed to act as a "Project Bible" if it is to be possible for the team to stay coherent, well focused, and on schedule. The panel do not intend that the project go through an elaborate new requirements analysis; rather that the current analysis be captured and summarised as soon as possible. The document should be concise and pragmatic rather than detailed. It should be used (a) to plan the remainder of the project, (b) to assess the required resources, and (c) to monitor progress. It should be a useful document for the CUC as well as CXC management. In Annex-B, we present some suggestions for what a requirements document might contain. With reference to Recommendation R4 (separation of database and catalog) the panel strongly suggest that requirements for the database and the catalog be clearly separated.

(R2) **Descope First Release**. The project is being too perfectionist in its planning for the first release. We strongly recommend that the first release stay on schedule but that its scope and capability be cut back with the aim of making the initial demands on the project as simple as possible. The panel did not dictate what this descope should be like, but did have a menu of suggestions :

- exclude high background regions
- exclude crowded fields
- process ACIS data only
- cut back on the live data objects, eg don't provide ARF and RMF
- ignore spectral fitting
- provide only most basic variability information
- keep the User Interface as crude and simple as possible

(R3) **Aim at multi-wavelength astronomer**. There are several potential "customers" for the CSC - e.g. X-ray astronomy experts, more general astronomers, high school students, and so on. In choosing the priority features of the system, the panel felt the most important customer was an astronomer wishing to do multi-wavelength work including Chandra data - not an X-ray expert, but reasonably skilled and knowledgeable.

(R4-5) **Distinguish database and catalog; take pipeline to faint limit**. These two recommendations go together, and are also linked to R8 in the next section. During the review there were lively discussions about whether the catalog could be continuously updated (as opposed to released at major intervals); about whether the flux limits being imposed were too conservative or the opposite; about how to design the pipeline to achieve various possible completeness characteristics; about how automated production could be; and about how important various aspects of the pipeline were. The panel believed that many of these issues could be resolved by distinguishing the general *database* from the final *catalog*, which is produced from the database in a separate further stage of filtering, merging, and quality assurance. The pipeline should then be run to the faint limit, and should have as many objects as possible; the pipeline does not therefore need to be the perfect final word. The database can be continuously updated; but the catalog is a well

characterised product and carries an implication of completeness and reliability. It should be released in carefully controlled, announced, and documented stages.

Note that many of the hardest issues that the CSC team were concerned with - e.g. merging source lists, correct choice of completeness criteria - should not be part of the basic pipeline, but part of a separate process of producing the catalog from the database - separate in terms of requirements, development, and processing. The processing can partly be automated, but will also require a semi-automated Quality Assurance phase - see recommendation R8.

(R6) **Scope UI as soon as possible**. The plans presented for the User Interface (UI) were worryingly vague, possibly ambitious, and presented as a kind of final "skin" placed on the project after the main build. By contrast we felt that UI conceptual design should be urgently completed as they should drive the backend functionality. The panel were aware that the team concerned are capable of producing an excellent UI; however we recommend that they complete a very simple first UI design as soon as possible.

(5) MEDIUM AND LONG TERM RECOMMENDATIONS

Several other recommendations were considered just as important as those of Section 4, but not as urgent.

R7 External solutions to a wider range of problems should be investigated.
R8 A **Quality Assurance Plan** should be produced
R9 Better techniques should be developed for dealing with extended emission
R10 The following should be future priorities
- R10a Merging of observations
- R10b Dealing with larger sources
- R10c Provision of background subtracted exposure corrected full images

R11 The team should carefully avoid interesting but low priority technical concerns

Again, some explanatory words are in order.

(R7) **Investigating external solutions**. The panel were concerned that the CSC team were often not aware of, or seemed to have ignored, existing solutions to many relevant problems. It may well be that these have in fact been considered but rejected for good reasons, but if so the reasoning was not apparent to the panel. Annex-C provides a lists of hopefully helpful references.

(R8) **Quality Assurance Plan**. The project team seemed to believe that they should design the pipeline to produce a final reliable catalog in as automated a fashion as possible. This contrasts with the experience of nearly all catalog projects, which always re-discover the importance of a final stage of filtering, merging and correcting which *cannot be fully predicted in advance*. This fits with our recommendation to conceptually separate database from catalog; a final distinct stage is needed to produce the catalog. This does not mean that the panel recommends manual inspection of every image or source; rather that a variety of manual spot checks be performed; that software to produce the catalog from the database be planned; that a variety of diagnostic tests be designed to be run on data products; and that staff effort be put aside to develop and operate this process. All these things can be loosely grouped together as "Quality Assurance" and need to be thought out and planned. We therefore recommend the production of a documented QA plan.

(R9) **Extended emission**. The CSC team were clearly aware that they were taking a relatively simple approach to dealing with extended emission, and this was agreed to be a reasonable start for

the first release. However, improved techniques should be investigated as soon as possible in readiness for later releases. Key issues include better characterisation of extended emission; how to find point sources on top of a varying background; and the possibility of characterising a surface brightness limit at multiple scales

(R10a) **Merging of observations**. For the first release, the panel agreed that the best approach is the simple one of extracting sources from each observation, and later merging these source lists (followed by filtering and quality control). However to get the full depth from the Chandra data, it will eventually be necessary to merge the observations themselves and/or extract a source list using joint information from several observations. This is obviously a difficult problem, but is a high priroity to achieve for future releases.

(R10b) **Dealing with larger sources**. The team described their intention to limit extended source detection to those sources no more than an arcminute across. The panel agreed that this approach is sensible for the first release, but that it misses a large amount of potential scientific value. Developing methods for locating and characterising much larger extended sources, should be a priority for future releases.

(R10c) **Supplying full images**. As Chandra is likely to be the premier imaging X-ray observatory in many of our professional careers, inclusion of full-field images in the L3 products should be a vital component. This requires that in conjunction within the development of CIAO there must be a reliable method for modeling the non-cosmic background, as well as exposure-correcting the images. For presentation purposes, the images will in general need to be adaptively smoothed.

(R11) **Avoiding low priority issues**. The panel felt that the CSC team were occasionally spending too much time analysing concerns that weren't scientifically the most important, or those most likely to make a big difference to the usefulness of the final product. We therefore recommend that the team carefully review priorities in human resource expenditure. A specific example is the issue of photon pile up, about which the team had thought very hard, but which made a difference in a tony minority of sources.

(6) OTHER RECOMMENDATIONS

The panel arrived at a number of other recommendations that the CSC will hopefully find helpful. These often concern areas where there is not yet a consensus of views, and/or considerable technical difficulties are involved. We therefore do not on the whole mandate specific technical solutions, but rather recommend further work. Individual members of the panel would be happy to provide further advice on specific issues.

Use Case studies. The panel felt that the Use Cases presented were too complex and challenging, and too focused on VO issues. The CSC needs to work through some very simple (catalog only) use cases; to consider a wider scientific range of cases (eg galactic astronomy, and those involving variability); and to pick a “Top Five” list. The project should be careful to draw simple direct conclusions about the basic design of the pipeline and database rather than elaborate downstream VO related possibilities.

Band Choice. The panel felt that further reasoned consideration should be given to choice of bandpasses, in particular bearing in mind that the source-finding bands may differ from the bands used for reporting colors in the catalog. The source-finding bands should be such as to maximise S/N, whereas the catalog-reporting bands should provide maximum astrophysical diagnostic power, and/or match bands previously used in the literature. The panel noted that for the

SDSS survey, the choice of bands was considered important enough, and difficult enough, that the logic behind their choice was written up as a refereed paper.

Spectral Fitting Options. While spectral fitting can obviously provide very useful information for a subset of sources, the panel were concerned that it could easily consume too much processing time and too much development effort in worrying about how to characterise it. As noted above the panel recommends that for the first release the project should consider avoiding spectral fitting completely. For later releases, further analysis of the likely science use cases should be performed before finalising the method of characterising the spectral energy distribution. For example, if a power law is not always a good model, using NH, index, and normalisation as three parameters may be more misleading that simply picking three colours. Any spectral characterisation should also be very careful to indicate when the characterisation is likely to be meaningless, rather than just quoting a formal fitting error.

VO compatibility. The panel felt the CSC must be careful to plan for VO compatibility in stages. As of today, this would mean being able to accept ADQL queries, and being able to export VOTable, and very little else. In due course other issues will become important - for example standardised authentication or new data model standards - but the project should not guess in advance what the VO world will stabilise upon.

Variability characterisation. The panel recognised the careful thought that had gone into the Gregory/Loredo method, but note that it is not well known in the community. Thus for a general user catalog we feel that it is important to include a more common test that is easy to interpret (such as a KS probability). For the first release, the variability characterisation should be as simple as possible. For later releases, some thought should be given to variability science use cases, so that some more advanced variability information, analogous to spectral fitting options, can be provided for those sources where this is appropriate. For example, assuming a standard spectral index for the power spectrum, one can derive an amplitude of variability which corrects for the effect of observation length.

Flux characterisation. This point is related to the one above. For the catalog, it would be preferable to calculate fluxes (ergs/cm**2/s) in a model independent manner, that is to divide the photons energies by an effective area rather than calculating a power law model. This approach allows one to calculate fluxes and errors from multiple observations, wheres calculating fluxes from a spectral fit will not.

Linking to observation database. Information from the original proposal and other sources could often be of importance - for example knowing that an observation was targeted at a cluster of galaxies. Such information needs to be propagated to the source database and the catalog.

Early Science Project. Some members of the panel suggested that the CSC team could design an actual science project that they would intend to carry out themselves, as this could be a very useful driver of functionality. There might of course be some community perception issues in doing this. The panel did not want come to a conclusion about whether it was the right thing to do or not, but do recommend that the CSC team discuss this possibility.

Incorporation delay. Some members of the panel suggested that instead of incorporating data into the database as soon as the proprietary period has expired, the project should consider using a delay of one year. This will result in a more uniform catalog, and will make sure a few "discrepant observations" are not incorporated into the archive.

Verification against other survey datasets. As a standard part of the validation/verification

process, we suggest the project should compare the CSC's database and final catalog output on specific ObsIDs to the output from other survey projects - eg Champs, Champlane - and/or from other catalog extraction systems - eg ACIS Extract, Kip Kuntz's analysis.

Photon Arrival Diagram. The project might want to consider supplying a plot of median energy vs. time or a "photon arrival diagram" for each source. For the Chandra Orion Ultradeep Project (COUP), Getman et al. 2005 (ApJS 160, 319) generated a 1600-page Atlas giving various plots, images, and properties for each source (see page 350 of that paper). A photon arrival diagram is given there.

Annex A : Review Participants

Review Panel

Andy Lawrence (chair)	Head of Physics, Edinburgh
Roc Cutri	Deputy Director, IPAC
Megan Donahue	Michigan State
Knox Long (CUC chair)	STScI
Robert Lupton	Princeton
Tom McGlynn	HEASARC, GSFC
Brian McLean	STScI
Marion Schmitz	NED, IPAC
Steve Snowden	XMM-Newton, GSFC
Leisa Townley (CUC member)	Penn State

Also Attending

Allyn Tennant	GSFC

CSC Project Presenting

Pepi Fabbiano
Ian Evans
Arnold Rots
Jonathan McDowell
Janet DePonte-Evans
Frank Primini
Kenny Glotfelty

Annex C : Suggestions for external solution investigation

Annex B : Initial Suggestions for requirements document.

Collected here are initial thoughts on the possible contents of a requirements document, as given to the CSC team by the review panel on Feb 10th. They have not been further edited here. Individual members of the review panel would be happy to elaborate further.

```
-- capture current scientific and technical analysis into a concise and
-- pragmatic requirements document against which the project will plan,
-- assess necessary resource, and monitor its progress.

All requirements apply to catalog not database

All requirements where appropriate should quote RMS error & e.g. 95% points on
requirements.

All quantities should have errors and units

Band definitions (number of bands; limits)

Source completeness for PSFs at a given total counts as a function of
bkgd and off axis distance

False source rate for PSFs at ditto

Specify source properties with requirements document for PSFs  (all
as function of ditto)

Property list to include, at a minimum, for release 0.

      Centroid
            relative accuracy
            absolute accuracy

      Energy flux within band -- please specify if with or without ARF/RMF
             1. PSF fluxes

             2. Aperture measures at 50%, 90%, 95% of circular PSF
             enclosed energy (corrected for detection regions)

      Measure of object size per-band or broad band,
      corrected for PSF size, for all sources

      Probability of being extended

      Variability
            probability of being variable
            noise corrected RMS variability

      Per-object data quality flags:
            E.g. EDGE, PILED, READOUT, CROWDED, problemsInProcessing

      Per-field data quality flags
            E.g. variableBackground, tooCrowded (e.g. SN 1006)


The catalog shall merge detection lists from different obsIds and filters,
and flag ambiguous cases.  Must maintain linkages to individual observations.

This requirements document shall specify how sophisticated the merge
must be for each release.  For release 0, the merged table shall handle all
straightforward cases.
```

As far as possible, the merged object table should carry average properties of its constituent observations. This requirements document shall define these averages.

This requirements document should be updated to define how CXC will include event-level coadds.

Annex C : Suggestions for external solution investigation

We suggest that all of the following would be useful for the CSC team to examine. It is not a complete list. Individual members of the panel would be happy to provide further suggestions.

External project web references

ACIS Extract User Guide
http://www.astro.psu.edu/xray/docs/TARA/ae_users_guide.html

XMM-Newton Serendipitous Source Catalog
http://xmmssc-www.star.le.ac.uk/

Andy Ptak's XASSIST software
http://xassist.pha.jhu.edu/xassist/index.jsp

Recent AE-based papers

Townsley et al. 2006a astro-ph/0601105 (30 Doradus diffuse regions)
Townsley et al. 2006b astro-ph/0601106 (30 Doradus point sources)
Getman et al. 2006, ApJS 160, 319 (COUP point sources)
Muno et al. 2006 astro-ph/0601627 (GalCen survey)

The latter paper includes an appendix on estimating completeness.

Other example papers on large Chandra catalogs

Muno et al. 2003, ApJ 589, 225 (GalCen deep data)
Lehmer et al. 2005, ApJS 161, 21 (Extended Deep Field South)
Virani et al. 2005, astro-ph/0506551 (another paper on the EDFS)
Alexander et al. 2002, AJ 126, 539 (2 Msec catalog on Deep Field North)
Yang et al. 2004, AJ 128, 1501 (Lockman Hole catalog)
Several papers by Kip Kuntz on Chandra observations of M101
ChaMPlane papers
ChaMP papers

The Lehmer et al paper tackles problems of matching sources across ObsIDs, and sensitivity across field.
.
.