Chandra Source Catalog Review

Architecture

Ian Evans
Chandra X-ray Center
February 8, 2006

Primary Architectural Design Goals

- Allow simple and quick access to the best estimates of the X-ray properties and *Chandra* data for individual sources with good scientific fidelity
- Directly support scientific analysis of individual source data
    - Acknowledges that it is almost always better to tailor analysis to individual datasets for the highest accuracy (for example, fine tuning source models for spectral fitting)
- Facilitate easy searches and analysis of a wide range of statistical properties for classes of X-ray sources
- Include all real X-ray sources detected in all publicly available *Chandra* datasets, while maintaining the number of false sources at an acceptable level
    - Processing must operate on full field of each *Chandra* dataset

Secondary Architectural Design Goals I

- Support manipulating the actual observational data for each X-ray source in addition to data recorded in tabular form
- Use all of the available data from multiple observations to extract the best estimates of the X-ray source properties
  - Recognize that the various source properties may be best determined from different observations
- Allow continual updating as new observations become public and as new calibrations and algorithms are developed

Secondary Architectural Design Goals I

- Make use of existing software infrastructure and data analysis tools to build and manipulate the catalog
- Where new software tools must be developed, do so in a manner that benefits the portable *Chandra* data analysis system (*CIAO*) users whenever possible
  - About a half dozen new tools have been developed that meet this criterion
- Leverage knowledge regarding problems, approaches, and algorithms developed by the broader X-ray community to improve the quality of the catalog and reduce development time
  - Examples include 1XMM and 2XMM source catalogs, **`acis_extract`** software tool, *ChaMP* and *ChaMPlane* surveys, **`Xassist`** software package

Derived Architectural Goals

- Catalog construction must run automatically on all public *Chandra* (imaging) datasets
  - Make most efficient use of limited resources
  - Implies
    - No manual selection of datasets
    - No manual initiation of processing
    - No manual processing steps
    - No manual verification and validation on a per-dataset basis
      - Will perform manual spot-checking
  - Automation enhances processing/catalog uniformity
  - Will perform manual characterization of the catalog statistical properties

Architectural Limitations of the Initial Release

- Observations not included in the initial release of the catalog
  - Moving target observations
  - ACIS "continuous clocking" mode observations
  - Transmission grating observations
  - (Some of) these limitations will be removed in future releases
- The initial release is designed to perform well for point or small-scale extended sources but will perform less well for very extended sources
  - Limited by largest wavelet scale size run during source detection processing, ~30"
  - Extended sources will be modeled as point sources convolved with an elliptical Gaussian

Improvements Expected in Future Catalog Releases

- Add support for additional observation types
  - Include background sources detected in moving target observations
  - Include transmission grating observations (0th order and spectral data products)
- Improve catalog performance for very extended sources
  - These require an alternative approaches for source detection and determination of source properties
    - Source detection would probably use a Voronoi tesselation and percolation procedure (very compute intensive!!)
    - Source morphologies could be characterized via shapelet analysis

Catalog Overview I

- Traditional source catalogs record properties only in *tabular form* (data tables)
  - These limit the user to searching the data that the catalog creators considered important (or simple combination of those properties)
- The *Chandra Source Catalog* incorporates data tables for source properties that require little interpretation, or where a common interpretation may be useful
  - Example of the former is sky coordinates (α, δ)
  - Example of the latter is power law spectral index fit to the data

Catalog Overview II

- The *Chandra Source Catalog* incorporates *live data objects*
  - Directly accessible data for individual sources
    - Actual data that can be manipulated for further analysis
  - Examples are the photon event data for a source, source image, source spectrum, and light-curve
  - Include source-specific calibration data, including the exposure map, ARF, RMF, and PSF at the source location, and *matched to the processed data* (i.e., created using common calibrations)
  - Maintained on a per-source, per-observation basis
  - The live data objects are a by-product of the pipeline processing necessary to populate the data tables

Catalog Overview III

- In addition to searching the tabular data, users will be able to manipulate the live data objects to
  - Perform scientific analyses of (sets of) sources without having to download entire observations
  - Search the data for signatures
  - Extract source property information alternate to that incorporated into the data tables
    - Example would be to extract colors in different energy bands

Master and Per-Observation Objects

- For each cataloged source the catalog includes one *master object* and one or more sets of *per-observation objects*
  - There is one set of per-observation objects for each observation (aka *Observation Interval* or *ObI)* in which the source was detected
    - Each set of objects includes all of the data about the source extracted from processing a single observation
  - The master object includes the best estimates of the source properties based on combining the information from the observations in which the source was detected
  - There are links between the master and per-observation objects for each source, so that all of the individual observation data for a source may be accessed

Data Object Summary

- Per-observation objects (subject to revision)

Observation	Observation information
Per-observation Source	Source properties extracted from the observation
Event	Photon event data
Image	Counts image, PSF, and exposure map
Lightcurve	Light curve (source and background)
Region	Source and background region information
Spectrum	PI spectrum (source and background), ARF, and RMF

- Master object

Source	Master source properties merged from all observations

Estimated Catalog Size

- An estimate of total source counts over the mission lifetime is needed to “size” the catalog processing and storage requirements
- Simple estimate based on source counts present in existing observations and scaling to 15 year mission, and *assuming no significant changes in detected source density* yields an estimated ***catalog size of ~400,000 sources***
  - Details of estimate provided elsewhere
- For estimating required resources, we therefore assume a catalog size of 1,000,000 sources

Non-Catalog Deliverables

- In addition to the catalog proper, we plan to make available to the user community a tool to support sensitivity analysis, and set of mosaics of interesting regions of the sky
  - Sensitivity analysis tool (initial release)
    - Will provide a mechanism for users to determine the limiting sensitivity for the catalog source detection process at any location within the fields of view of all processed observations (i.e., not just at source locations)
  - Mosaics (future releases)
    - Image (and event?) mosaics for interesting “key” regions
      - LMC, M 31, Orion, Sag A etc. (TBD)
    - Sensitivity mosaic
    - Mosaics will include links to detailed catalog source information
    - Need to identify granularity (pixel scale) on the sky most appropriate for image and sensitivity mosaics

Catalog Version Control

- The *Chandra Source Catalog* will be subject to continual updating as observations become public
  - New master source entries will be *added*
  - Existing master source entries may be *revised* if they are included in the field of view of a newly processed observations
  - Existing master source entries may be *deleted*
    - Because the PSF varies across the field of view, a single “source” that was detected well off-axis (big PSF) in one observation may be re-identified as multiple sources in a new on-axis (small PSF) observation
- Need to maintain catalog history and version information!

History and Snapshots

- Per-observation objects are dependent on a single observations, and so are stable against processing newly public observations
  - Must be updated if the observation is reprocessed
- Master objects must be updated if the source is detected in a newly processed observation
  - Data volume is *small* so maintaining a history is feasible
  - History mechanism enables the state of catalog at any specified date/time to be viewed
    - Users will be able to access the catalog view for any specified date/time as well as the current (continually updated) “live” catalog view for those who need access to the latest data
- Propose to release catalog “snapshots” as appropriate
  - Snapshot is an alias for a predefined date/time

Catalog Reprocessing I

- Reprocessing observations through the per-observation pipelines is supported by the catalog architecture
  - Individual observations could be reprocessed if needed (e.g., due to processing errors that affect a limited number of observations)
- "Bulk" reprocessing of all observations would be resource intensive
  - Circumstances under which the CXC would do this have not been identified, but could conceivably include major improvements to calibrations or algorithms (i.e., similar criteria to bulk Level 0 – 2 reprocessing)
    - Could conceivably leverage off such reprocessing; however most of the computation in catalog processing is not duplicated in Level 0 – 2 reprocessing, so this is not a significant factor

Catalog Reprocessing II

- *Constant improvements imply that the catalog sources in the live catalog will have been processed heterogeneously with respect to both calibrations and algorithms*
  - Inevitable while the mission is still in progress
- Users will be able to select only sources and datasets that were processed with specific versions of the software and calibration products, for studies where homogeneity of processing is critical

Catalog Production Overview

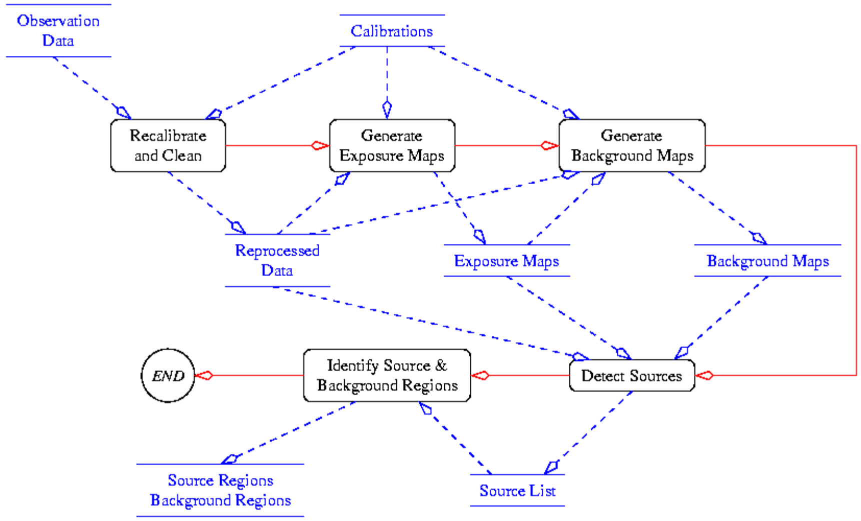
- Catalog production is split into two principal steps
  - Per-observation processing
    - Triggered by dataset becoming non-proprietary
    - Populates per-observation objects in the catalog for each detected source
  - Merge processing
    - Triggered by population of per-observation objects for a dataset
    - Identifies matching sources currently in the catalog (if any)
    - Merges existing per-observation objects for matching sources with new per-observation objects and creates/updates master objects for those sources

Per-Observation Processing Overview: Detect Sources I

- Source detection is performed separately for each observation and energy band
  - For both instruments, source detection is performed using a broad-band energy filter
    - 200 – 7500 eV for ACIS
    - 0 : 254 PHA for HRC
  - In addition, for ACIS, source detection is performed in soft, medium, and hard energy bands
    - More sensitive than broad-band if the source spectrum is strongly peaked in a single band (e.g., super-soft sources)
    - Band energies selected based on study of spectral response (EA and instrumental features), & comparison with commonly used bands in existing source catalogs and the literature

      200 – 500 eV (soft)
      500 – 2000 eV (medium)
      2000 – 7500 eV (hard)

Per-Observation Processing Overview: Detect Sources II



Per-Observation Processing Overview: Detect Sources III

- Reprocess dataset through Level 1 equivalent processing
  - Ensures *best currently available calibrations/algorithms* used
    - In routine operations, this occurs when the dataset becomes non-proprietary, typically ~1 year after original processing
  - Uses more conservative data cleaning criteria than standard data processing to minimize false source rate
    - Predicated on assumption that a user who retrieves an entire observation (Level 1 and 2) can selectively clean data based on their science goals, but in Level 3 data cleaning is performed prior to source detection so a catalog user cannot selectively clean data
    - Example: More rigorous background flare rejection

Per-Observation Processing Overview: Detect Sources IV

- Generate exposure and background maps in each band
  - Correct for instrumental effects that raise the background
    - Example: ACIS readout streak correction
- Detect sources in each band using wavelet algorithm
- Identify source and local background regions for each detected source for use in subsequent steps

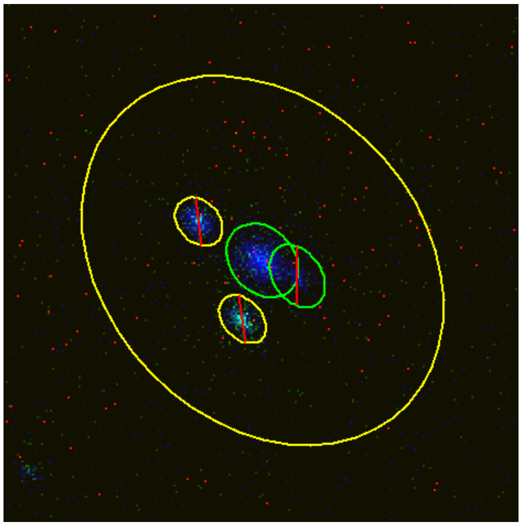
Wavelet Source Detection I

- Compared *CIAO* **`wavdetect`** (wavelet), *CIAO* **`celldetect`** (sliding cell), and **`SExtractor`** source detectors
  - **`SExtractor`** was modified to use Poisson statistics
  - Simulations and experiments using a wide sample of real data show that **`wavdetect`** has the best source detection rate and lowest false source rate, especially far off-axis
- Wavelet source detection is based on a *hypothesis test* rather than a *flux significance test*
  - A source is identified if the observed counts cannot be due to background fluctuations

Wavelet Source Detection II

- Wavelet detection algorithm (conceptual)
  - Counts image is correlated with a Mexican-Hat wavelet at a set of *user-specified* scales to determine a correlation value for each image pixel
  - Threshold correlation value is computed for each pixel based on the background estimate at that pixel and a *user-specified* tolerance for false sources
  - All pixels with correlation values greater than the threshold are identified as source pixels
- Detectability is a function of the local background
  - Equivalently, flux sensitivity limit depends on local background

Example Source and Background Regions

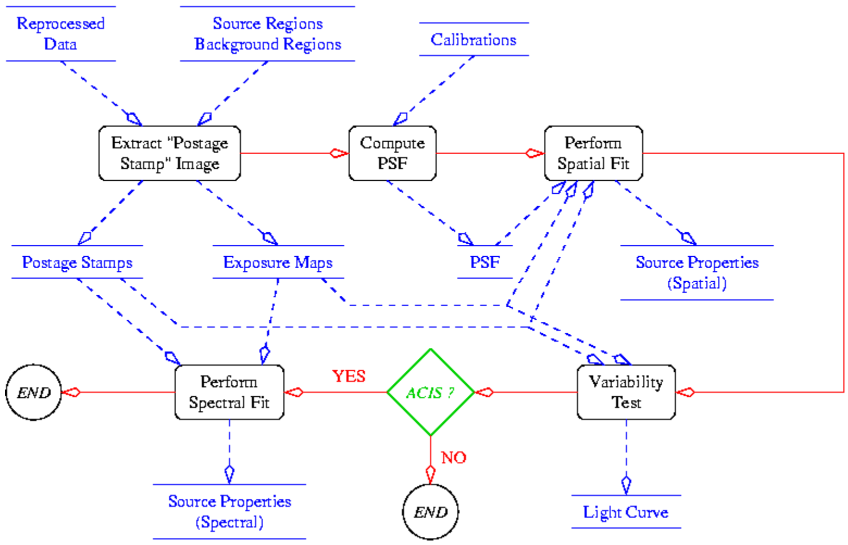


- Source region of centered source is the intersection of the included and excluded green regions
- Corresponding background region is the intersection of the included and excluded yellow regions and excludes the green regions
- Region dimensions are based on the source size determined by `wavdetect`

Per-Observation Processing Overview: Analyze Sources I

- Source properties are analyzed separately for each detected source
- Processing is performed for all of the energy bands (broad and soft/medium/hard), even if the sources was detected in only one band
  - Ensures that upper limits can be determined for the remaining bands

Per-Observation Processing Overview: Analyze Sources II



Per-Observation Processing Overview: Analyze Sources III

- Extract photon events from the rectangle bounding the background region
  - Construct “postage stamp” image, exposure map, fluxed image (photons cm^{-2} s^{-1}) and flux-error image
- Compute PSF at the source position from ray-trace and apply aspect blur
  - PSF computed at effective monochromatic energy of band
- Perform 2-D spatial fit to source
- Apply *a* variability test and create a light-curve
- If the detector is ACIS, then extract the source spectrum, compute the RMF and ARF, and perform a spectral fit

Why Perform A 2-D Spatial Fit?

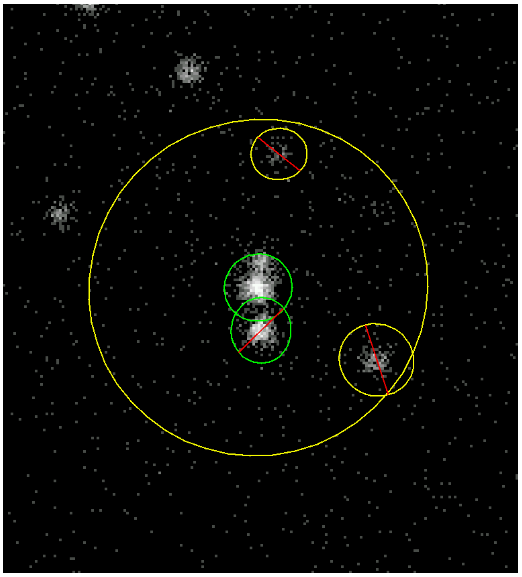
- Purpose is to improve the quality/reliability of source detections and determination of source properties
  - We have demonstrated that applying the 2-D fitting improves over the raw **`wavdetect`** results, and are currently working to improve the robustness of the 2-D fitting implementation

- *We do not impose the assumption that the detected source is a point-source when performing the 2-D fit, in order to ensure better characterization of close double and extended sources*

How Does A 2-D Spatial Fit Help?

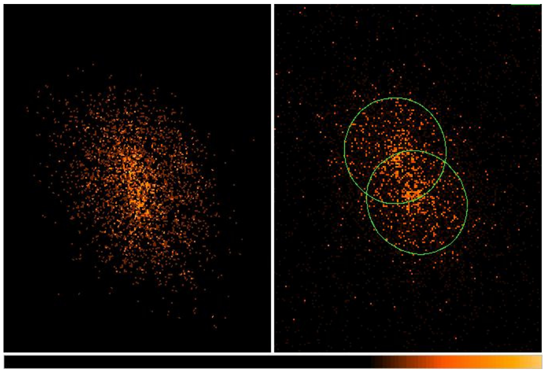
- Issues are primarily for sources detected well off-axis
  - Closely separated sources may not be reliably distinguished
    - Missed by `wavdetect`
      - Simulations show that ~50% of close sources pairs (<4” separation) are missed for off-axis angles $\theta > 5'$
    - Detected by `wavdetect` but eliminated when merging results from multiple wavelet scales
    - At 20' off-axis, ~35% of sources overlap at the 90% encircled energy radius for a density of 35 uniformly distributed sources per field
  - PSF substructure detected as distinct sources by `wavdetect`
- Instrumental features can result in detection of false sources
  - Such features often are often narrower than the size of the PSF and can be discriminated against by 2-D fit

Closely Separated Sources



- Source region of centered source is the intersection of the included and excluded green regions
- Corresponding background region is the intersection of the included and excluded yellow regions and excludes the green regions
- Central source appears to be double, but “upper” source was not identified in detected source list
- 2-D spatial fitting may recover such sources

Single Off-Axis Source



- *Right:* Single point source in ρ Oph observed at 15' off-axis is detected as two sources by **`wavdetect`**
- *Left:* Point spread function generates using SAO-SAC ray-trace at same location demonstrates that the source is single
  - 2-D fitting with a PSF model would correctly fit a single source

2-D Spatial Fit Steps I

- 2-D spatial fit steps
  - Fit the data with two source models:
    - A single PSF model appropriate for the location on the detector plus a constant background
    - A single PSF model appropriate for the location on the detector, convolved with an elliptical Gaussian, plus a constant background
      - The starting guesses for the fits are determined by 1-D fits to the marginal distributions
  - If the first model is a better fit, then the source is a single, isolated point source
  - If the second model is a better fit, then the source is either extended, or may be a close double

2-D Spatial Fit Steps II

- 2-D spatial fit steps (continued)
  - In the latter case, we further perform a fit with a source model that includes two point sources plus an elliptical Gaussian
  - If the fit with two point sources is better, then the “source” is a close double and we catalog both sources and extract the source properties separately (to the extent possible)
  - If a satisfactory fit is still not achieved, then the source is flagged as “complex”
- The fit results in source position coordinates and errors, the parameters and parameter errors for the elliptical Gaussian components, and the integrated fluxes and flux errors for the source and background components

Time Variability

- Identify sources that are potentially, or certainly, variable so a user can select them for further analysis
  - Objective is not to analyze the time variability in detail
- Per-observation source analysis applies variability criteria within a single observation (inter-observation variability analysis performed as part of merge processing)
  - Can be performed in counts space
  - Construct a variability index based on a Bayesian analysis yields odds ratio and multi-resolution light curve
    - Uses Gregory-Loredo algorithm
      - More computationally efficient than a Bayesian Blocks algorithm, and requires fewer assumptions
  - Compute the sigma scatter of points in the light curve around the average for use in the merge pipeline determination of inter-observation variability

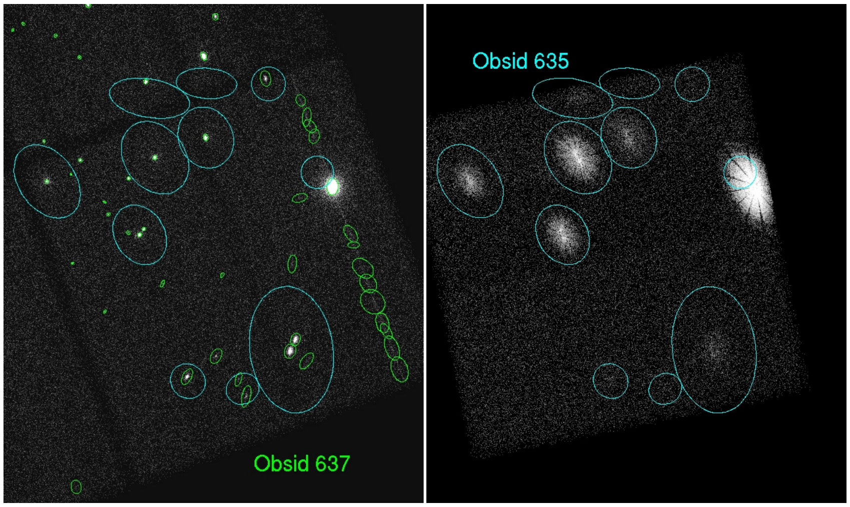
Spectral Model Fitting

- For ACIS data, fit a spectral model to the source spectrum
  - Source and background regions derived from source detection
  - Compute the weighted ARF and RMF at source location on detector
  - Correct the ARF for the fraction of the PSF outside of the source region as a function of energy
  - Initially fit broad-band data with $F_\nu \sim \nu^{-1}$ power law plus Galactic absorption where N_H column is determined from existing surveys (fitting normalization only)
  - If sufficient counts (TBD), in addition fit the spectral index
  - If sufficient counts (TBD), repeat the fit in the soft, medium, and hard bands separately
- The fitted model spectrum is used to convert flux in photons cm^{-2} s^{-1} to ergs cm^{-2} s^{-1}

Merge Processing

- Generates master objects for sources identified during per-observation processing of a new dataset
  - If a new per-observation source matches any per-observation sources currently in the catalog, redetermines the master object source properties for the matched source by combining information from the individual per-observation datasets
  - Result of merge must be independent of processing order
- The merge pipeline is currently in the definition stage
  - The first implementation will construct merged source properties through weighted averages or taking the properties from the “best” observation
  - Subsequent implementations will merge observations that are obtained using the same (α, δ) and roll angle (within a predetermined tolerance) prior to computing source properties

Merging Different Observations

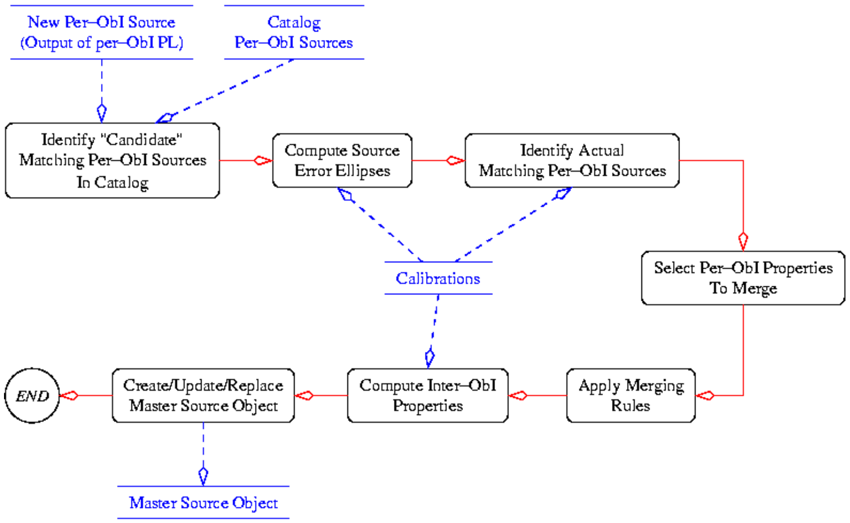


Merge Processing Overview I

The following information is subject to change based on analysis of performance on simulated and real datasets

- Sources that are observed in multiple observations will be considered to be the same source if their deconvolved source position ellipses overlap
  - Determined by combining the extent of the fitted source region (including errors) with the fitted position uncertainty ellipse

Merge Processing Overview II



New Per–ObI Source
(Output of per–ObI PL)

Catalog
Per–ObI Sources

Identify "Candidate"
Matching Per–ObI Sources
In Catalog

Compute Source
Error Ellipses

Identify Actual
Matching Per–ObI Sources

Calibrations

Select Per–ObI Properties
To Merge

Apply Merging
Rules

Compute Inter–ObI
Properties

Create/Update/Replace
Master Source Object

END

Master Source Object

Merge Processing Overview III

- Possible combinations of sources
  - Simple case
    - Single “old” source matches single “new” source
      - Master source information is updated
  - Complex cases
    - Single “old” source, multiple “new” sources
      - Old master source may be “retired” and multiple new master sources are added to the catalog
    - Multiple “old” sources, single “new” source
    - Not all source properties can be updated unambiguously for each of the multiple sources
      - If there is ambiguity a “confused source” flag will be set

Merge Processing Overview IV

- Examples of master source properties determined by averaging properties measured in individual observations:
  - Source position
    - Average of the individual positions derived from all the observations and bands, weighted by S/N
  - Band fluxes and source size
    - Error-weighted averages of the individual observation data
- Example master source property determined by taking the "best" properties measured in individual observations:
  - Spectral fit parameters
    - Taken from the individual observation with the highest source significance

Merge Processing Overview V

- Some master source properties must be determined by performing additional computations
  - For example, the master source variability information is a combination of the intra-observation variability properties for each observation (determined by the per-observation processing) with inter-observation variability properties that must be established by the merge processing
    - Inter-observation variability
      - Must be performed in flux space
      - Compare average count rates and sigma scatter and determine whether they are consistent with a constant flux density

User Interface Overview

- *Detailed UI requirements are currently being developed*
- UIs will provide ability to perform queries on data tables and retrieve matching rows and associated objects
  - Retrieval formats to be evaluated, but will include VO compliant formats and FITS format data files that are compatible with the *CIAO* and other X-ray data analysis software
- Future enhancements
  - Virtual column support
  - Query API to enable user-written scripts to perform queries and retrieve/analyze data
- Plan to support VO compliant query language (ADQL) and workflows when appropriate standards are established