Chandra Source Catalog Merge Pipeline Specification

December 7, 2006

Ian Evans (ievans@cfa.harvard.edu)

Inputs:

- TBD data for each source-by-observation object associated with the current ObI.
- TBD data for candidate matching source-by-observation objects associated with other ObIs.
    - Candidate matching source-by-observation objects associated with other ObIs are identified using a cone search around the position of each source-by-observation object associated with the current ObI.
    - *We assume that any necessary updates are completed and any locking that is required to prevent race conditions is in place [how relevant is this concern?].*
- *If the source-by-observation objects associated with the current ObI are replacing the source-by-observation objects associated with a prior version of the current ObI* (*i.e.*, the current ObI has been reprocessed), TBD data for each source-by-observation object associated with the *prior version* of the current ObI.
- *In all cases, only source-by-observation objects that have passed all ObI level and source-by-observation level quality assurance checks [see part VI (Quality Assurance Requirements) of the* Chandra Source Catalog *Requirements document], and that satisfy the criteria identified in subsection 2.2 of section 2 (Criteria for Source Inclusion) of part I of the* Chandra Source Catalog *Requirements document, shall be included as inputs to the merge pipeline.*

Outputs:

- TBD data in parameter file format (TBR) for each master source object to update master source list.

Conceptual Algorithm:

- *Identify sets of matching sources (for reprocessing information, first see "Reprocessing" below)*

  For the set $\mathcal{S}$ of source-by-observation objects associated with the current ObI:

  1. Identify the sets $\mathcal{M}1$, $\mathcal{M}2$, ... of candidate matching source-by-observation

objects associated with other ObIs (where the suffixes 1, 2, ... identify each individual ObI) by performing cone searches of radius r around the world coordinates of each member of the set $\mathcal{S}$. For convenience, we will designate the union of the sets $\mathcal{M}1$, $\mathcal{M}2$, ... as the set $\mathcal{M}$.

- The value of the parameter r shall be chosen to be "large enough" to ensure that the neighbors-of-neighbors search will succeed.

2. Compute the *overlap ellipses* for each member of the sets $\mathcal{S}$ and $\mathcal{M}$.

   - We equate the overlap ellipse with the source position error ellipse, scaled by a fudge factor parameter of order unity.

     - The fudge factor parameter is likely to be a function of instrument and off-axis angle. The initial value should be unity everywhere.

     - Ideally we would equate the overlap ellipse with the *source ellipse*, which is the convolution of the deconvolved source size with its associated uncertainty and the source position error ellipse. This is not done because we expect that the deconvolved source size and associated uncertainty will not be well determined in the general case.

     - *We* assume *(a) that most sources will be compact* and *(b) that the source position error ellipse of any single off-axis detection that resolves into multiple sources on-axis will overlap the source position error ellipses of the on-axis detections.* This assumption must be verified experimentally.

   - Two sources are deemed to "overlap" *if and only if* the spatial intersection of their overlap ellipses is not empty.

3. *When identifying matching sources, multiple source-by-observation objects associated with a single ObI shall always be treated as distinct sources, even if they overlap.*

4. Identify the subset $\mathcal{N}$ of the set $\mathcal{S}$ of source-by-observation objects associated with the current ObI that do not overlap any member of the set $\mathcal{M}$ of candidate matching source-by-observation objects associated with other ObIs. $\mathcal{N}$ *comprises the set of source-by-observation objects that must be added to the master source list as new master source objects [assuming that they satisfy the criteria identified in subsection 2.1 of section 2 (Criteria for Source Inclusion) of part I of the* Chandra Source Catalog *Requirements document].*

5. Identify all members of the set union $\{ \mathcal{S}, \mathcal{M} \}$ that comprise *(partial) unambiguous matches*. A (partial) unambiguous match is identified between a source a that is observed in ObI 1 and a source b that is observed in ObI 2 ($1 \neq 2$) *if and only if* (1) a overlaps b; (2) a does not overlap any other source observed in ObI 2; and (3) b does not overlap any other source observed in

ObI *1* .

6. Identify all members of the set of (partial) unambiguous matches that comprise the subset $\mathcal{U}$ of *unambiguous matches*. An unambiguous match is identified between sources *a* , *b* , *c*, ... that are observed in ObIs *1* , *2* , *3* , ... when all pairs ([*a* , *b*], [*a* , *c*], [*a* , ...], [*b* , *c*], [*b* , ...], ...) of sources are individually (partial) unambiguous matches. $\mathcal{U}$ *comprises the set of source-by-observation objects that correspond to master source objects already present in the master source list that must be merged and updated.*

7. Identify all members of the set union { $\mathcal{S}$, $\mathcal{M}$ } that comprise the subset $\mathcal{C}$ of *ambiguous (confused) matches.* An ambiguous (confused) match is identified when a source *a* that is observed in ObI *1* overlaps multiple sources *b* , *c* , ... that *either* (a) are observed in a single ObI *2* , *or* (b) consist of (partial) unambiguous matches such that there is at least one ObI that is common amongst the (partial) unambiguous matches for all of the sources *b* , *c* , ... $\mathcal{C}$ *comprises the set of ambiguous (confused) source-by-observation objects that must be* flagged as confused*, and* linked *but not* merged *to the corresponding master source objects in the master source list.*

    - Note that in the case (b) above, source *a* is not required to overlap all of the individual sources that comprise each of the (partial) unambiguous matches. For example, *if* source *a* , observed in ObI *1* , overlaps sources *b* and *c* , observed in ObIs *2* and *3* , respectively, *and* source *c* is a (partial) unambiguous match with source *b1* observed in ObI *2* , *then* source *a* is an ambiguous (confused) match with sources *b* and *c* / *b1* .

8. Once the set $\mathcal{C}$ of ambiguous (confused) source-by-observation objects is identified, steps 4–6 should be re-applied to the set { $\mathcal{S}$, $\mathcal{M}$ } - $\mathcal{N}$ - $\mathcal{U}$ - $\mathcal{C}$ to identify additional members of the sets $\mathcal{N}$ and $\mathcal{U}$ that were previously missed because they were overlapped by one or more ambiguous (confused) source-by-observation objects.

9. At this point, the set { $\mathcal{S}$, $\mathcal{M}$ } - $\mathcal{N}$ - $\mathcal{U}$ - $\mathcal{C}$ should consist of source-by-observation objects that overlap at least two other sources that were observed in different ObIs and that do not overlap each other. For the time being we designate members of this set $\mathcal{H}$ as *ambiguous (too hard) matches*. More sophisticated algorithms or manual intervention will be required to disambiguate the source-by-observation object links.

    - We need to gather statistics about the rate of ambiguous (too hard) matches, as well as assess patterns that may be amenable to automated resolution by improved algorithms. For the time being, the appropriate linkages for members of the set $\mathcal{H}$ will need to be determined by manual intervention.

- *Create/modify master source list entries*

    1. Each member of the set $\mathcal{N}$ corresponds to a single source observed only in the current ObI (and therefore should not be linked to any existing master source objects in the master source list). Each member of the set $\mathcal{N}$ that satisfies the criteria identified in subsection 2.1 of section 2 (Criteria for Source Inclusion) of part I of the *Chandra Source Catalog* Requirements document will result in the creation of a new master source object, with parameters *merged* from that single source-by-observation object observation.

    2. Each member of the set $\mathcal{U}$ corresponds to a single source observed in the current ObI and in one or more previously processed ObIs. The current source-by-observation object should not be linked to any existing master source objects in the master source list, and all of the previously processed source-by-observation objects should be linked to the same existing master source object. Each member of the set $\mathcal{U}$ will result in the modification of an existing master source object, with parameters *merged* from the all of the source-by-observation object observations.

    3. Each member of the set $\mathcal{C}$ corresponds to a source-by-observation object that matches multiple distinct sources observed in either the current or previously processed ObIs. These source-by-observation objects may be linked to zero or more existing master source objects in the master source list. Each member of the set $\mathcal{C}$ will be *marked as confused*, and the corresponding source-by-observation objects must be *linked* to the master source objects of each distinct source that is overlapped. *If* the corresponding source-by-observation objects were previously *merged* into any master source objects, *then* the confused source-by-observation objects must be *unmerged* from those master source objects. *If* this last step would remove the only source-by-observation object corresponding to a master source object, *then* the master source object must be *deleted*.

    4. Populate the data quantities that comprise each new/revised master source object. Unless otherwise specified, each quantity should be populated from the equivalent quantities extracted from the set of source-by-observation objects that are to be *merged* (not just *linked*) into the master source object, using the specified rules. Values not determined (for example, because none of the source-by-observation objects have spectral fits, or because there are no HRC source-by-observation objects to populate HRC "**`W`**" energy band fluxes) shall be populated with an identifiable **`NULL`** value (TBR).

        - These quantities should be populated with the error-weighted means of the equivalent quantities extracted from the source-by-observation objects to be merged. Source-by-observation objects for which the value of **`pileup_warning`** is > TBD (for ACIS observations) should be

excluded from the merge *unless* all source-by-observation objects to be merged have **pileup_warning** > TBD.

ra, **dec**.

- These quantities should be populated with the error-weighted means of the equivalent quantities extracted from the source-by-observation objects to be merged. For ACIS science energy bands, source-by-observation objects for which the value of **pileup_warning** is > TBD should be excluded from the merge *unless* all ACIS source-by-observation objects to be merged have **pileup_warning** > TBD. In the latter case, set **pileup_flag** = **TRUE**, otherwise set **pileup_flag** = **FALSE**.

  photflux_aper, **photflux_aper90**,
  flux_aper, **flux_aper90**,
  flux_powlaw_aper, **flux_powlaw_aper90**,
  flux_bb_aper, **flux_bb_aper90**.

- These quantities should be populated by propagating the errors of the equivalent quantities extracted from the source-by-observation objects that are included in the merge.

  err_ellipse_r0, **err_ellipse_r1**,
  err_ellipse_ang,
  photflux_aper_err, **photflux_aper90_err**,
  flux_aper_err, **flux_aper90_err**,
  flux_powlaw_aper_err, **flux_powlaw_aper90_err**,
  flux_bb_aper_err, **flux_bb_aper90_err**.

- These quantities should be populated with the equivalent quantities extracted from the ACIS source-by-observation object that has the largest value of **flux_significance** and for which **pileup_warning** ≤ TBD. *If* no ACIS source-by-observation objects have **pileup_warning** ≤ TBD, *then* the ACIS source-by-observation object that has the largest value of **flux_significance** should be used.

  alpha, **flux_powlaw**, **nh_powlaw**,
  alpha_err, **flux_powlaw_err**, **nh_powlaw_err**,
  kt, **flux_bb**, **nh_bb**,
  kt_err, **flux_bb_err**, **nh_bb**, **nh_bb_err**.

- These quantities should be populated as specified.

  If the master source object has previously been released to a catalog view, *then* the existing value of **name** shall be used. Otherwise:

  name = **"CXO J<hhmmss.s><+ddmmss>"**

where **<hhmmss.s>** = **ra** in hhmmss.s format, **<+ddmmss>** signed **dec** in [+/-]ddmmss format, where the last digits **<hhmmss.s>** and **<+ddmmss>** are *truncated* (*not* rounded).

gal_b, **gal_l** = Galactic latitude, longitude corresponding to (**ra**, **dec**).

extent_flag = **TRUE** if all of the values of **extent_flag** extracted from the source-by-observation objects to be merged are non-zero, otherwise **FALSE**.

conf_flag = **TRUE** if any of the values of **conf_flag** extracted from the source-by-observation objects to be merged have coded values 1 (Multiple sources identified within the source detection region) or 2 (Source detection region overlaps another source detection region) set, *or* if any source-by-observation objects to be *linked* are members of the sets $\mathcal{C}$ or $\mathcal{H}$, otherwise **FALSE**.

var_flag = **FALSE** if all of the values of **var_flag** extracted from the source-by-observation objects to be merged are zero *and* the computed value of the Inter-observation Variability Index (**var_inter_index**, see below) is zero, otherwise **FALSE** (TBR).

significance = The maximum value of **flux_significance** extracted from all science energy bands of the source-by-observation objects to be merged.

ks_intra_prob = The maximum values of **ks_prob** extracted from each science energy band of the source-by-observation objects to be merged.

kp_intra_prob = The maximum values of **kp_prob** extracted from each science energy band of the source-by-observation objects to be merged.

var_intra_prob = The maximum values of **var_prob** extracted from each science energy band of the source-by-observation objects to be merged.

var_intra_index = The maximum values of **var_index** extracted from each science energy band of the source-by-observation objects to be merged.

var_intra_sigma = The maximum values of **var_sigma** extracted from each science energy band of the source-by-observation objects to be merged.

- These quantities should be populated with the Kaplan-Meier medians of the equivalent quantities extracted from all science energy bands of the source-by-observation objects to be merged:

  major_axis, **minor_axis**, **pos_angle**.

- These quantities should be populated by propagating the errors of the equivalent quantities extracted from all science energy bands of the source-by-observation objects to be merged:

  major_axis_err, **minor_axis_err**, **pos_angle_err**.

- These quantities should be computed as specified:

  hard_<x><y> =
  [**flux_aper**(**<x>**) - **flux_aper**(**<y>**)] / **flux_aper**(**b**)

  for **<x><y>** = **hm**, **hs**, **hu**, **ms**, **mu**, **su**.

  hard_<x><y>_err = propagated error from corresponding **hard_<x><y>** computation.

  nh_gal = Computed with **COLDEN** using **NRAO** option (Dickey & Lockman 1990) at (**ra**, **dec**).

- These quantities should be computed as specified in section 1.1.7.2 (Inter-observation Source Variability) of part I of the *Chandra Source Catalog* Requirements document.

  var_inter_prob, **var_inter_index**, **var_inter_sigma**.

- These quantities should be populated with values that record the total number of ObIs with the corresponding telescope/instrument configurations that are represented by the set of source-by-observation objects that are *linked* to the master source object:

  acis_num, **acis_hetg_num**, **acis_letg_num**, **hrc_num**, **hrc_hetg_num**, **hrc_letg_num**.

- These quantities should be populated with values that record the total good time in seconds of ObIs with the corresponding telescope/ instrument configurations that are represented by the set of source-by-observation objects that are *linked* to the master source object:

  acis_time, **acis_hetg_time**, **acis_letg_time**, **hrc_time**, **hrc_hetg_time**, **hrc_letg_time**.

- *Reprocessing*

  The reprocessing case is similar to the normal case, except that in addition, the set of

source-by-observation objects associated with the prior version of the current ObI must be *unmerged* and *unlinked* from any master source objects as part of the merge processing.

1. Each source-by-observation object associated with the current version of the current ObI is matched to zero or more source-by-observation objects associated with the prior version of the current ObI using a similar process to the steps above.

    - *If* more than one source-by-observation object associated with the current/prior ObI overlaps a source-by-observation object associated with the prior/current ObI, *then* the overlapping source-by-observation objects with the *closest* matching source positions shall be considered to be unambiguous matches.

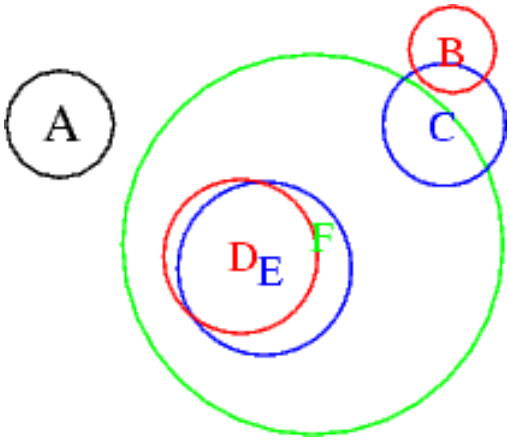
    - *If* a source-by-observation object associated with the prior ObI does not overlap any source-by-observation objects associated with the current version of the current ObI, *then* the master source object properties must be redetermined from the source-by-observation objects associated with other ObIs (if any) that are linked to the master source object. *If* unmerging the source-by-observation object associated with the prior version of the current ObI would remove the *only* source-by-observation object corresponding to a master source object, *then* the master source object must be *deleted*.

2. Once the set of source-by-observation objects associated with the prior version of the current ObI have been identified for unmerging and unlinking from any master source objects, then merge processing proceeds as specified above, with the following exception.

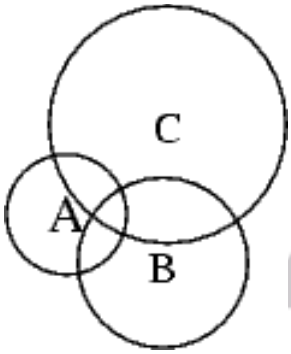
    - *If* a source-by-observation object associated with the current version of the current ObI is identified as belonging to the set $\mathcal{N}$ in merge processing, *and* if it is matched to a source-by-observation object associated with the prior version of the current ObI, *and if* the corresponding master source object has previously been released to a catalog view, *then* the value of **`name`** assigned to the master source object shall not be redetermined. This scenario corresponds to the common case of a source observed once in a single ObI that undergoes merge processing and is subsequently reprocessed.)

Source Matching Examples:

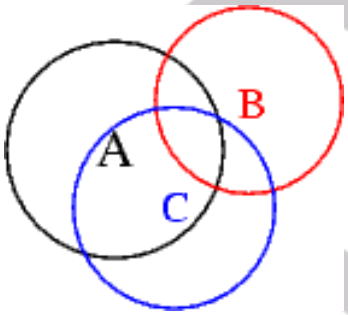
- In each example below, black circles represent source-by-observation object sources associated with the current ObI, while differently colored circles represent source-by-observation object sources associated with other ObIs (each ObI is a different color).



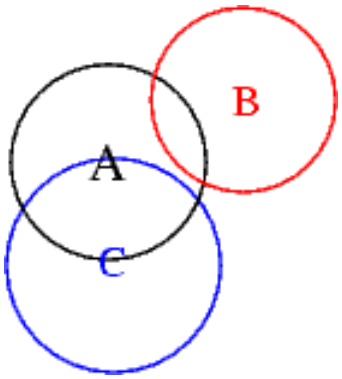
In this example, source A will be promoted to be a new master source, since it does not overlap any other source in any ObI. Sources B and C will be merged together to form a single master source, since they comprise an unambiguous match once source F [which is an ambiguous (confused) match] is accounted for. Sources D and E similarly comprise an unambiguous match. Source F is an ambiguous (confused) match, and will be *linked* but not *merged* into the two master source objects resulting from the B/C and D/E merges.



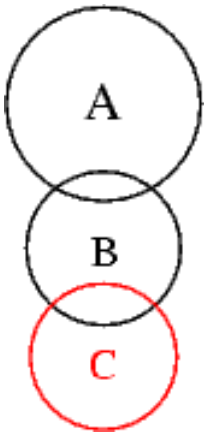
In this example, sources A, B, and C are distinct even though they overlap, because they all belong to the same ObI.



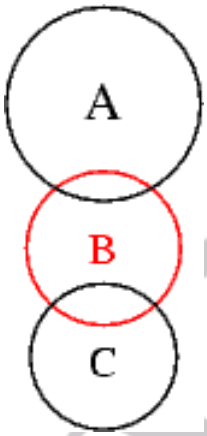
In this example, sources A, B, and C comprise an unambiguous match, since they all overlap each other and do not overlap any other sources.



In this example, source A is *not* an unambiguous match to sources B and C, since B and C belong to different ObIs and do not overlap. This is the ambiguous (too hard) case. Since sources A, B, and C all belong to distinct ObIs, there is no way to determine unambiguously how to handle this case (*e.g.*, A and B distinct, merge C with B? A and C distinct, B ambiguous (confused)?, B and C distinct, merge A with B?)



In this example, sources A and B are distinct, since they belong to the same ObI. Sources B and C form an unambiguous match.



In this example, sources A and C are distinct, since they belong to the same ObI. Source B is an ambiguous (confused) match. It will be *linked* but not *merged* to the two master sources that correspond to sources A and C, and so will not contribute to the properties of either master source. This is distinct from the ambiguous (too hard) case because we can identify A and C as distinct master sources.