CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

An X-ray Astrophysicist Looks at ObsCore

Ian Nigel Evans

Chandra X-ray Center

Center for Astrophysics | Harvard & Smithsonian

IVOA Interop

2024 Nov 16

ObsCore And High Energy Astrophysics Data

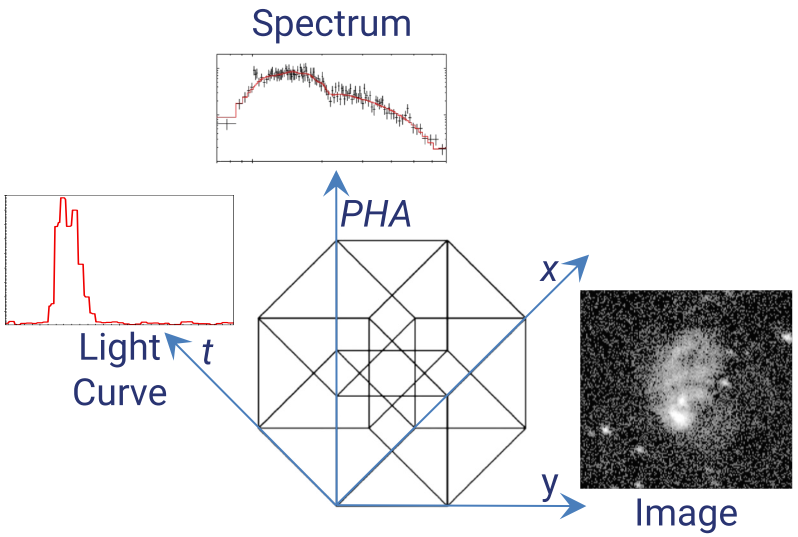
- *ObsCore* is an IVOA standard for data discovery — but how well does it work for high-energy astrophysics (HEA) data?
- Perspective of a working X-ray astrophysicist not (just) a data provider
  - Can I search for the data products I'm looking for effectively?
- Typical HEA experiments detect individual particles (*e.g.*, *Chandra* detects *X-ray photons*)
- Use *Chandra X-ray Observatory* data as an example to investigate
  - Focus on a few examples
- Two main categories of *Chandra* data products
  - Archival single-observation datasets
  - *Chandra Source Catalog (CSC)* data products
- *Chandra* science data products post telemetry decom are recorded primarily in FITS format

CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

The HEA Data Hypercube

- Each *event* records a (typically) 4-D set of observables that map to physical properties (*i.e.*, α, δ, t_{TT}, E)

- A set of events (*e.g.*, from an observation) is termed an *event list*



- Event list is an efficient way to store a sparse photon list
  - A typical *Chandra* observation stored as a non-sparse pixelated 4-D cube would require $O(10^{13})$ voxels

- We try not to pixelate the data until necessary for specific analysis
  - Select only the events of interest
  - Binning loses information — photon spatial positions are measured with subpixel resolution for *Chandra* instruments

- *Chandra* data analysis requires multiple additional data products

What’s An Observation?

- The ObsCore recommendation doesn’t define the term “observation”
- We define an observation in the traditional sense for individual *Chandra* archival observation data products
  - An observation is a single science exposure obtained with the telescope pointing at a target of interest
  - The longest possible single exposure duration for *Chandra* is ~190 ks

- ~50% of *Chandra* Source Catalog data products combine data from more than one *Chandra* observation
- ObsCore recommends treating these “advanced data products” as a new “observation”
  - Assign obs_id = stack_id for stack-based products and obs_id = name for master source-based products

Chandra Archival Observation Data Products

- ~25 types of data products, ~25–60 files per observation, depending on instrument, mode, and exposure
- ~25,000 observations in the current archive, so ~800,000 total files
- ***Typically downloaded as a package*** for an observation for data analysis
- Using ObsTAP we provide access to a tar package that includes these data products as a set for data analysis
- Most individual data products are not accessible via ObsTAP (only L2 event list and center, full images)

Data Product	dataproduct_type	Data Product	dataproduct_type
Photon event list (L1, L1.5, L2)	event	Exposure statistics (ACIS)	timeseries?
Images (center, full)	image	GTI filter	timeseries?
Bias images (ACIS)	image	Bad pixel regions	?
PHA spectrum (ACIS, HRC+TG)	spectrum	Mask	?
Aspect solution (+ OBC solution)	timeseries	Field of View	?
Aspect quality	timeseries	Parameter block (ACIS)	?
Ephemerides (spacecraft, lunar, solar)	timeseries	ARF (TG)	?
Mission time line	timeseries	RMF (TG)	?
Deadtime factors (HRC)	timeseries	V&V report and summary (PDF format)	?

Chandra Source Catalog Data Products

- ~38 types of data products
- ~90 million total files
- ***Generally used individually or multiples of same type*** for selected sources/detections
- Using ObsTAP we provide access to stack images only

27% per single observation
19% per observation stack
30% per observation detection
16% per observation stack detection
8% per source

Data Product	dataproduct_type	Data Product	dataproduct_type
Photon event list (obs, stack, obs det, stack det)	event	Aperture photometry MPDFs (obs det, stack det, src)	?
Images (obs, stack, obs det, stack det)	image	Detection fit MCMC draws (obs det, stack det)	?
Background images (obs, stack)	image	Bayesian blocks properties (src)	?
Exposure maps (obs, stack, obs det, stack det)	image	Detection list (stack)	?
Pixel mask (obs)	image	Extended source region (obs, src)	?
Point spread function (obs det)	image	Bad pixel regions (obs)	?
Limiting sensitivity (stack)	image	ARF (obs det)	?
PHA spectrum (obs det)	spectrum	RMF (obs det)	?
Light curve (obs det)	timeseries	Source region (obs det, stack det)	?
Aspect solution (obs)	timeseries	Field of View (obs, stack)	?
Aspect histogram (obs)	?		

Data Product Type

- ObsCore includes a limited set of dataproduct_type values
- ~50% of *Chandra* data products don't conform to existing data product type classifications
- Could set dataproduct_type = 'NULL' and use dataproduct_subtype, but this doesn't work so well for global data discovery

- Generic type "measurements" could be useful but is restricted by caveat

  Note that "measurements" extends the set of accepted values for dataproduct_type in ObsCore 1.0. This extension is meant to expose derived data products together with *the progenitor observation dataset.* (emphasis added)

  which is not desirable for CSC data products

- *Advanced data products would benefit greatly from a wider array of carefully selected data product types*

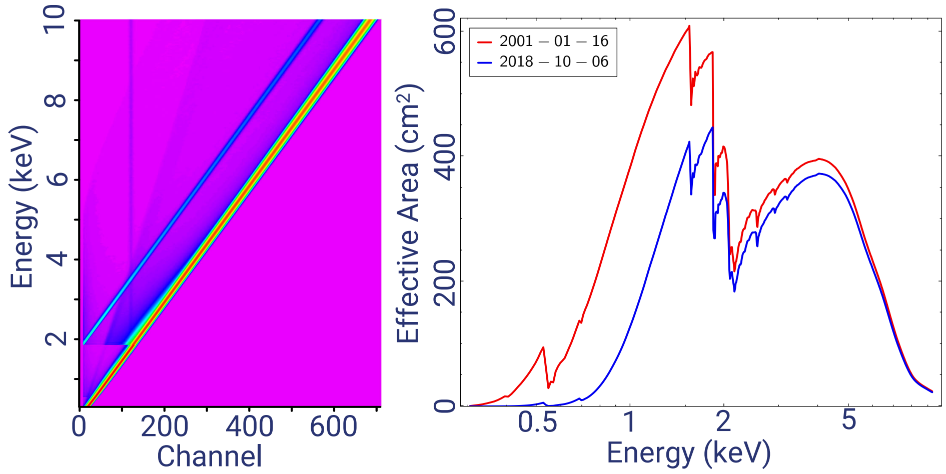
Calibration Level = 1.5

- ObsCore suggested classifications include
  - **Level 1**: Instrumental data in a standard format
  - **Level 2**: Calibrated, science ready data with the instrument signature removed

Calibration Level = 1.5

- ObsCore suggested classifications include
  - **Level 1**: Instrumental data in a standard format
  - **Level 2**: Calibrated, science ready data with the instrument signature removed

ACIS I-3 Aimpoint RMF and ARF



Left: RMF gives probability that a photon with a given energy will be detected in a given detector channel
Right: ARF gives effective area as a function photon energy
Both depend on location on the detector and observation epoch

- Calibrated HEA event lists typically include calibrated event spatial positions and times and are considered "science ready"
- However, the spectral axis typically does not have the instrument signature removed
  - Mapping from energy to PHA channel is probabilistic and depends on the responses
  - For *Chandra*, the ARF for an energy band depends on the (unknown) source spectrum and the RMF depends on selection of events (because of spacecraft dither)

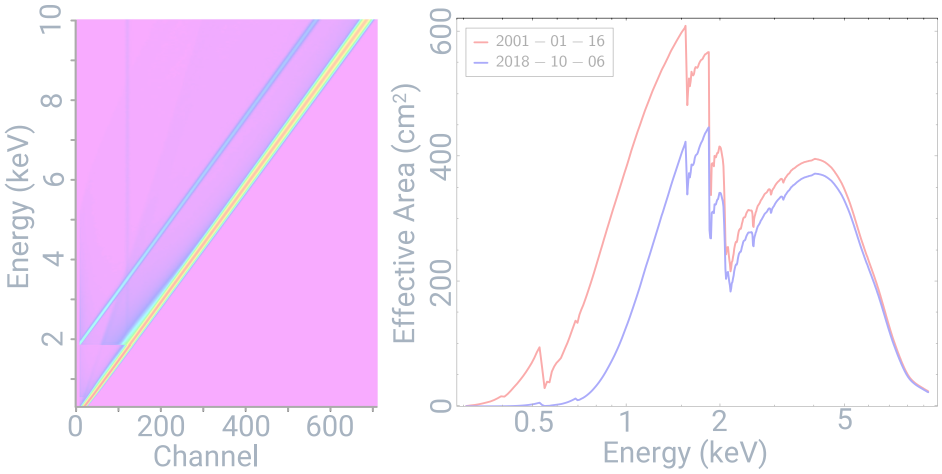
⇒ *Needs input from scientist*

CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

Calibration Level = 1.5

- ObsCore suggested classifications include
  - **Level 1**: Instrumental data in a standard format
  - **Level 2**: Calibrated, science ready data with the instrument signature removed

ACIS I-3 Aimpoint RMF and ARF



Left: RMF gives probability that a photon with a given energy will be detected in a given detector channel
Right: ARF gives effective area as a function photon energy
Both depend on location on the detector and observation epoch

- Calibrated HEA event lists typically include calibrated event spatial positions and times and are considered "science ready"

- However, the spectral axis typically does not have the instrument signature removed
  - Mapping from energy to PHA channel is probabilistic and depends on the responses
  - For Chandra, the ARF for an energy band depends on the (unknown) source spectrum and the RMF depends on selection of events (because of spacecraft dither)

  ⇒ *Needs input from scientist*

- For *Chandra*, we *choose* to set calib_level = 2 for these event lists

CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

Observable Axes o_ucd(s)

- Unlike an image whose observable is the quantity stored in each pixel, event lists typically include *multiple* observables for each event
    - HEA event lists include one event per detected particle, and many record a spatial position (2 axes), a time, and a spectral measure
    - *Chandra* event lists include many more than these 4 columns (*e.g.*, additional coordinate systems such as chip or detector, event grade, event status information, …)
    - Like event lists, HEA data products are often recorded as FITS BINTABLES (and possibly multi-HDU BINTABLES) so the presence of multiple observables in a single data product is not uncommon

Observable Axes o_ucd(s)

- Unlike an image whose observable is the quantity stored in each pixel, event lists typically include *multiple* observables for each event
    - HEA event lists include one event per detected particle, and many record a spatial position (2 axes), a time, and a spectral measure
    - *Chandra* event lists include many more than these 4 columns (*e.g.*, additional coordinate systems such as chip or detector, event grade, event status information, …)
    - Like event lists, HEA data products are often recorded as FITS BINTABLES (and possibly multi-HDU BINTABLES) so the presence of multiple observables in a single data product is not uncommon

- The ObsCore recommendation in this case is that o_ucd be left NULL unless a specific axis should be highlighted
    - This is not very satisfactory because it hides the details of the data content — some HEA experiments may not have spectral resolution, or may only have a single spatial axis, or may measure polarization, …

CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

Observable Axes o_ucd(s)

- Unlike an image whose observable is the quantity stored in each pixel, event lists typically include *multiple* observables for each event
  - HEA event lists include one event per detected particle, and many record a spatial position (2 axes), a time, and a spectral measure
  - *Chandra* event lists include many more than these 4 columns (*e.g.*, additional coordinate systems such as chip or detector, event grade, event status information, …)
  - Like event lists, HEA data products are often recorded as FITS BINTABLES (and possibly multi-HDU BINTABLES) so the presence of multiple observables in a single data product is not uncommon

- The ObsCore recommendation in this case is that o_ucd be left NULL unless a specific axis should be highlighted
  - This is not very satisfactory because it hides the details of the data content — some HEA experiments may not have spectral resolution, or may only have a single spatial axis, or may measure polarization, …

- *HEA would benefit greatly from a way to represent the presence of multiple observable axes*

Spectral Bounds em_min, em_max

- HEA typically expresses spectral quantities in units of *eV* (*keV*, *MeV*, *GeV*, *TeV*)
  - Units of *m* are very HEA-unfriendly

- The radio extension proposed recommendation includes example use cases in ADQL like

```
… WHERE 299792458 / em_max > 1.0e+9
```

  taking advantage of $\nu = c / \lambda$ where everyone knows the exact value of *c* in units of $m\ s^{-1}$

Spectral Bounds em_min, em_max

- HEA typically expresses spectral quantities in units of *eV* (*keV*, *MeV*, *GeV*, *TeV*)
  - Units of *m* are very HEA-unfriendly
- The radio extension proposed recommendation includes example use cases in ADQL like

```
… WHERE 299792458 / em_max > 1.0e+9
```

taking advantage of $\nu = c / \lambda$ where everyone knows the exact value of *c* in units of m s^{-1}

- Can we do the same with $E = hc / \lambda$?
- Let's do an experiment — who can tell me the value of *hc* in units of eV m ?

Spectral Bounds em_min, em_max

- HEA typically expresses spectral quantities in units of *eV* (*keV*, *MeV*, *GeV*, *TeV*)
  - Units of *m* are very HEA-unfriendly
- The radio extension proposed recommendation includes example use cases in ADQL like

  ```
  … WHERE 299792458 / em_max > 1.0e+9
  ```

  taking advantage of $\nu = c / \lambda$ where everyone knows the exact value of *c* in units of m s^{-1}

- Can we do the same with $E = hc / \lambda$?
- Let's do an experiment — who can tell me the value of *hc* in units of eV m ?

  ~1/806554.3937

- We should consider whether HEA-friendly values such as *energy_min*, *energy_max* would be preferable

CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

Time Bounds t_min, t_max

- ObsCore defines t_min (t_max) as the minimum (maximum) start time for data products that are combinations of multiple frames

- This definition may not be very useful for advanced data products
  - Some CSC data products have t_min to t_max spanning >20 yr (but very sparsely!)

- Can we encode (t_min, t_max) for the list of observations (others have suggested using TMOC)?
  - For HEA datasets a similar mechanism could be used to represent GTIs or STIs

Flexible Definitions

- Some ObsCore elements are expressly left to the data provider to decide what makes sense

- For other elements the level of flexibility is unclear

- Example: Central Coordinates
  - Section 4.10 defines (s_ra, s_dec) as the ICRS (RA, Dec) "… of the center of the observation"
  - Telescope pointing/optical axis? Where the best image quality is found
  - Instrument center? Instrument doesn't have to be centered in the FoV
  - What about cases where there are cut-outs (*e.g.*, windows) that are not centered on either of the above?
  - Appendix B.6.1.2 uses the wording "… used to identify a reference position (typically the center) of an observation …"
  - This is more flexible, and what we assume for *Chandra*

HEA Data May Be Different (1)

- For some HEA experiments many quantities are energy dependent (*e.g.*, s_fov, s_resolution, em_resolution, …) or depend on location within the FoV
  - Example: The *Chandra* PSF size varies by a factor ~50× across the FoV (and also depends on energy)
    - s_resolution is not very robust
    - s_resolution_min, s_resolution_max may be helpful
    - How do I query for datasets that have at least a certain spatial resolution at the location of my source?
- How do we associate the energy (or energy range) or off-axis angle that is relevant to the quantities to support queries?

- For non-pixelated data ObsCore recommends setting axes lengths <*x*>_xel to −1
  - The equivalent dimensionality for event lists is the *number of events*
  - This quantity is important for data discovery (scales as data size, and perhaps S/N)
- *Suggest adding* ev_number *for HEA data*

HEA Data May Be Different (2)

- Example: Data Product Type
    - ObsCore defines *spectrum* as “Any dataset for which spectral coverage is the primary attribute”
    - Great! Chandra PHA spectra meet this definition!
    - However, the IVOA data product types vocabulary defines *spectrum* as “Flux or magnitude as a function of spectral coordinates”
    - “Flux” is not defined but the standard astronomical definition of flux is energy flux (SI units W m^{-2})
    - In the optical/IR magnitude and energy flux density are tightly related $m = -2.5 \log f + const$
    - Chandra PHA spectra do not satisfy this definition: they are in units of counts (which may be mapped to a photon flux, but the actual photon energies are not determined)
    - The ObsTAP “List For Observables” describes “Flux” (phot.flux) as a “Photon Flux”, but then specifies units of W m^{-2}, which is an energy flux rather than a photon flux
    - None of these would help if the messenger is (*e.g.*) neutrinos instead of photons

- *How do we ensure that IVOA recommendations/definitions/… broadly cover the wide range of astronomical research and do not have unintentional biases for a particular waveband or type of data?*

Conclusions

- General
  - ObsCore recommendations work reasonably well for single *Chandra* observations but less so for advanced data products (especially those derived from multiple observations)
  - Limited set of dataproduct_type classifications don't map well to many types of data products
  - Event lists and many advanced products may include multiple observable axes
  - Recommendations and definitions should be flexible enough to enable data discovery for various wavebands and messengers

- HEA-specific
  - HEA event lists typically map to calibration level 1.5 but recommendation is flexible
  - Spectral units of *m* are not HEA friendly
  - Several elements don't work optimally for HEA data due to dependencies on energy etc.
  - Number of events is an important measure for event lists

dataproduct_type	obs_creator_did	s_dec	s_calib_status	t_stat_error	em_resolution
dataproduct_subtype	obs_release_date	s_fov	s_stat_error	em_xel	em_stat_error
calib_level	obs_publisher_did	s_region	s_pixel_scale	em_ucd	o_ucd
target_name	publisher_id	s_resolution	t_xel	em_unit	o_unit
target_class	bib_reference	s_xel1	t_refpos	em_calib_status	o_calib_status
obs_id	data_rights	s_xel2	t_min	em_min	o_stat_error
obs_title	access_url	s_ucd	t_max	em_max	pol_xel
obs_collection	access_format	s_unit	t_exptime	em_res_power	pol_states
obs_creation_date	access_estsize	s_resolution_min	t_resolution	em_res_power_min	instrument_name
obs_creator_name	s_ra	s_resolution_max	t_calib_status	em_res_power_max	proposal_id