Large-scale Data Projects / Research Data Alliance 5th Plenary

CHANDRA X-RAY OBSERVATORY

The *Chandra* Data Archive: Data Linking and Data Mining

Arnold Rots
Archive Astrophysicist
Smithsonian Astrophysical Observatory
8 March 2015

Summary

Challenges:

Collecting and curating the data

Linking data to publications and data to data

Metadata extraction

Data discovery

Finding exactly what you need

Issues:

Convincing people to submit data; compound data objects

Persistent identifiers; interoperability among repositories

Automated, complete, and correct

Where are the registries for heterogeneous repositories?

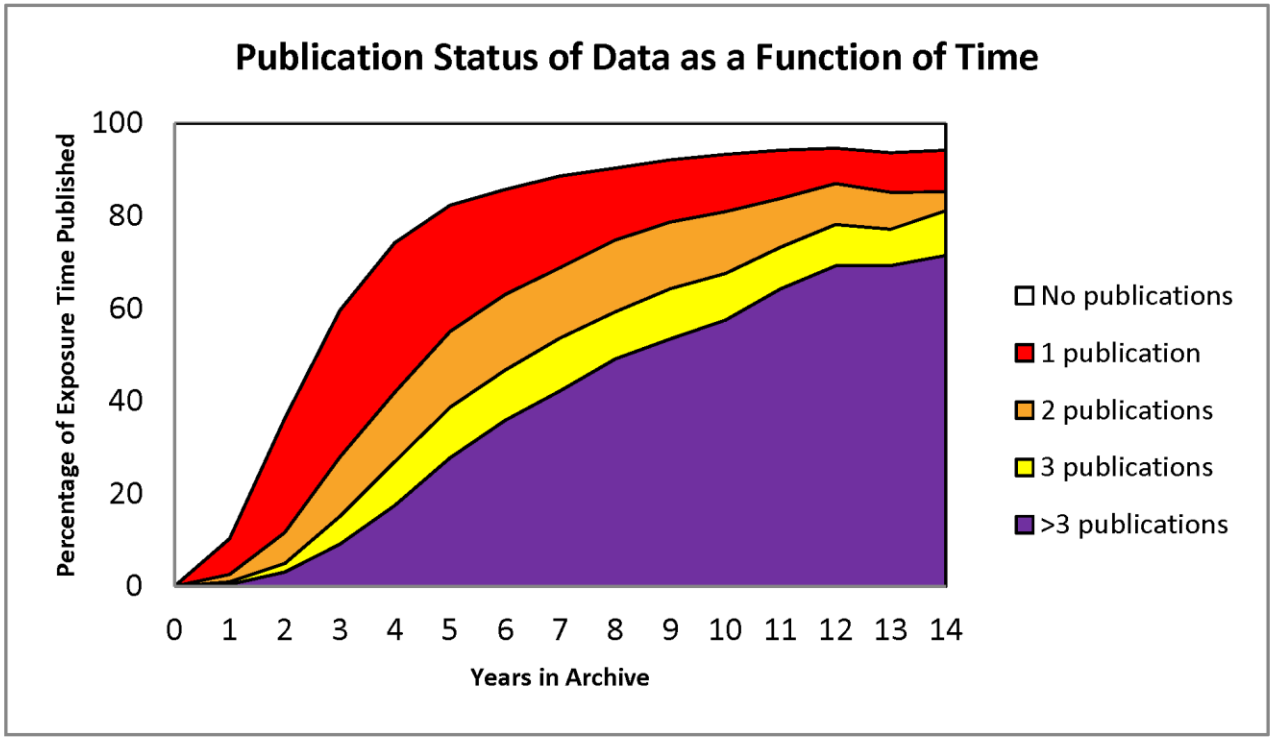
No more and no less: drilling down with identifiers

- *Chandra* X-ray Observatory
  - One of NASA's Great Observatories, sibling of the *Hubble* Space Telescope, launched in 1999
  - Operated by Smithsonian Astrophysical Observatory under contract with NASA
  - Covers the electro-magnetic spectrum from 0.5 to 10 keV (25 – 1.2 Å)
  - Spatial resolution better than 1 second of arc
  - In an eccentric 60-hour orbit that reaches ⅓ the way to the moon

The *Chandra* X-ray Observatory

- *Chandra* Data Archive
  - Contains all the mission's data for the full life-cycle of the data, Levels 0 to 3 of processing
  - Higher level (value-added) products may be contributed by users
  - Multiple reprocessing runs prompted by improved calibration information
  - Users have little interest in older data versions
  - Each dataset (observation) is tagged with a persistent identifier
  - Standard astronomical file format (FITS)

- Complete bibliography
  - Links datasets to publications (and vice versa)
  - Joins publication metadata with observation metadata
  - Research tool: simultaneous browsing of data archive and literature
  - Provides useful performance metrics beyond numbers of papers and citations
- Enabled by the Astrophysics Data System
  - Also operated by SAO
  - Indexes entire astronomical literature


Publication Status of Data as a Function of Time
Percentage of Exposure Time Published
100
80
60
40
20
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Years in Archive
No publications
1 publication
2 publications
3 publications
>3 publications

Collecting and Curating the Data: Requirements

- Collecting and curating the data
  - Convince people to submit data to persistent repositories
  - Managing a variety of data object types:
    - text, tables, graphs, images, multi-dimensional image cubes
  - Handling compound data objects
    - Multi-file datasets
    - Publication, table, row, number
  - Provenance

Collection and Curation Requirements: PIDs

- Linking data-to-publications and data-to-data
  - Use of persistent identifiers
    - Reduces the problem to linking one identifier to an other
  - Interoperability among repositories is required to make data-to-data linking meaningful
    - It seems obvious to require interoperability to be discipline-specific
    - But there is virtue in inter-disciplinary interoperability

Collection and Curation Requirements: Metadata

- Metadata extraction
  - Standardized metadata
    - Expected to be discipline-specific, but see previous comment
  - Automatic upon submission
    - … or it will become a mess
    - Extracting numeric metadata is tougher than extracting keyword-based metadata
  - Complete and correct
    - Goes without saying; it implies consistency
- This is a tough nut to crack, *imho*

Data Discovery: Issues

- Data discovery
  - Relatively easy for homogeneous (in content and format) repositories
    - These need a single record in a registry
  - There is no good infrastructure, yet, for heterogeneous repositories
    - Requires comprehensive registry services
  - Support for queries based on multi-dimensional numeric parameter spaces
    - Google does not work
    - Related to the issue noted in metadata extraction

Persistent Identifiers: Issues

- Data access with variable level of granulation
  - Meaning: no less *and* no more than what the user wants, e.g.:
    - A full article *vs* a table from the article *vs* a single table cell
    - A compound dataset *vs* a single file *vs* part of a file
  - People may not yet be asking for it, but it will be coming
  - Ability to expand identifiers dynamically
    - A fixed set if identifiers won't do
  - Parameterized fragments in PIDs?
- This is a challenge, *imho*

- Of particular interest in RDA
  - Metadata (particularly automated extraction of numeric metadata)
  - PIDs (and the ability to drill down)
  - Registries (especially for heterogeneous repositories)