DOI-Enabled Discovery and Credit: an ADS Perspective

Alberto Accomazzi
aaccomazzi@cfa.harvard.edu | *@aaccomazzi*
NASA Astrophysics Data System | *@adsabs* | *https://ui.adsabs.harvard.edu*

IVOA Interop | 5/10/2023

Context

ADS is primarily a literature database, and as such it does not aim to be an index for all research data products.

However, one of ADS's goals is to make relevant data products discoverable from the literature, whenever feasible.

Some types of data which are of most interest to ADS:

- Datasets "close" to publications, either as DBF, supporting archival links, or citations, as they supplement the science presented therein; examples include VizieR catalogs, text-mined Zenodo links, archival data links, data citations
- Reference catalogs, collections, and services, which are highly used and (possibly) cited; examples include 2MASS, WISE, CSC, etc.
- (to a somewhat lesser extent) Observations linked to proposals; examples include obs. proposals from JWST, CXO, XMM, etc.

What is indexed, linked in ADS

Indexed (an actual database record, searchable)

- The scholarly literature of interest to Astronomers
- VizieR records, IVOA standards, observing and funding proposals
- Software products: ASCL records, software packages cited via DOI
- Soon: cited data products, other research objects such as notebooks

Indexed records are curated scholarly research objects.
They are discoverable and citable via ADS, and their metrics are tracked

Linked (resource accessible from a record via a link)

- Data Products hosted by external collaborators (Archives, SIMBAD, NED)

Linked data collections can be used as a filter in ADS,
and to evaluate impact of linked data products

What's the difference

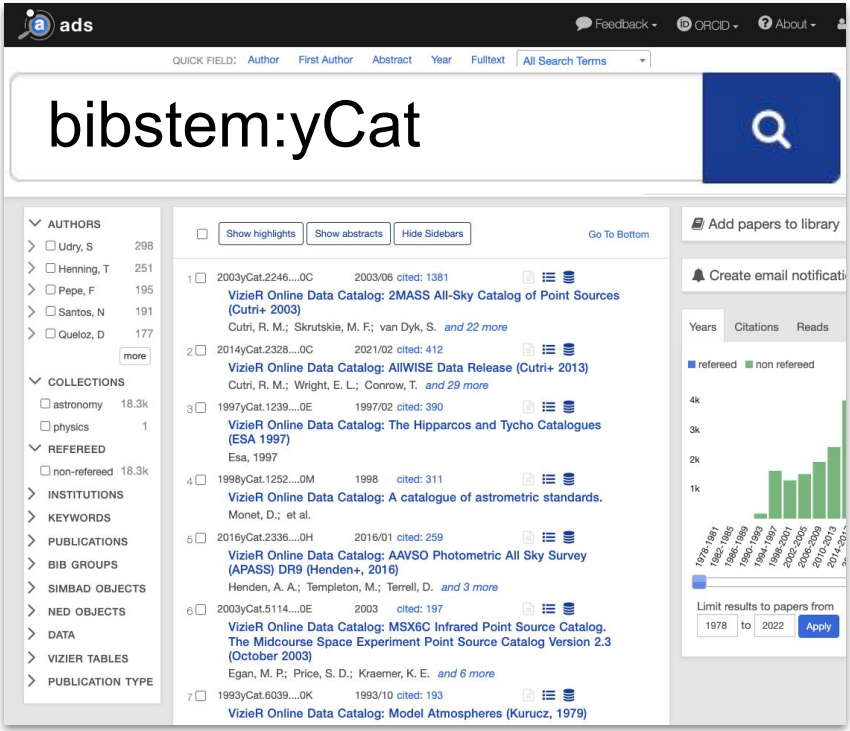
Indexed Dataset

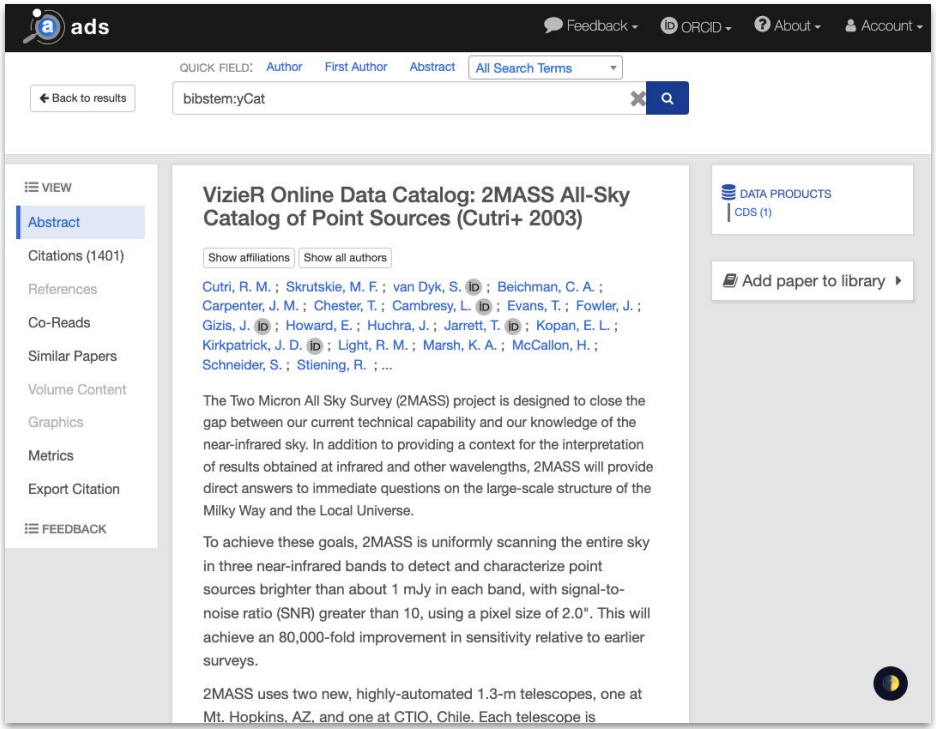
- ADS has a record corresponding to the dataset
- Dataset has higher level of discoverability (retrieved by e.g. ADS author search)
- Dataset has ADS metrics associated with it
- Data is accessible from paper via citation and data link

Linked Dataset

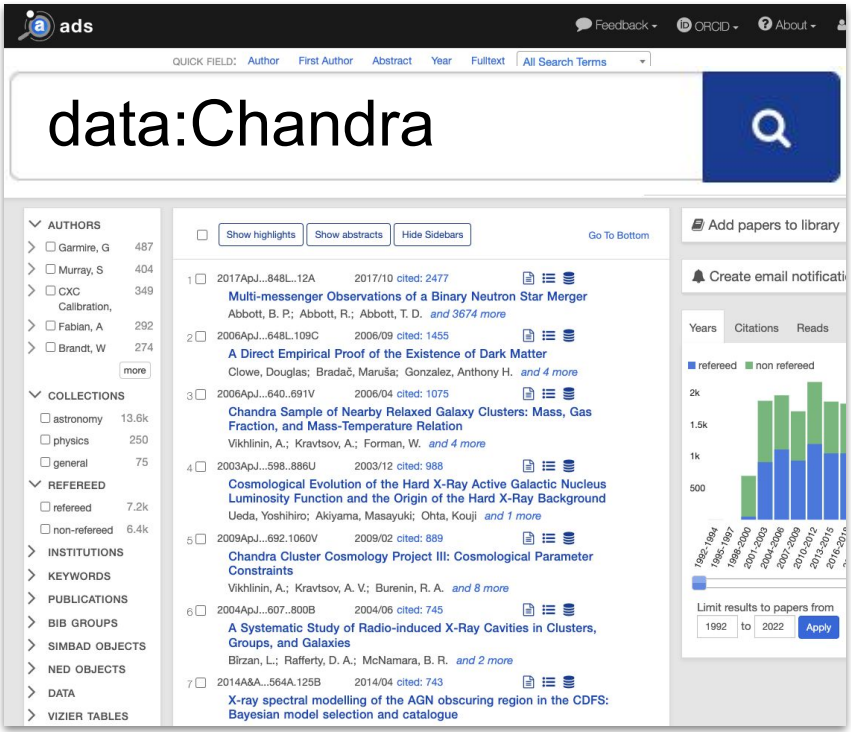
- ADS does not have a record corresponding to the dataset
- Papers associated with dataset typically part of a linked data collection (e.g. Chandra, IRSA, MAST)
- Only metrics available are via associated paper metrics
- Data is accessible from paper via data link

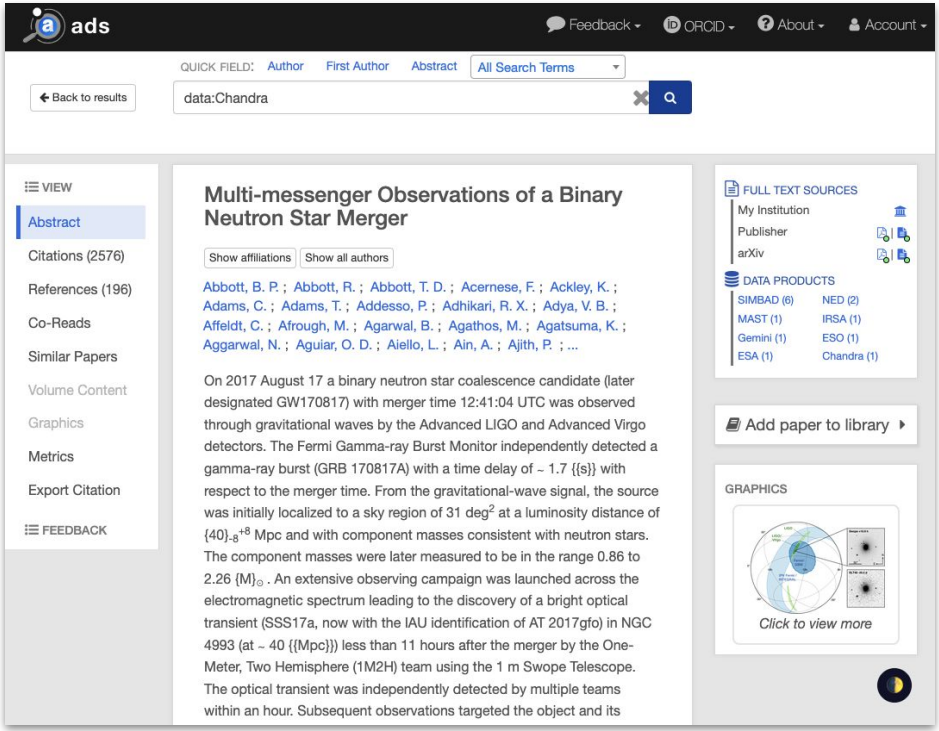
Indexed Dataset Example: Vizier records





Linked Dataset Example: Chandra data collection





Ingestion Policy (still evolving)

What is/will be indexed in ADS

- Curated, high-level datasets with good metadata (registered DOI) and clear authorship information
- Research data collections that have shown reuse value (initially via citations, i.e. `citation_count > 1`)

What will not be indexed in (but possibly linked from) ADS

- Every data product out there
- Every single version of a data product
- Data collections created for bundling purpose (e.g. MAST user generated DOIs)

The larger ecosystem of Discovery

Data Archives/users register data products using DOIs and rich metadata which includes provenance	→	ADS (and others) can discover and index/link data products via text mining of DAS/reference sections
Data Archives/editors provide clear instructions for acknowledgment (“cite as”/”acknowledge as”)	→	Authors/referees know how to cite/mention the relevant data products when writing papers
ADS provides discovery capabilities and metrics for linked/indexed data products	→	Data Archives can use the ADS to discover mentions/citations of their data and related impact metrics

Food for thought

- How does the data indexing & linking policy outlined here fit the needs of our community?
- How does it help you, as an data archivist / publisher / scientist?
- Is there a need for a disciplinary index of data products cited / mentioned in the literature beyond what is described here?

Backup Slides

Credit vs. Discovery

For ADS, credit means “whoever is listed in the author field”

- Contributor model for data products is much more complex than authorship for bibliographic records
  - 20 contributor types in DataCite schema
  - No appetite to replicate this in ADS, so some kind of mapping needed
  - No decisions have been made, waiting to see what the community adopts/decides for metadata registration
- Multiple implications, including first-authorship
  - What is the “right” order when flattening a set of lists?
  - Is “curator” an important role for discovery in ADS?
  - Note: ADS already has “publisher” and “editor” fields in its schema, but used with different semantics