

DOI-Enabled Discovery and Credit: an ADS Perspective

Alberto Accomazzi

aaccomazzi@cfa.harvard.edu | [@aaccomazzi](https://twitter.com/aaccomazzi)

NASA Astrophysics Data System | [@adsabs](https://twitter.com/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

The screenshot shows the VizieR search results page for the query 'bibstem:yCat'. The search bar at the top contains the query and a magnifying glass icon. Below the search bar, there are buttons for 'Show highlights', 'Show abstracts', and 'Hide Sidebars', along with a 'Go To Bottom' link. The main content area displays a list of search results, each with a checkbox, a title, a year, and a citation count. The results are:

- 1. 2003yCat.2246....0C 2003/06 cited: 1381
VizieR Online Data Catalog: 2MASS All-Sky Catalog of Point Sources (Cutri+ 2003)
Cutri, R. M.; Skrutskie, M. F.; van Dyk, S. *and 29 more*
- 2. 2014yCat.2328....0C 2021/02 cited: 412
VizieR Online Data Catalog: ALLWISE Data Release (Cutri+ 2013)
Cutri, R. M.; Wright, E. L.; Conrow, T. *and 29 more*
- 3. 1997yCat.1239....0E 1997/02 cited: 390
VizieR Online Data Catalog: The Hipparcos and Tycho Catalogues (ESA 1997)
Esa, 1997
- 4. 1998yCat.1252....0M 1998 cited: 311
VizieR Online Data Catalog: A catalogue of astrometric standards.
Monet, D.; et al.
- 5. 2016yCat.2336....0H 2016/01 cited: 259
VizieR Online Data Catalog: AAVSO Photometric All Sky Survey (APASS) DR9 (Henden+ 2016)
Henden, A. A.; Templeton, M.; Terrell, D. *and 3 more*
- 6. 2003yCat.5114....0E 2003 cited: 197
VizieR Online Data Catalog: MSX6C Infrared Point Source Catalog. The Midcourse Space Experiment Point Source Catalog Version 2.3 (October 2003)
Egan, M. P.; Price, S. D.; Kraemer, K. E. *and 6 more*
- 7. 1993yCat.6039....0K 1993/10 cited: 193
VizieR Online Data Catalog: Model Atmospheres (Kurucz, 1979)

On the left side, there is a navigation menu with categories like 'AUTHORS', 'COLLECTIONS', 'REFEREED', 'INSTITUTIONS', 'KEYWORDS', 'PUBLICATIONS', 'BIB GROUPS', 'SIMBAD OBJECTS', 'NED OBJECTS', 'DATA', 'VIZIER TABLES', and 'PUBLICATION TYPE'. A 'more' button is visible under the 'AUTHORS' section. On the right side, there are buttons for 'Add papers to library' and 'Create email notification'. A small bar chart shows the number of citations over time, with a legend for 'referred' (blue) and 'non referred' (green).

The screenshot shows the VizieR record page for the query 'bibstem:yCat'. The search bar at the top contains the query and a magnifying glass icon. Below the search bar, there are buttons for 'Back to results' and 'Show affiliations', 'Show all authors'. The main content area displays the record details for 'VizieR Online Data Catalog: 2MASS All-Sky Catalog of Point Sources (Cutri+ 2003)'. The record is categorized as 'Abstract' and has 1401 citations. The authors listed are Cutri, R. M.; Skrutskie, M. F.; van Dyk, S.; Beichman, C. A.; Carpenter, J. M.; Chester, T.; Cambresy, L.; Evans, T.; Fowler, J.; Gizis, J.; Howard, E.; Huchra, J.; Jarrett, T.; Kopan, E. L.; Kirkpatrick, J. D.; Light, R. M.; Marsh, K. A.; McCallon, H.; Schneider, S.; Stiening, R.; ... The abstract text reads: 'The Two Micron All Sky Survey (2MASS) project is designed to close the gap between our current technical capability and our knowledge of the near-infrared sky. In addition to providing a context for the interpretation of results obtained at infrared and other wavelengths, 2MASS will provide direct answers to immediate questions on the large-scale structure of the Milky Way and the Local Universe. To achieve these goals, 2MASS is uniformly scanning the entire sky in three near-infrared bands to detect and characterize point sources brighter than about 1 mJy in each band, with signal-to-noise ratio (SNR) greater than 10, using a pixel size of 2.0". This will achieve an 80,000-fold improvement in sensitivity relative to earlier surveys. 2MASS uses two new, highly-automated 1.3-m telescopes, one at Mt. Hopkins, AZ, and one at CTIO, Chile. Each telescope is

On the right side, there are buttons for 'DATA PRODUCTS' (CDS (1)) and 'Add paper to library'. On the left side, there is a navigation menu with categories like 'VIEW', 'References', 'Co-Reads', 'Similar Papers', 'Volume Content', 'Graphics', 'Metrics', 'Export Citation', and 'FEEDBACK'.

Linked Dataset Example: Chandra data collection

The screenshot shows the ADS search results for the query 'data:Chandra'. The search bar at the top contains the text 'data:Chandra' and a magnifying glass icon. Below the search bar, there are several filters and options: 'Show highlights', 'Show abstracts', 'Hide Sidebars', and 'Go To Bottom'. The main content area displays a list of search results, each with a checkbox, a citation key, a date, a citation count, and a title. The first result is 'Multi-messenger Observations of a Binary Neutron Star Merger' by Abbott, B. P.; Abbott, R.; Abbott, T. D. and 3674 more, dated 2017/10, cited 2477 times. Other results include 'A Direct Empirical Proof of the Existence of Dark Matter', 'Chandra Sample of Nearby Relaxed Galaxy Clusters: Mass, Gas Fraction, and Mass-Temperature Relation', 'Cosmological Evolution of the Hard X-Ray Active Galactic Nucleus Luminosity Function and the Origin of the Hard X-Ray Background', 'Chandra Cluster Cosmology Project III: Cosmological Parameter Constraints', 'A Systematic Study of Radio-induced X-Ray Cavities in Clusters, Groups, and Galaxies', and 'X-ray spectral modelling of the AGN obscuring region in the CDFS: Bayesian model selection and catalogue'. On the left side, there are navigation menus for 'AUTHORS', 'COLLECTIONS', 'REFEREED', 'INSTITUTIONS', 'KEYWORDS', 'PUBLICATIONS', 'BIB GROUPS', 'SIMBAD OBJECTS', 'NED OBJECTS', 'DATA', and 'VIZIER TABLES'. On the right side, there are options to 'Add papers to library' and 'Create email notifications'. A bar chart at the bottom right shows the number of papers over time, with a legend for 'referenced' (blue) and 'non referenced' (green).

The screenshot shows the abstract page for the paper 'Multi-messenger Observations of a Binary Neutron Star Merger'. The title is prominently displayed at the top. Below the title, there are options to 'Show affiliations' and 'Show all authors'. The abstract text begins with the authors' names: 'Abbott, B. P.; Abbott, R.; Abbott, T. D.; Acernese, F.; Ackley, K.; Adams, C.; Adams, T.; Addesso, P.; Adhikari, R. X.; Adya, V. B.; Affeldt, C.; Afrough, M.; Agarwal, B.; Agathos, M.; Agatsuma, K.; Aggarwal, N.; Aguiar, O. D.; Aiello, L.; Ain, A.; Ajith, P.; ...'. The abstract text describes the discovery of a binary neutron star coalescence candidate (later designated GW170817) with merger time 12:41:04 UTC, observed through gravitational waves by the Advanced LIGO and Advanced Virgo detectors. The Fermi Gamma-ray Burst Monitor independently detected a gamma-ray burst (GRB 170817A) with a time delay of ~ 1.7 s with respect to the merger time. From the gravitational-wave signal, the source was initially localized to a sky region of 31 deg^2 at a luminosity distance of $\sim 40\text{-}8^+ Mpc$ and with component masses consistent with neutron stars. The component masses were later measured to be in the range 0.86 to $2.26 (M_{\odot})$. An extensive observing campaign was launched across the electromagnetic spectrum leading to the discovery of a bright optical transient (SSS17a, now with the IAU identification of AT 2017gfo) in NGC 4993 (at ~ 40 [Mpc]) less than 11 hours after the merger by the One-Meter, Two Hemisphere (1M2H) team using the 1 m Swope Telescope. The optical transient was independently detected by multiple teams within an hour. Subsequent observations targeted the object and its

On the left side, there are navigation options: 'VIEW' (Abstract, Citations (2576), References (196), Co-Reads, Similar Papers, Volume Content, Graphics, Metrics, Export Citation) and 'FEEDBACK'. On the right side, there are options to 'Add paper to library' and 'FULL TEXT SOURCES' (My Institution, Publisher, arXiv) and 'DATA PRODUCTS' (SIMBAD (6), NED (2), MAST (1), IRSA (1), Gemini (1), ESO (1), ESA (1), Chandra (1)). At the bottom right, there is a 'GRAPHICS' section with a thumbnail image and the text 'Click to view more'.

Ingestion Policy (still evolving)

What is/will be indexed in ADS

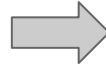
- Curated, high-level datasets with good metadata (registered DOI) and clear authorship information
- Research data collections that have shown reuse value (initially via citations, i.e. `citation_count > 1`)

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)

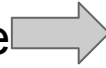
The larger ecosystem of Discovery

Data Archives/users register data products using DOIs and rich metadata which includes provenance



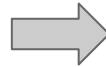
ADS (and others) can discover and index/link data products via text mining of DAS/reference sections

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