ADS Services in support of the Discovery, Management and Evaluation of Science Data

Alberto Accomazzi
ADS Program Manager and PI - @aaccomazzi
ESO/ESA Science Operations 2015 - 24 November 2015

Overview

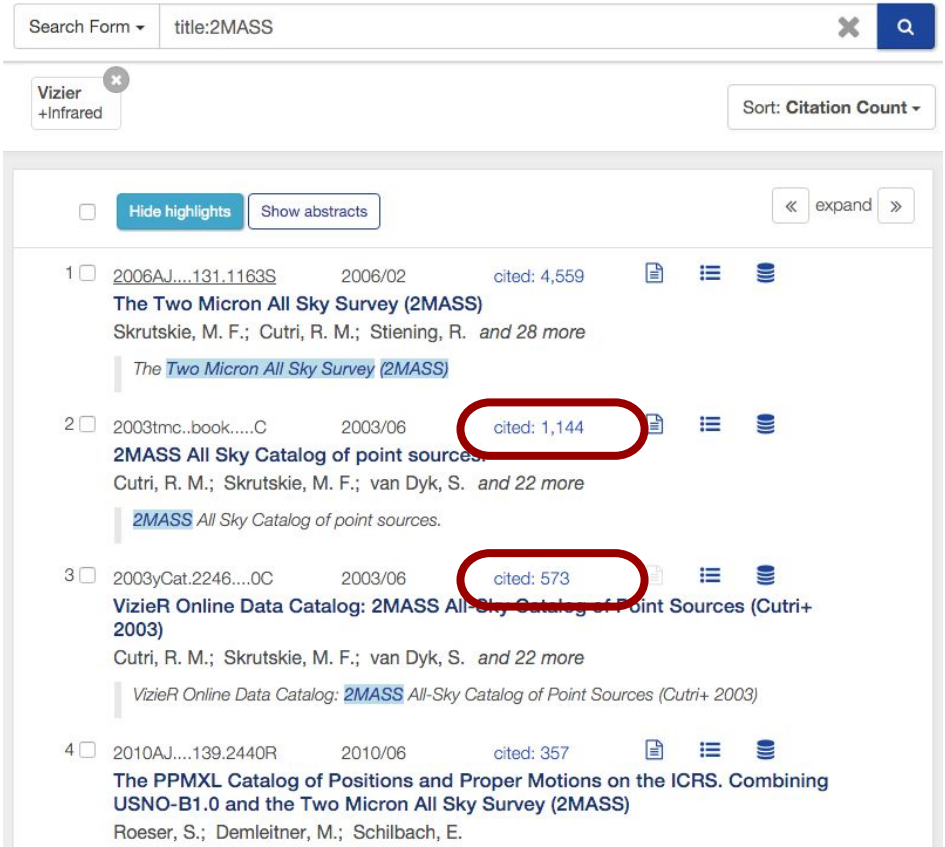
- “Data” indexing in ADS:
  - High-level data products
  - Observing Proposals
  - Institutional and Telescope Bibliographies
  - Links to data products
- Challenges in Scholarly Publishing
  - Data citations
  - Software citations
- Upcoming Challenges in Data Management
  - Agency OA mandates
  - Identity management
  - Thought and Predictions

What ADS Indexes

- Bibliographic metadata (authors, title, abstract) from multiple sources (arXiv and publishers) - 10M records
- Full-text documents from most Astronomy & Physics journals - 4.5M records
- High-level data products appearing in journal articles (mostly Vizier catalogs) - 10K records
- Observing proposals - 36K records
- Bibliographic groups - 330K records
- Data links - 300K records
- Software entries from the Astrophysics Source Code Library (ASCL) - 1K records

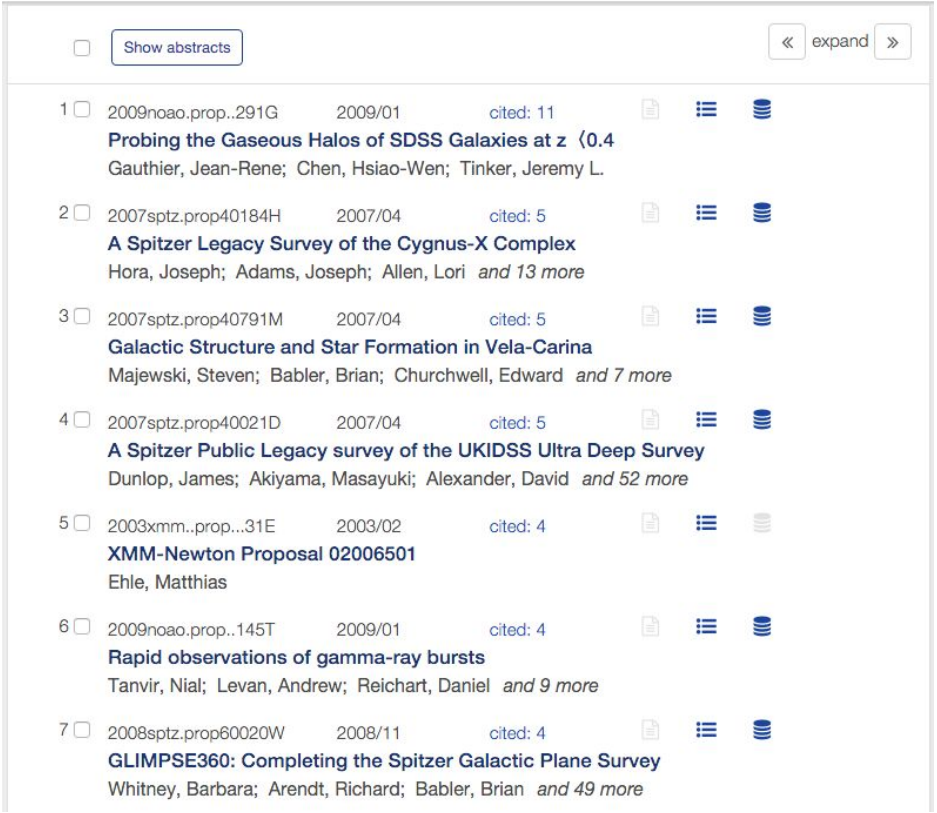
High-level Data Products Indexed in ADS

- Important datasets are often described in “data” papers
- But can also be available as electronic catalogs
- Greatest majority are from Vizier (close to 10,000 records)
- Once in ADS, they become easily discoverable, citable
- This is how our community has dealt with “data citation” all along



Observing Proposals Indexed in ADS

- Proposals contain early descriptions of current and ongoing science activities
- They provide a direct link to existing or planned observations
- HST, IUE, CXC, NOAO, XMM, KOA, Spitzer, ATNF, Subaru, …
- 36,000 records, 38,000 data links, 300 citations
- Ongoing ingest rate is 1,000 records/year



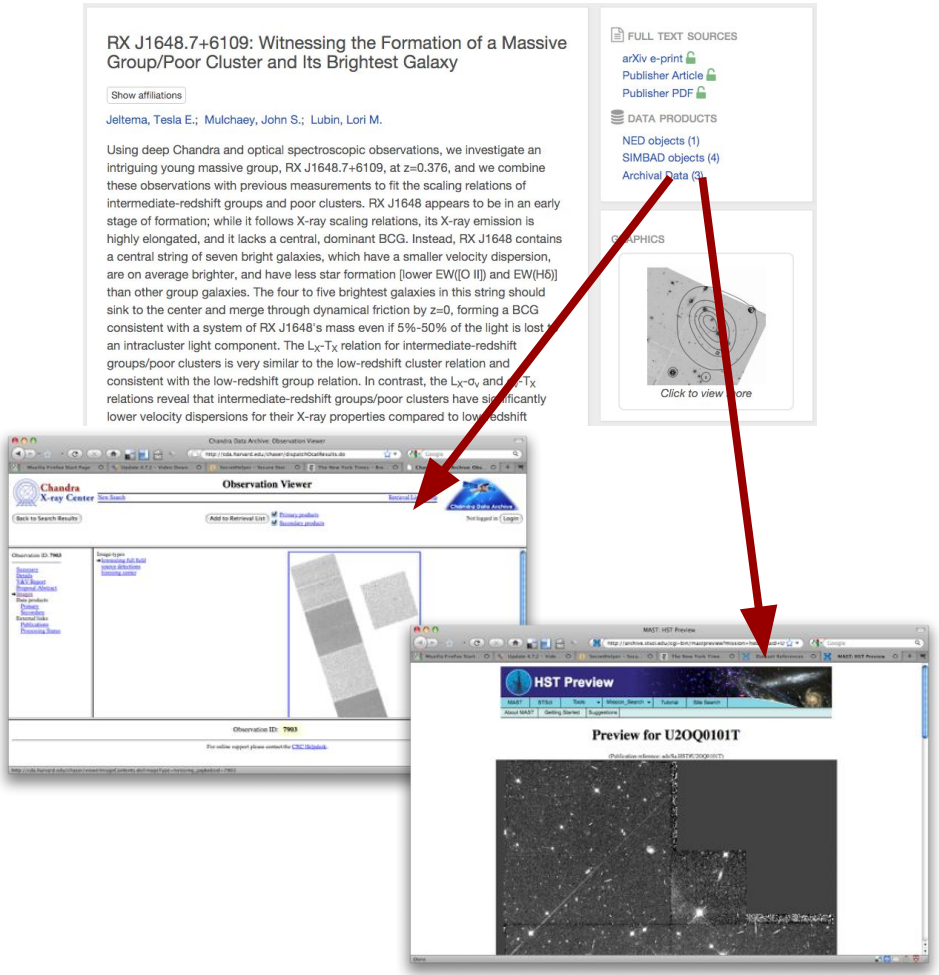
Bibliographies

- Institutional bibliographies, highlighting scientific output from research center or project
- “Telescope” bibliographies, identifying papers related to their data products
- About 30 bibliographic groups so far, over 330K records
- Help with scientific evaluation of projects and institutions, but also useful in disambiguation

ALMA	ISO	ROSAT
ARI	IUE	SDO
CfA	JCMT	SMA
CFHT	Keck	Spitzer
Chandra	Leiden	Subaru
ESO	LPI	Swift
Gemini	Magellan	UKIRT
Herschel	NOAO	USNO
HST	NRAO	XMM

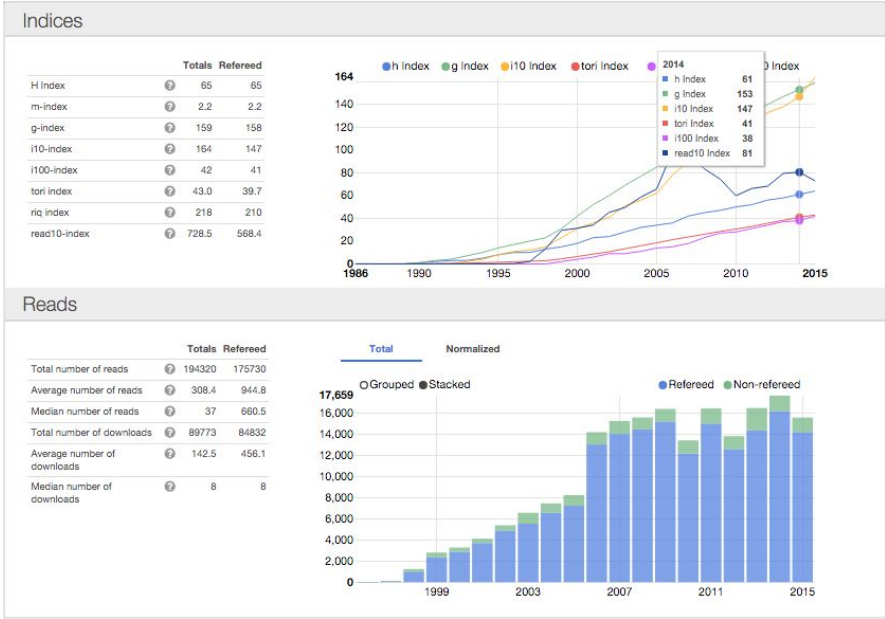
Data Links

- Have existed between Data Centers and ADS since 1994
- Maintained by librarians, data archivists, harvested by ADS
- Bibcode-URL pairs, linking to either individual observations or aggregates
- Often part of data center's bibliographies, used to compute metrics

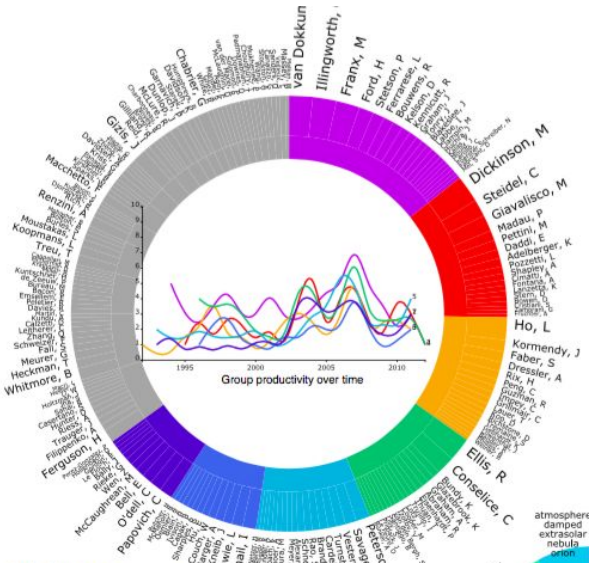


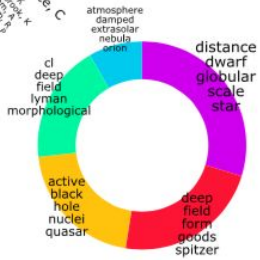
Benefits

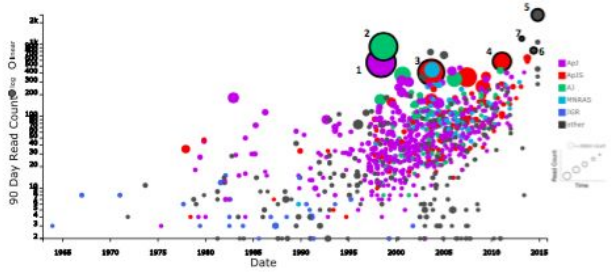
- Search and filter by bibliography or data property:
  “exoplanets and bibgroup:CfA”
- Find multi-wavelength papers and access archival data:
  “data:CXO and data:HST”
- View paper-based metrics for people and projects

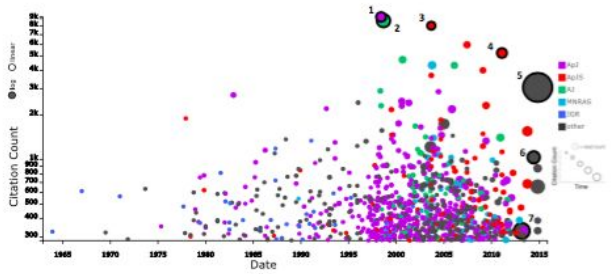


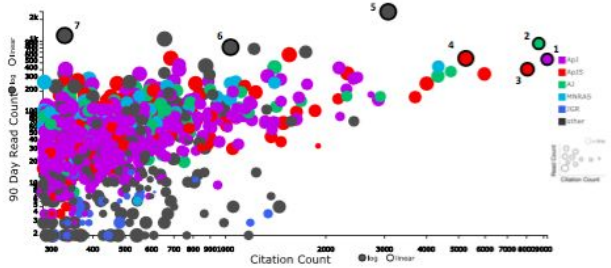
Metrics & Analytics





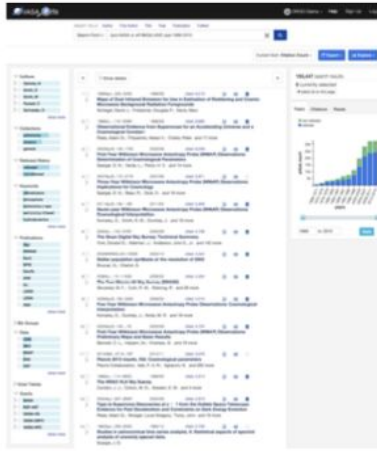






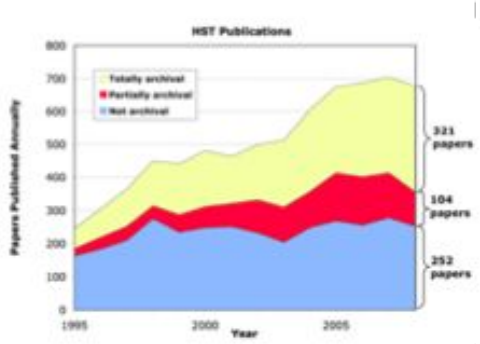
A sample of the 1,000 top-cited papers in ADS acknowledging NASA funding or affiliation.

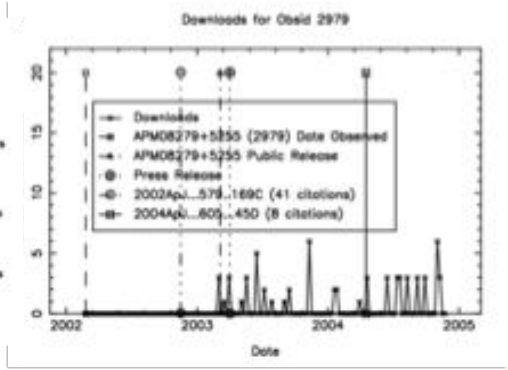
1. **Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds**
   Schlegel et al, 1998ApJ...500..525S
   Citations: **9216**, Reads: **559**
2. **Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant**
   Riess et al, 1998AJ....116.1009R
   Citations: **8693**, Reads: **928**
3. **First-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters**
   Spergel et al, 2003ApJS..148..175S
   Citations: **8040**, Reads: **407**
4. **Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Interpretation**
   Kumatsu et al, 2011ApJS..192...18K
   Citations: **5253**, Reads: **577**
5. **Planck 2013 results. XVI. Cosmological parameters**
   Plank Collaboration, 2014A&A...571A..16P
   Citations: **3075**, Reads: **2563**
6. **Detection of B-Mode Polarization at Degree Angular Scales by BICEP2**
   Bicep2 Collaboration, 2014PhRvL.112x1101A
   Citations: **1035**, Reads: **824**
7. **emcee: The MCMC Hammer**
   Foreman-Mackey et al, 2013PASP..125..306F
   Citations: **329**, Reads: **1200**

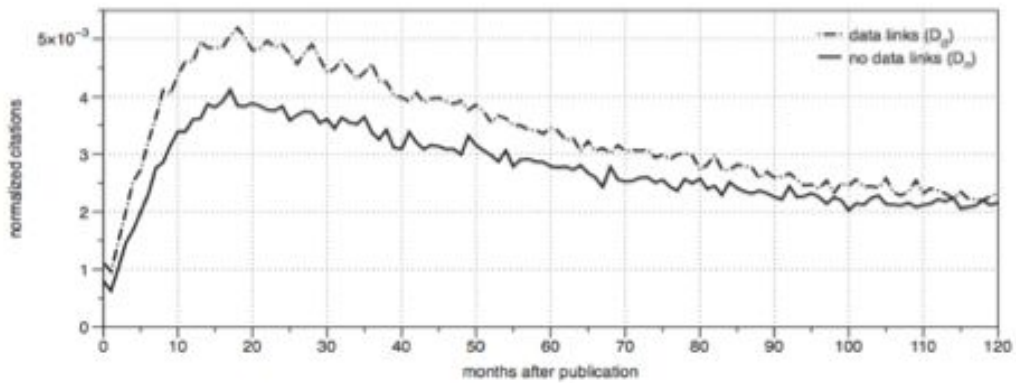


Benefits to Archives: Linked Data Advantage

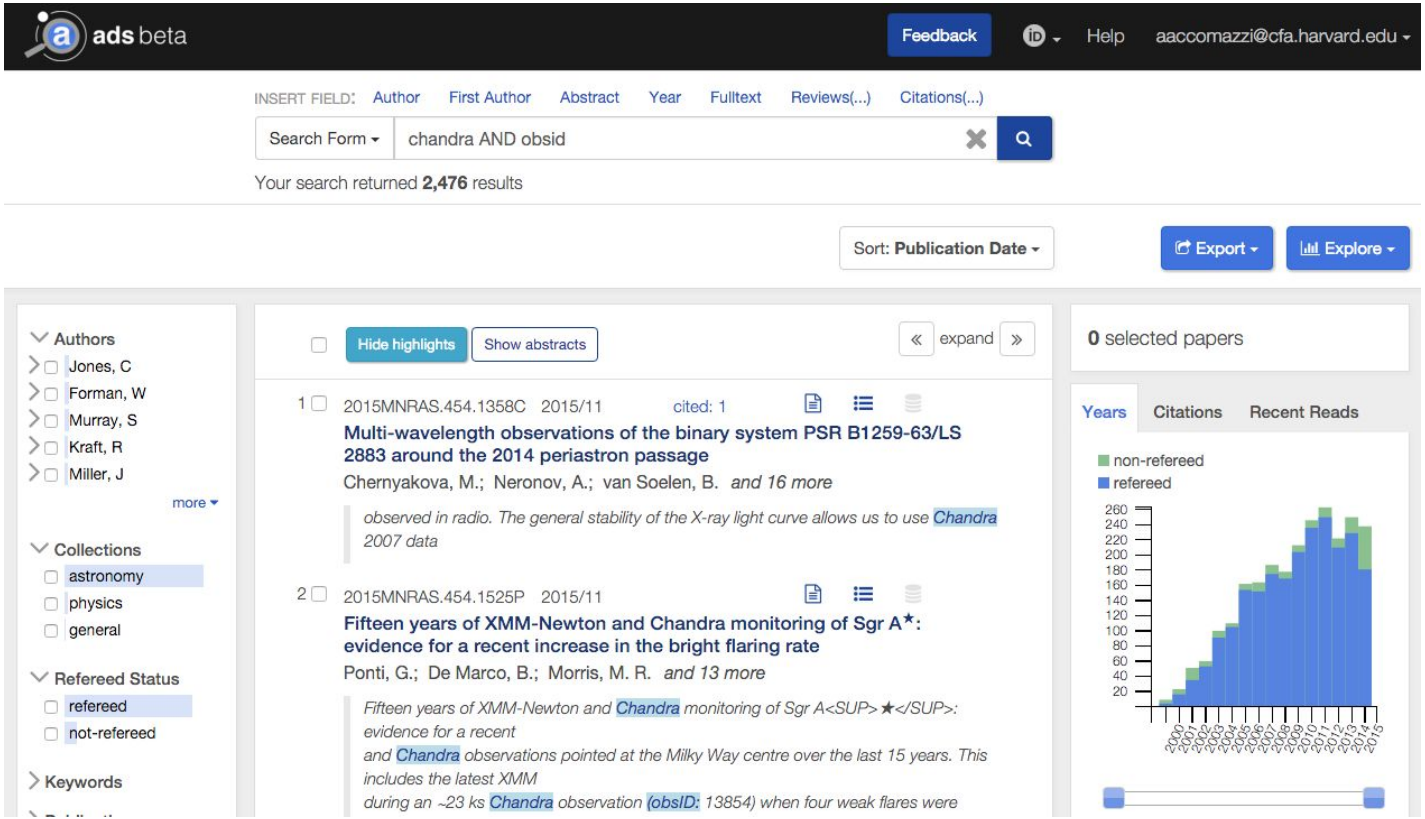
- Well-archived, well-linked data is often shared, re-used (White et al, 2009)
- Data usage increases upon publication of papers (Winkelman et al, 2006)
- Well-linked papers receive more citations (Henneken & Accomazzi, 2012; Dorch et al, 2015)







https://ui.adsabs.harvard.edu

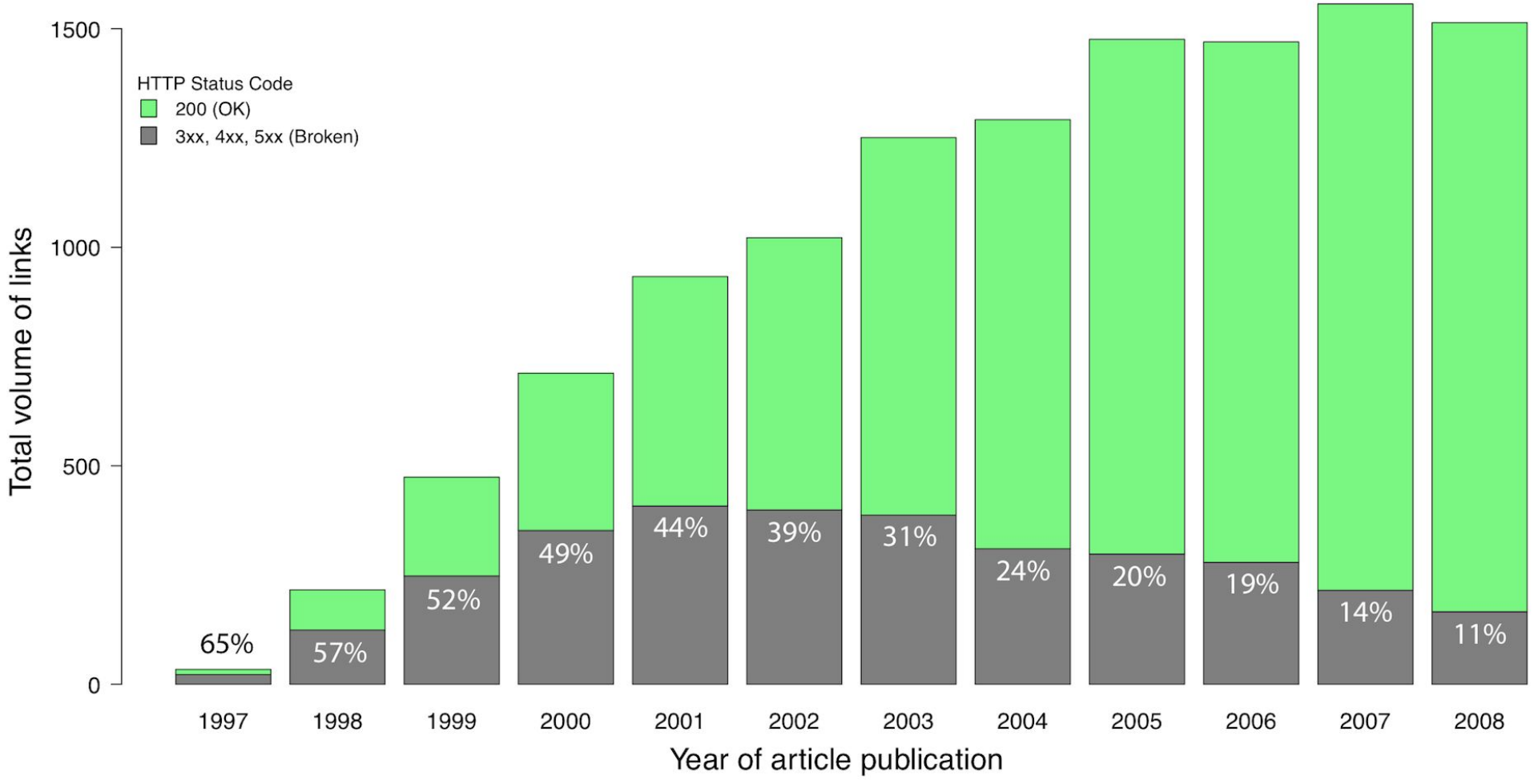


https://github.com/adsabs/adsabs-dev-api

Data Citation

- In the IVOA we have been talking about “data publication,” a necessary step towards the goal of enabling data citation, which is the practice of providing a formal reference to a data product the way we currently do for papers.
- In Astronomy, this has not been a pressing need for a number of reasons:
  - High-value data has been curated and made available from a number of well-managed archives
  - Joint curation efforts have made it possible to create and maintain paper/data links
  - We have had a well-established practice of using papers as a proxy for citing important datasets
  - We have a history of pre- and post-publication curation efforts which alleviate the need for this
- But some needs are going unmet and some problems remain unresolved
  - When citing via data paper impossible, fragile solutions such as URL in footnote are used which lead to “link rot:” http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0104798
  - Not having a robust way to capture citations/links to data upfront leads to far more work for curators, librarians post-publication (“what data was used?”)

Broken links in AAS journals (Pepe et al, 2014)



Joint Declaration of Data Citation Principles (2014)

1. **Importance** - “data is first class citizen in scholarly discourse”
2. **Credit and Attribution** - “provide scholarly credit to contributors”
3. **Evidence** - “whenever a claim relies upon data, the data should be cited”
4. **Unique Identification** - “citation should include a unique identifier to data”
5. **Access** - “data citation should facilitate access to data and metadata”
6. **Persistence** - “identifiers, metadata, and data disposition should persist”
7. **Specificity and Verifiability** - citation should facilitate identification of, access to, and verification of the specific data that support a claim”
8. **Interoperability and Flexibility** - “citation methods should be flexible enough to accommodate different practices, but still provide technical interoperability”

https://www.force11.org/group/joint-declaration-data-citation-principles-final

Required Effort

The main reasons why we are talking about this is to enable repeatability (scientific goal) while also providing proper credit and attribution (data management requirement). Both of these are worthy goals but don't come for free. In order to find out whether your data is cited we need the following things:

- A persistence layer over your data products, with the capability to deposit and update the corresponding metadata with a registration authority
- Buy-in from publishers allowing data products to be listed as references so that there is a consistent community standard for data citations
- A discovery system which identifies these citations and publishes their information (ADS does this for papers and high-level data products so far)

What data products should be citable?

- High level data products associated with a paper (e.g. Vizier catalogs):
  Straightforward (authorship, metadata inherited from paper)
- Data Catalogs (e.g. 2MASS, Wise, etc) as a “whole:”
  Pretty straightforward, need to figure out proper authorship rules
- Individual data catalog tables, specific releases:
  Possibly useful, but maybe not strictly necessary
- Data collections (e.g. all observations from an archive analyzed in a paper or series of papers, currently an ongoing MAST prototype project):
  Useful, although requires infrastructure to capture collection and metadata
- Individual ObsIds (pointed observations): Useful, also requires infrastructure

Software Citation

- Similar to the “data citation” issue, although we are seeing publishers adopt consistent practices based on ASCL ids
- With the indexing in ADS of the ASCL, citing software is as easy as cutting and pasting a bibtex entry
- However, a number of issues still exist:
  - Acceptance: ensuring that publishers will allow the submissions of software references
  - Versioning: how do capture citation to specific software versions in support of repeatibility
  - Preservation: how to ensure that the software is properly stored and available in the future
  - Attribution: since software changes in time, the author list will also change
- If we can solve the data citation issue, we will be well posed to solve the software citation issue as well

https://ui.adsabs.harvard.edu/#search/q=bibstem%3A%22ascl%22&sort=citation_count+desc

1,146 results

Search Form ▾ | bibstem:"ascl"

AUTHORS
- Jenness, T
- Bertin, E
- Berry, D
- Gammie, C
- Portegies Zwart, S

more ▾

COLLECTIONS
- astronomy

REFEREED
- refereed
- not-refereed

KEYWORDS

PUBLICATIONS

BIB GROUPS

DATA

VIZIER TABLES

GRANTS

Show abstracts

1 2012ascl.soft03003C 2012/03 cited: 41
spec2d: DEEP2 DEIMOS Spectral Pipeline
Cooper, Michael C.; Newman, Jeffrey A.; Davis, Marc *and 2 more*

2 2012ascl.soft08004S 2012/08 cited: 28
PyKE: Reduction and analysis of Kepler Simple Aperture Photometry data
Still, Martin; Barclay, Tom

3 2013ascl.soft05002P 2013/05 cited: 23
pynbody: N-Body/SPH analysis for python
Pontzen, Andrew; Roškar, Rok; Stinson, Greg *and 1 more*

4 2011ascl.soft05003C 2011/05 cited: 23
The DTFE public software: The Delaunay Tessellation Field Estimator code
Cautun, Marius C.; van de Weygaert, Rien

5 2012ascl.soft05004P 2012/05 cited: 19
Turbospectrum: Code for spectral synthesis
Plez, B.

6 2010ascl.soft10068B 2010/10 cited: 16
SWarp: Resampling and Co-adding FITS Images Together
Bertin, Emmanuel

7 2011ascl.soft09022H 2011/09 cited: 9
Synspec: General Spectrum Synthesis Program
Hubeny, Ivan; Lanz, Thierry

0 selected

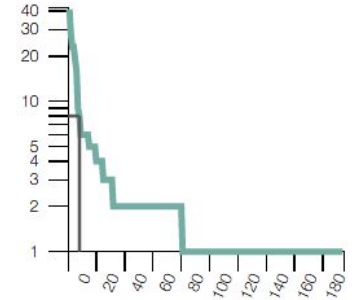
Add papers to library

Years | Citations | Recent Reads

total number of citations : **481**

H-Index for results: 8

Y-axis: linear ○ log ◉



Limit to top 1146 most cited Apply

Big Data management vs. big Data Management

- We are all excited about Big Data, and are preparing for its management
- But are we ready for the increasing demands being imposed on us by funding agencies, government mandates, technological advances?
  - Example: in US, OSTP mandate requires agencies to implement OA for publications and data. 24 agencies, a dozen different plans to comply with, whose problem is it?
  - Data management plans required by NASA and NSF still largely “fuzzy,” although moving in right direction. Data management plans for ground-based data in US lag behind.
  - Institutions, research agencies, countries are requiring scientists to report research output using systems like ORCiD as a way to automate evaluation
  - Journals such as PLOS ONE require the specification of Contributor Roles in papers (http://dictionary.casrai.org/Contributor_Roles); how should this affect evaluation?
- ADS offers *some* help...

ORCiD Integration in ADS

- ORCiD (Open Researcher and Contributor ID) aims to tackle the problem of author disambiguation by assigning unique ids to contributors (http://orcid.org)
- ADS users can now “claim” papers, i.e. create their bibliography.
  This exists already but the UI will be updated based on early feedback.
- ADS is indexing ORCiDs collected when authors submit papers to publishers.
  This covers journals that send this information to ADS (e.g. AAS)
- ADS now allows users to find people by searching for their ORCiDs.
  Search for publisher-provided ORCiD works, but limited in coverage; searching of user claims available shortly.
- ADS will allow users to discriminate ORCiD-paper associations.
  Differentiate between user claims and authoritative mappings.

Your search returned **184** results

Sort: **Publication Date** | Export | Explore

AUTHORS
- Accomazzi, A
- Kurtz, M
- Grant, C
- Murray, S
- Eichhorn, G

more

COLLECTIONS
- astronomy
- general
- physics

REFEREED
- refereed
- not-refereed

KEYWORDS

PUBLICATIONS

Show abstracts

1 2015ASPC..495..401C 2015/09
ADS 2.0: New Architecture, API and Services
Chyla, R.; Accomazzi, A.; Holachek, A. *and 7 more*

2 2015IAUGA..2257768A 2015/08
The NASA Astrophysics Data System joins the Revolution
Accomazzi, Alberto; Kurtz, Michael J.; Henneken, Edwin *and 7 more*

3 2015IAUGA..2257639R 2015/08
Bibliographies and Data Archives: a Rich Data Mining Tool
Rots, Arnold H.; Winkelman, Sherry; Accomazzi, Alberto

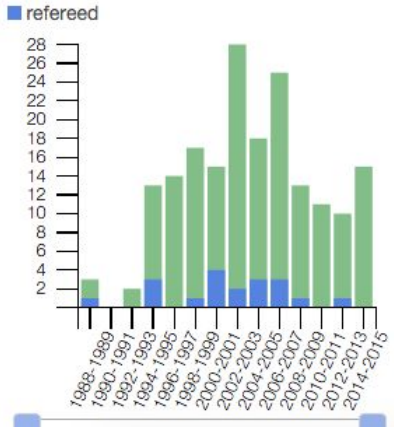
4 2015IAUGA..2257982A 2015/08
Curation and integration of observational metadata in ADS
Accomazzi, Alberto

5 2015ASPC..492..208G 2015/04 cited: 2
Enabling Meaningful Affiliation Searches in the ADS
Grant, C. S.; Thompson, D. M.; Chyla, R. *and 6 more*

0 selected

Add papers to library

Years | Citations | Recent Reads



ads beta

Feedback | ORCID | Learn | Account

INSERT FIELD: Author | First Author | Abstract | Year | Fulltext | Reviews(...)

Search Form | author:"accomazzi, alberto"

Your search returned **184** results

Signed in to ORCID as
Alberto Accomazzi

ORCID Mode off on

View my ORCID papers in ADS

View my ORCID profile

Log out from ORCID in ADS

Export | Explore

AUTHORS
- Accomazzi, A
- Kurtz, M
- Grant, C
- Murray, S
- Eichhorn, G

more

COLLECTIONS
- astronomy
- general
- physics

REFEREED
- refereed
- not-refereed

KEYWORDS

PUBLICATIONS

Show abstracts

1 2015ASPC..495..401C 2015/09
ADS 2.0: New Architecture, API and Services
Chyla, R.; Accomazzi, A.; Holachek, A. *and 7 more*

2 2015IAUGA..2257768A 2015/08
The NASA Astrophysics Data System joins the Revolution
Accomazzi, Alberto; Kurtz, Michael J.; Henneken, Edwin *and 7 more*

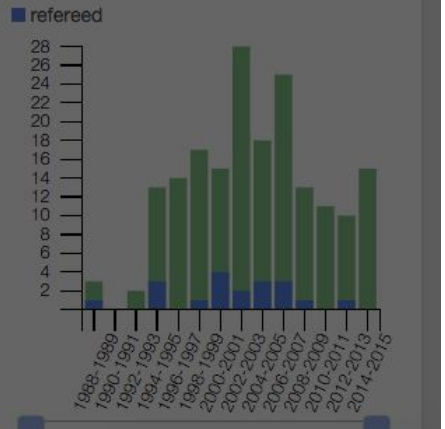
3 2015IAUGA..2257639R 2015/08
Bibliographies and Data Archives: a Rich Data Mining Tool
Rots, Arnold H.; Winkelman, Sherry; Accomazzi, Alberto

4 2015IAUGA..2257982A 2015/08
Curation and integration of observational metadata in ADS
Accomazzi, Alberto

5 2015ASPC..492..208G 2015/04 cited: 2
Enabling Meaningful Affiliation Searches in the ADS
Grant, C. S.; Thompson, D. M.; Chyla, R. *and 6 more*

Add papers to library

Years | Citations | Recent Reads



INSERT FIELD: Author First Author Abstract Year Fulltext Reviews(...) Citations(...)

Search Form | author:"accomazzi, alberto"

Your search returned **184** results

Sort: **Publication Date** | Export | Explore

AUTHORS
- Accomazzi, A
- Kurtz, M
- Grant, C
- Murray, S
- Eichhorn, G

more

COLLECTIONS
- astronomy
- general
- physics

REFEREED
- refereed
- not-refereed

KEYWORDS

PUBLICATIONS

BIB GROUPS

Show abstracts

1 2015ASPC..495..401C 2015/09
ADS 2.0: New Architecture, API and Services
Chyla, R.; Accomazzi, A.; Holachek, A. *and 7 more*
Actions | In ORCID

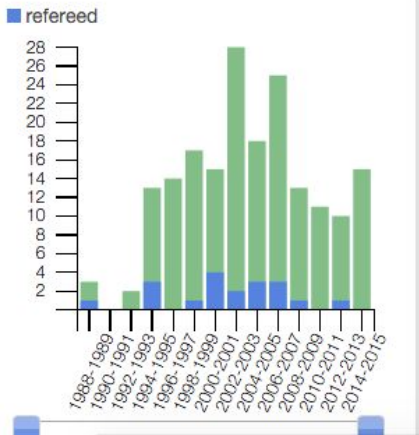
2 2015IAUGA..2257768A 2015/08
The NASA Astrophysics Data System joins the Revolution
Accomazzi, Alberto; Kurtz, Michael J.; Henneken, Edwin *and 7 more*
Actions | In ORCID
- update in ORCID
- delete from ORCID

3 2015/08
B chives: a Rich Data Mining Tool
Sherry; Accomazzi, Alberto
Actions | In ORCID

0 selected

Add papers to library

Years Citations Recent Reads



ads beta

Feedback | ORCID | Learn | Account

INSERT FIELD: Author | First Author | Abstract | Year | Fulltext | Reviews(...) | Citations(...)

Search Form | orcid:0000-0002-4110-3511

Your search returned **19** results

Sort: **Publication Date** | Export | Explore

AUTHORS
- Accomazzi, A
- Henneken, E
- Kurtz, M
- Grant, C
- Luker, J

more

COLLECTIONS
- astronomy
- general
- physics

REFEREED
- refereed
- not-refereed

KEYWORDS

PUBLICATIONS

RIB GROUPS

Show abstracts

1 2015IAUGA..2257639R 2015/08
Bibliographies and Data Archives: a Rich Data Mining Tool
Rots, Arnold H.; Winkelman, Sherry; Accomazzi, Alberto

2 2015A&C....11..190H 2015/06
The Virtual Astronomical Observatory: Re-engineering access to astronomical data
Hanisch, R. J.; Berriman, G. B.; Lazio, T. J. W. *and 4 more*

3 2015ASPC..492..204F 2015/04
Management of the Unified Astronomy Thesaurus
Frey, K.; Erdmann, C.; Accomazzi, A. *and 4 more*

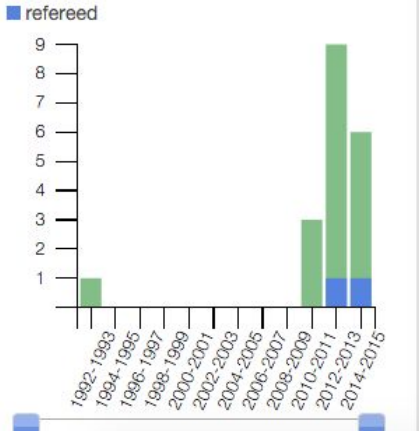
4 2015ASPC..492..150T 2015/04
Saving the Orphaned Astronomical Literature
Thompson, D. M.; Henneken, E. A.; Grant, C. S. *and 6 more*

5 2015ASPC..492...85E 2015/04 cited: 1
Using ADS for Creating Bibliographies of Research Institutions

0 selected

Add papers to library

Years | Citations | Recent Reads



The Bigger Picture

- Increasingly we are asked to identify things in a unique, unambiguous, machine-readable way:
  - Publications (bibcodes, DOIs)
  - People (ORCiDs) - ready for launch Q1 2016
  - Institutions (Ringgold IDs, ISNIs) - working on affiliation normalization
  - Funding sources (FundRef) - extracted some grant ids via text mining
  - Datasets, software (DataCite DOIs) - working with publishers on policies, funding opportunities
  - Facilities ...
- And we need to keep track of how it is all linked together
- The only way to achieve this is to publish, aggregate, and share this information. It’s good for science and it’s good for management.

Final Thoughts and Predictions

- Discipline-specific solutions will allow us to move faster but won't take us as far. We should pay attention to initiatives within RDA.
- Mandates will force us to provide global solutions to what once were local problems, but as a discipline we are well positioned to deal with this
- The need to quantify and evaluate will drive us further into the collection and publishing of metrics and analytics (cf. "altmetrics")
- The issues of software and data citation will take a little while to sort out, but we will get there by the end of the decade
- New forms of publications such as blogs might enter the scholarly discourse; ultimately the community will decide what belongs in scientific publications

Thanks!

- ADS “bumblebee” UI: https://ui.adsabs.harvard.edu
- ADS API: https://github.com/adsabs/adsabs-dev-api
- ADS Blog: https://adsabs.github.io/blog/
- IVOA Intro to DOIs: https://docs.google.com/a/cfa.harvard.edu/presentation/d/1yo9kJz01umlwkE_uGeSnBjjbWniFCx8tBXeD9AWDpys/edit?usp=sharing
- ADS/ORCID integration plan: https://docs.google.com/presentation/d/1g1zhyOJkQb5XspDSI1rMOCafYLl6EFIDKWU8PcLznQ4/edit?usp=sharing