

Data Citations in Astronomy

Alberto Accomazzi

DC&P Session
IVOA Interop Meeting
Pune, India
17 October 2011

The buzz about Publishing Data

- The rest of the world has discovered the importance of Publishing and citing data
- Data Citation Practices and Standards workshop:
http://sites.nationalacademies.org/PGA/brdi/PGA_064019
<http://datacite.org/node/30>
- Lots of issues come into play: Archival, Preservation, Nomenclature, Persistence, Attribution, Discovery
- Issue now pressing because of mandates, funding requirements

The General Thinking

Smit, E (2011) Abelard and Héloïse: *Why Data and Publications Belong Together*. D-Lib Magazine doi:10.1045/january2011-smit

- Journals to require availability of underlying research material as an editorial policy
- More careful treatment of submitted digital research data by those who accept them
- Ensure data is stored, curated and preserved in trustworthy places
- Ensure links (bi-directional) and persistent identifiers between data and publications
- Establish uniform citation practices of data
- Develop data-publications and quality standards

Links between Publications and Data Products

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

Online Archives

Abstract

Using deep Chandra and optical spectroscopic observations, we investigate an intriguing young massive group, RX J1648.7+6109, at $z=0.376$, and we combine these observations with previous measurements to fit the scaling relations of intermediate- and high groups and poor clusters. RX J1648 appears to be in an early stage of formation. It follows $Y_{\text{500}}-z_{\text{500}}$ relations, its Y_{500} emission is highly absorbed, and it lacks a central AGN. Instead, RX J1648 contains a central source of energy that is not a point source, which

MAST: HST Preview

HST Preview

Preview for U2OQ0101T

(Publication reference: [ads/5a35TFU2OQ0101T](#))

SAO/NASA ADS Abstract Service

Links for 2008ApJ...685..138J

- [European Southern Observatory HST Proposal \(ST-ECF\)](#)
- [Multimission Archive at STScI](#)
- [MAST References \(HST\)](#)
- Resource at [oda.harvard.edu](#)
- [Chandra Data Archive ObsIDs 8472, 7903](#)

Chandra X-ray Center

Observation Viewer

Observation ID: 7903

Image types: [Viewing full field](#), [Source detections](#), [Viewing center](#)

“Data” records in ADS

The screenshot shows the ADS website interface. The browser address bar contains the URL: `labs.adsabs.harvard.edu/ui/cgi-bin/topicSearch?q=2MASS&qtype=IMPORTANT&db_key=AST&db_key=PRE&arxiv_sel=astro-ph`. The page title is "2MASS - Most cited". On the right side of the page, there are buttons for "Top 34 results", "View as...", and "Export to...". Below the title, there is a filter section for "Vizier: Infrared [X]". On the left side, there is a "FILTER BY:" section with several categories: "Authors" (listing Kirkpatrick, J (33), Burgasser, A (24), Reid, I (21), Liebert, J (20), Gizis, J (18)), "Keywords", "Archives" (listing CDS (47), NED (44), ESO (18), MAST (10), Chandra (2)), "Missions", "SIMBAD Objects", and "Vizier Tables". The main content area displays a list of search results. The first result is "2006AJ....131.1163S The Two Micron All Sky Survey (2MASS)" by Skrutskie, M. F.; Cutri, R. M.; Stiening, R.; Weinberg, M. D.; Schneider, S.; Carpenter, J. M.; Beichman, C.; Capps, R.; Chester, T.; Elias, J.; and 21 coauthors. The second result is "2003tmc..book....C 2MASS All Sky Catalog of point sources." by 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.; and 15 coauthors. The 23rd result is "2003yCat.2246....0C 2MASS All-Sky Catalog of Point Sources (Cutri+ 2003)" by 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.; and 15 coauthors. The 25th result is "2004AJ....127.3553K Near-Infrared Photometry and Spectroscopy of L and T Dwarfs: The Effects of Temperature, Clouds, and Gravity" by Knapp, G. R.; Leggett, S. K.; Fan, X.; Marley, M. S.; Geballe, T. R.; Golimowski, D. A.; Finkbeiner, D.; Gunn, J. E.; Hennawi, J.; Ivezić, Z.; and 19 coauthors. The 27th result is "2002ApJ...564..421B The Spectra of T Dwarfs. I. Near-Infrared Data and Spectral".

2,156 citations

812 citations

249 citations

Machine Readable Tables (MRTs) published with articles

Table 1 [CITED IN TEXT](#) | [Machine-readable \(MRT\)](#) | [Virtual Observatory \(VOT\)](#) | [TYPESET IMAGE](#) Go to: [Table 2](#)

Cleaned Cluster Catalog

Object	R.A. 2000	Decl.	<i>V</i>	Type	S ^a	Ap ["]	P ^b	C ^c
<i>SH01</i>	0:32:41.44	40:01:41.4	15.82				G	
G001-MII	0:32:46.53	39:34:40.6	13.75		B		B	
G002-MIII	0:33:33.77	39:31:18.9	15.81		B		B	
<i>BH01</i>	0:34:11.92	39:24:11.6						
B290	0:34:20.94	41:28:18.1	17.14		P		B	

Notes.

- ^a Source of velocity: HS = this paper; B = Barnby et al. (2000); P = Perrett et al. (2002).
- ^b Source of photometry: L = this paper; B = Barnby et al. (2000); G = Galieti et al. (2007), H = Huxor et al. (2005).
- ^c Source of classification as a cluster: S = spectrum from this paper indicates a cluster; L = LGS image indicates nonstellar; H = *HST* image indicates a cluster; objects with blank entries in this column and in the velocity source column should still be considered "candidates."
- ^d Not a cluster in Barnby et al. (2000).
- ^e Not a cluster in Crampton et al. (1985).
- ^f Not a cluster in Racine (1991).
- ^g Dubath & Grillmair (1997) probably observed a different object.
- ^h Not a cluster in Racine & Harris (1992).

“Data” linking in AAS journals



Data Set Linking

Introduction

Formatting Data Sets in
AAS Submissions

Data Set Linking vs.
Supplementary Materials

Data Set Verification Tool
Data Set Verification Tips
ADS Data Set Verification
and Resolution Services

Participating Data Centers

AAS^{Te}X Information

AAS^{Te}X Home Page
Object Linking
Facility Keywords

Print-friendly page

The AAS^{Te}X Package for Manuscript Preparation

Data Set Linking

In partnership with the NASA Astronomical Data System (ADS) and several NASA data centers, the AAS has a new project to allow authors to tag data sets from [participating data centers](#) in their papers using the AAS^{Te}X “\dataset” macro. Data sets tagged with the “\dataset” macro will appear in the electronic edition linked to a name resolver at ADS that will take readers to the data sets themselves. The extra time required of authors to determine data set identifiers and [tag them appropriately in their articles](#) will have numerous and far-reaching benefits:

- Readers will get immediate access to the data used in the paper.
- Data centers will be able to quickly and efficiently construct links from the data back to the electronic journal papers allowing researchers a seamless transition between the electronic journals and the data centers
- Papers with these tags will have a higher visibility to readers and researchers. Greater visibility means these authors are more likely to be read and cited.

Authors of AAS journals are welcome to tag individual data sets from participating data centers in their papers. To take advantage of data set linking, authors first need to get the unique identifier for each of their data sets. These identifiers take the form “*ADS/FacilityId#PrivateId*” where *FacilityId* is the facility acronym and **PrivateId** is the identifier given to each data set by the facility archive. The participating data centers have different ways of providing these identifiers. Further information on the current list of participating data centers and their data set linking resources can be found on the [Participating Data Centers](#) page. Authors may verify their identifiers before submission using the [AAS Data Set Verification Tool](#) to query the ADS name resolver. See [Formatting Data Sets in AAS Journals](#) for instructions on how to tag data sets in AAS^{Te}X.

Identifiers included in accepted manuscripts will be verified during copy editing with the ADS resolver tool. Accompanying text may be edited for clarity. Authors will have an opportunity during the editing process to correct unresolved identifiers or add clarifying text. Identifiers that do not resolve at ADS will not be linked in the electronic edition.

Last updated: 19 May 2011
aastex-help@aaas.org

What's missing?

THE ASTRONOMICAL JOURNAL, 131:2921–2933, 2006 June
© 2006. The American Astronomical Society. All rights reserved. Printed in U.S.A.

THE COMPLETE SURVEY OF STAR-FORMING REGIONS: PHASE I DATA

NAOMI A. RIDGE,¹ JAMES DI FRANCESCO,² HELEN KIRK,^{2,3} DI LI,^{1,4} ALYSSA A. GOODMAN,¹ JOÃO F. ALVES,⁵
HÉCTOR G. ARCE,⁶ MICHELLE A. BORKIN,⁷ PAOLA CASELLI,⁸ JONATHAN B. FOSTER,¹ MARK H. HEYER,⁹
DOUG JOHNSTONE,^{2,3} DAVID A. KOSSLYN,¹ MARCO LOMBARDI,⁴ JAIME E. PINEDA,¹
SCOTT L. SCHNEE,¹ AND MARIO TAFALLA¹⁰

Received 2005 November 8; accepted 2006 February 22

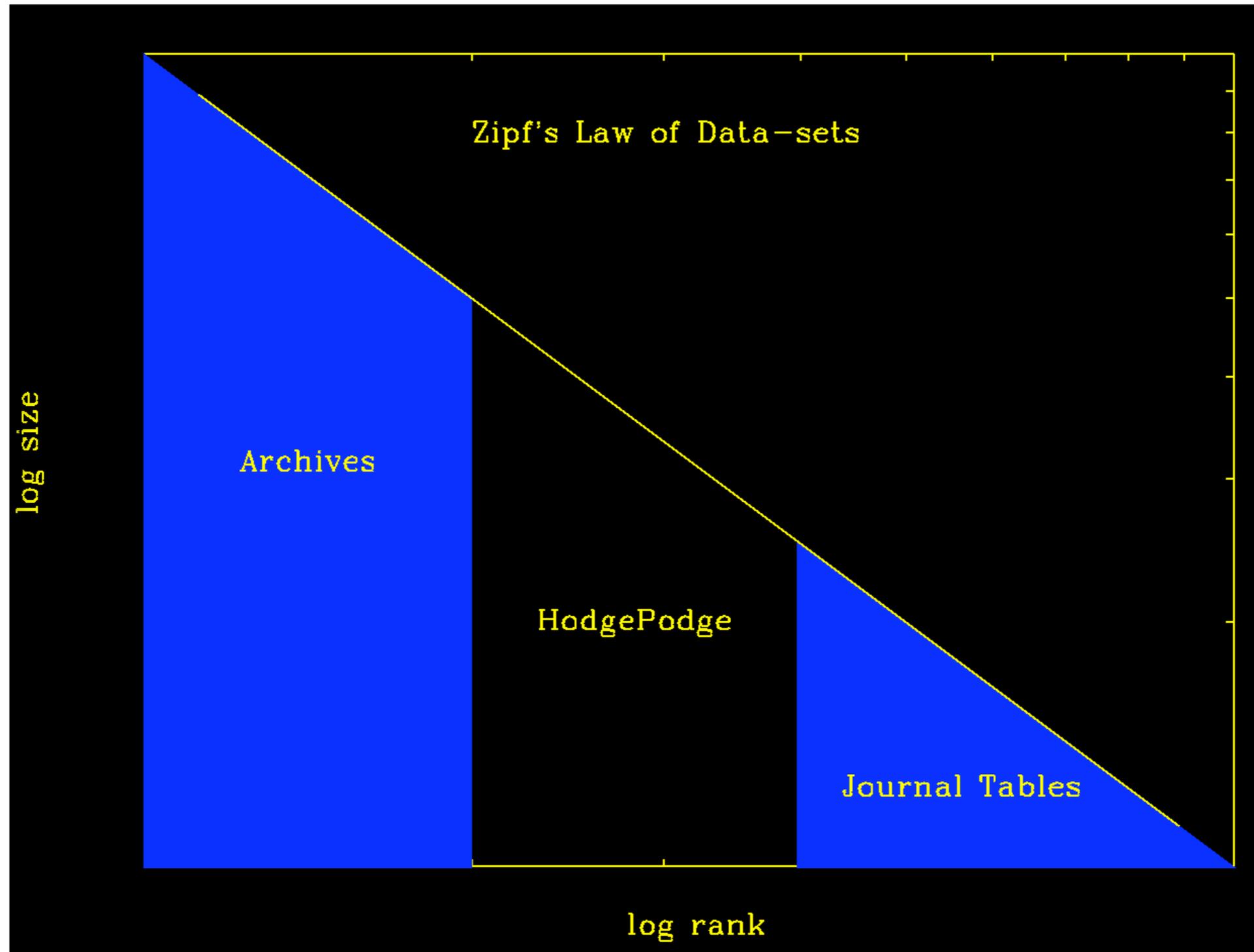
ABSTRACT

We present an overview of data available for the Ophiuchus and Perseus molecular clouds from Phase I of the COMPLETE Survey of Star-Forming Regions. This survey provides a range of data complementary to the *Spitzer* Legacy Program “From Molecular Cores to Planet Forming Disks.” Phase I includes the following: extinction maps derived from the Two Micron All Sky Survey (2MASS) near-infrared data using the NICER algorithm; extinction and temperature maps derived from *IRAS* 60 and 100 μm emission; H I maps of atomic gas; ^{12}CO and ^{13}CO maps of molecular gas; and submillimeter continuum images of emission from dust in dense cores. Not unexpectedly, the morphology of the regions appears quite different depending on the column density tracer that is used, with *IRAS* tracing mainly warmer dust and CO being biased by chemical, excitation, and optical depth effects. Histograms of column density distribution are presented, showing that extinction as derived from 2MASS NICER gives the closest match to a lognormal distribution, as is predicted by numerical simulations. All the data presented in this paper, and links to more detailed publications on their implications, are publicly available at the COMPLETE Web site.

Key words: ISM: clouds — stars: formation — surveys

¹¹ The COMPLETE Web site is at <http://www.cfa.harvard.edu/COMPLETE>.

Where is the Data?



Credit: Todd Vision, Michael Kurtz

Summary of Current Practices

1. Links to on-line data catalogs are listed in reference section (citation)
2. Links to MRTs live alongside with the article
3. Links to individual archival observations (alas few and far in between) are in-lined in HTML article (mention)
4. Links to everything else are a hodge-podge (institutional websites, private pages, “write to us”)

Credit

1. On-line data catalogs: **author gets citation**
2. Journal-published MRTs: **author usually gets citation**
3. Archival observations: **archival gets inbound link, no formal “citation”**
4. Everything else: **currently lost, but should count as citation**

What's missing?

1. On-line data catalogs: better citation guidelines (“cite as”), persistent IDs (may be DOIs, ARKs)?
2. Journal-published MRTs: tracking of re-use in online catalogs, maybe component DOIs?
3. Archival observations: revamp of dataset citation proposal, creation of robust registry
4. Everything else: community-supported repository, with persistent IDs, citation guidelines

Some Tough Questions

- Who should maintain a repository for user-curated data products?
- Who should maintain the registry of all data products, ensuring persistence, multiple copies?
- Who mints and who registers the persistent IDs for data products?
- Who enforces data citation policies and guidelines?

User-generated Data Products

- Fall through the cracks: too big to be “supplementary material” but an integral part of the underlying data products described in paper
- Not just a bunch of tables, FITS files; often full-fledged websites with an underlying database
- Some prototypes now exist:
 - Data Conservancy / arXiv
 - The CfA Astro DataVerse project

Registry of Data Products

- Not the same thing as the VO registry!
- Resources are data products, not services
- Need to deal with long-term persistence, multiple copy problem
- DOI model can work for describing “simple” dataset, should be good enough for datasets published in the literature
- Datasets need to be properly described via registration of metadata

Minting and Registering of IDs

- Archives responsible for the preservation of a data product should create identifiers, register them with metadata (DOI model)
- IDs should be as archive agnostic as possible, but may include some branding if desired (but beware of future-proofing issues)
- It should allow us to map/retrofit existing identifiers into new scheme, e.g.

[ADS/Sa.CXO#obs/123](#)

=> [10.1234/ADS/Sa.CXO;obs/123](#)

[ivo://CDS.VizieR/J/ApJ/715/429](#)

=> [10.1234/CDS.VizieR/J/ApJ/715/429](#)

Publishing and Citation Policies

- Not really part of IVOA's area of expertise...
- BUT as a DC&P issue IVOA should support strong policies in favor of data publication
- Actual refereeing and publication best handled by societies and journals; this is no small task and needs careful consideration
- We (VO archives) should be ready to support publishers' decisions, technical requirements

Upcoming

- ADASS BoF will involve input from community
- AAS is considering supporting a data repository effort, possible opportunity for IVOA involvement
- Need to involve A&A and MNRAS on both data citation and repository issues
- Not clear what resources are available to implement any of this
- Is there a need/use for an IVOA note?