



Everything Everywhere All At Once: How to publish astronomical data

Minh Huynh | 20 July 2023

Australia's National Science Agency





How to publish data

Today I'll discuss:


- Why Data Repositories
- Astro Specific Data Tips
- CSIRO Data Access Portal (DAP) and CASDA





The Old Way

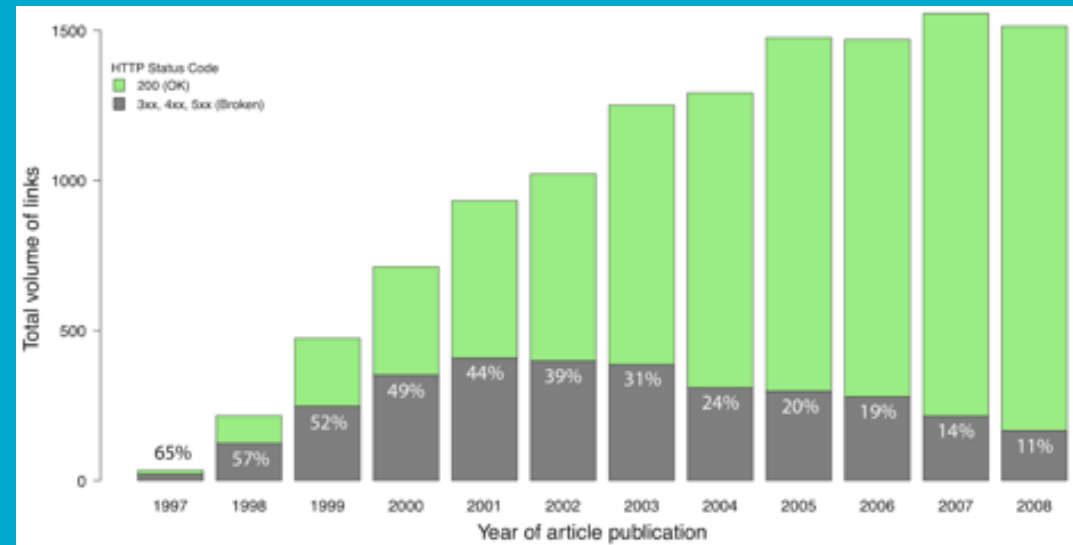
- Data linked to in a paper with URL
- Data available from bespoke website
- Or data may not be available at all

A diagram on the right side of the slide illustrates data sharing methods. It features a central light purple rectangular box containing the text 'dropbox', 'email', 'URL', and 'google-drive' stacked vertically. The word 'dropbox' is in a grey sans-serif font, 'email' is in a purple serif font, 'URL' is in a green sans-serif font, and 'google-drive' is in a green sans-serif font. This central box is partially overlapped by two solid blue rectangular blocks: one above and one below it, creating a cross-like shape.



The Old Way

- One study found that 44% of data links published in the astronomical literature were broken, when examined a decade later (in 2011)
 - papers published by the American Astronomical Society
- “How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers” (Pepe et al. 2014, PLoS ONE), <https://doi.org/10.1371/journal.pone.0104798>



Total volume of external links in AAS articles published between 1997 and 2008, color coded by HTTP status code. Green bars represent accessible links (200), grey bars represent broken links. Analysis done in 2011



Why Repositories and Archives

- Data preservation
 - core trust seal
- Discovery
 - Persistent identifier (DOI) for citation
 - Good description of metadata
 - Interoperability and VO services





Why Repositories and Archives

- Makes data more FAIR (findable, accessible, interoperable and reusable)
 - <https://ardc.edu.au/resource/fair-data/>
- Journals now ask for a Data Availability Statement
 - DOIs are persistent link the dataset used
 - Allows for citations to be tracked
- Some evidence data availability leads to more citations
<https://arxiv.org/abs/1111.3618> (ADASS XXI 2011)



Examples of Repositories and Archives

- VizieR (CDS)
 - catalogues, mostly from published papers
 - can have associated images and spectra
- NASA Extragalactic Database (NASA/IPAC)
 - catalogues
 - positions, photometry, source sizes, and redshifts
 - derived results, e.g. Galactic extinction, velocity corrections
- Skyview (NASA/Goddard)
 - Images, large surveys
- AAO Data Central



Astro-specific tips on data

- See “Best Practices for Data Publication in the Astronomical Literature” Chen et al. 2022 (DOI: 10.3847/1538-4365/ac6268)
- Advice around nomenclature, astrometry, photometry, etc.
- See Checklist in Appendix A

THE ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES, 200:5 (13pp), 2022 May
© 2022. The Author(s). Published by the American Astronomical Society.
<https://doi.org/10.3847/1538-4365/ac6268>

OPEN ACCESS

Best Practices for Data Publication in the Astronomical Literature

Tracy X. Chen¹, Marion Schmitz¹, Joseph M. Mazzarella¹, Xiuqin Wu¹, Julian C. van Eyken², Alberto Accomazzi³, Rachel L. Akeson², Mark Allen⁴, Rachael Beaton⁵, G. Bruce Berriman², Andrew W. Boyle², Marianne Brouty⁴, Ben H. P. Chan¹, Jessie L. Christiansen², David R. Ciardi⁷, David Cook¹, Raffaele D’Abrusco⁶, Rick Ebert¹, Cren Frayer¹, Benjamin J. Fulton², Christopher Gelino², George Helou¹, Calen B. Henderson², Justin Howell⁶, Joyce Kim¹, Gilles Landais⁸, Tak Lo¹, Cécile Loup⁹, Barry Madore^{7,8}, Giacomo Monari², August Muench⁹, Anaís Oberio⁴, Pierre Ocvirk⁴, Joshua E. G. Peek^{10,11}, Emmanuelle Perret⁴, Olga Pevunova¹, Solange V. Ramirez⁷, Luisa Rebull⁶, Ohad Shemmer¹², Alan Smale¹³, Raymond Tam², Scott Terek¹, Doug Van Orsow^{13,14}, Patricia Vannier⁴, and Shin-Ywan Wang¹

¹ Caltech/IPAC-NED, Mail Code 100-22, Caltech, 1200 E. California Boulevard, Pasadena, CA 91125, USA; chenx@caltech.edu
² Caltech/IPAC-NISSci, Mail Code 100-22, Caltech, 1200 E. California Boulevard, Pasadena, CA 91125, USA
³ Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA
⁴ Centre de Données astronomiques de Strasbourg, Observatoire de Strasbourg, 11, rue de l’Université, 67000 Strasbourg, France
⁵ Department of Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, NJ 08544, USA
⁶ Caltech/IPAC-IRSA, Mail Code 100-22, Caltech, 1200 E. California Boulevard, Pasadena, CA 91125, USA
⁷ The Observatories, Carnegie Institution for Science, 813 Santa Barbara Street, Pasadena, CA 91101, USA
⁸ Department of Astronomy & Astrophysics, University of Chicago, 5640 S. Ellis Avenue, Chicago, IL 60637, USA
⁹ American Astronomical Society, 1667 K Street NW, Suite 800, Washington, DC 20006, USA
¹⁰ Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
¹¹ Department of Physics & Astronomy, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA
¹² Department of Physics, University of North Texas, Denton, TX 76203, USA
¹³ NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA
¹⁴ Business Integra, Inc., 6550 Rock Spring Drive #600, Bethesda, MD 20817, USA

Received 2021 June 2; revised 2022 March 21; accepted 2022 March 28; published 2022 May 4

Abstract


We present an overview of best practices for publishing data in astronomy and astrophysics journals. These recommendations are intended as a reference for authors to help prepare and publish data in a way that will better represent and support science results, enable better data sharing, improve reproducibility, and enhance the reusability of data. Observance of these guidelines will also help to streamline the extraction, preservation, integration and cross-linking of valuable data from astrophysics literature into major astronomical databases, and consequently facilitate new modes of science discovery that will better exploit the vast quantities of panchromatic and multidimensional data associated with the literature. We encourage authors, journal editors, referees, and publishers to implement the best practices reviewed here, as well as related recommendations from international astronomical organizations such as the International Astronomical Union for publication of nomenclature, data, and metadata. A convenient Checklist of Recommendations for Publishing Data in the Literature (Appendix A) is included for authors to consult before the submission of the final version of their journal articles and associated data files. We recommend that publishers of journals in astronomy and astrophysics incorporate a link to this document in their Instructions to Authors.

Unified Astronomy Thesaurus concepts: Astronomy databases (83); Astronomical reference materials (90); Astronomical object identification (87); Astrometry (80); Astronomical coordinate systems (82); Photometry (1234); Observatories (1147); Telescopes (1689)



Notes on nomenclature

See IAU recommendations at <https://cds.unistra.fr//Dic/iau-spec.html>



**Specifications concerning designations
for astronomical radiation sources outside the solar system**

How to refer to a source or designate a new one from the Working Group Designations of IAU Commission B2

Contents

- 1. General recommendations
- 2. Case of existing designations
- 3. Creation of new designations
 - 3.1 Acronym
 - 3.2 Sequence
 - 3.2.1 Use of coordinates
 - 3.3 Specifier
 - 3.4 Punctuation and special characters
 - 3.5 Examples
 - 3.5.1 Examples of complete designations
 - 3.5.2 Examples of improper designations
 - 3.6 Helpful Hints
- 4. Advice on designations
- 5. Further information

Short title: IAU Recommendations for Nomenclature

Key words: Designations IAU

(Version August 2018)



Notes on nomenclature

See IAU recommendations at <https://cds.unistra.fr//Dic/iau-spec.html>

The designation of an astronomical source should consist of the following parts :
Acronym ^ Sequence

NGC 205 , PKS 1817–43, CO J0326.0+3041.0, TYC 1234–545–1

Avoid traps!

- Short acronym (minimum 3 letters now)
- Acronym already used (check <https://cds.unistra.fr/cgi-bin/Dic>)
- Rounding instead of truncating coords in name

How to publish at CSIRO

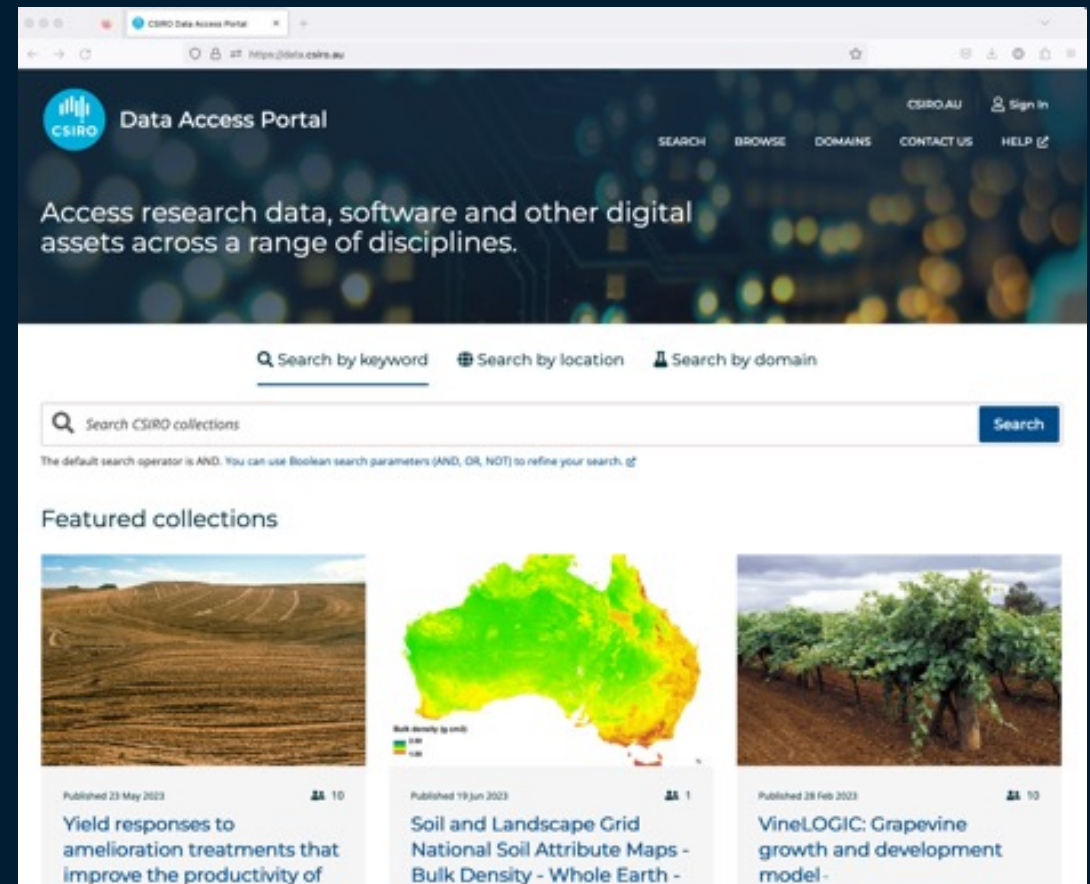




The Data Access Portal

The Data Access Portal is CSIRO's enterprise-wide data repository:

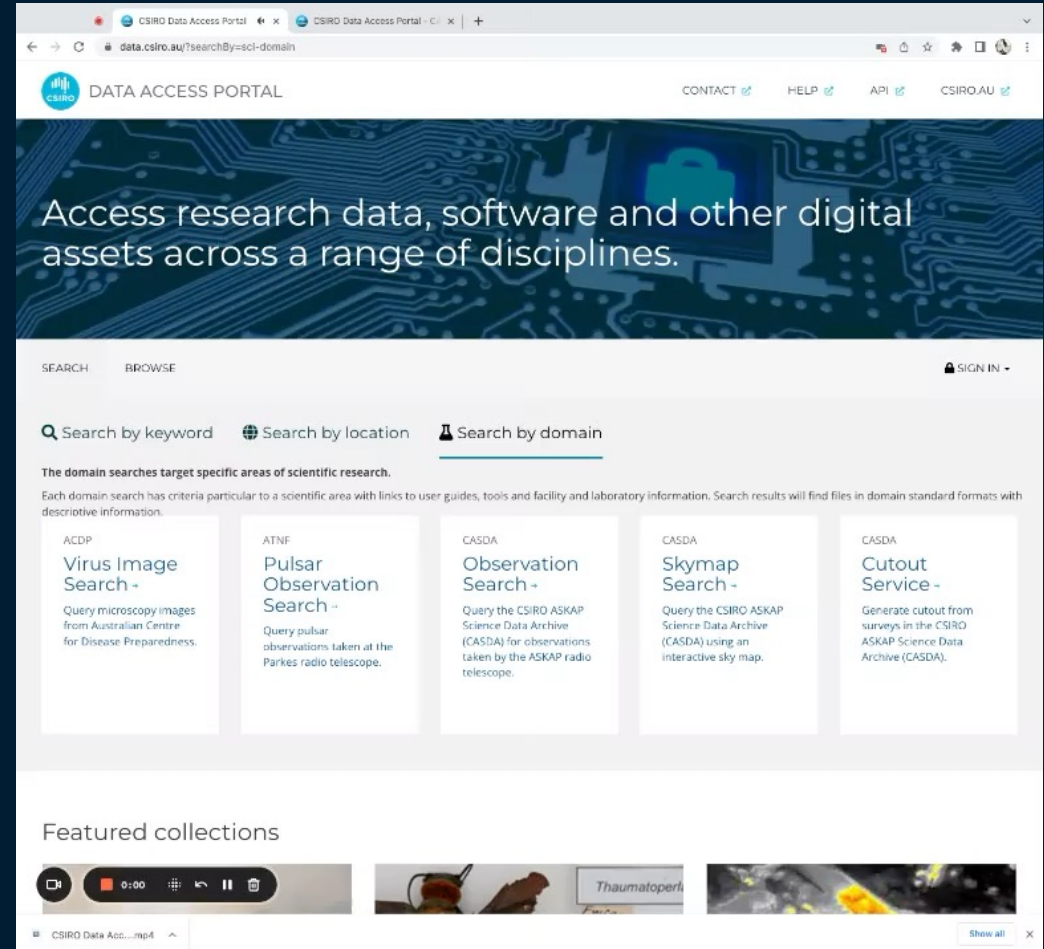
- Publicly released in 2011, many updates since then
- Contains research data (and software) from all CSIRO research areas
- Mission is to provide access to CSIRO's research data, enable reusability, and ensure long-term preservation and persistence.



<https://data.csiro.au>

The CSIRO ASKAP Science Data Archive (CASDA):

- Domain of the CSIRO Data Access Portal (DAP), uses DAP front end
- Archives science-ready data products from ASKAP:
 - Images/cubes, spectra, catalogues, calibrated visibilities from ASKAP ops team
 - **Value-added data from science teams**

A screenshot of the CSIRO Data Access Portal (DAP) website. The page has a dark blue header with the CSIRO logo and 'DATA ACCESS PORTAL' text. Below the header is a large banner with a circuit board background and the text 'Access research data, software and other digital assets across a range of disciplines.' The main content area is white and features a search bar with options for 'Search by keyword', 'Search by location', and 'Search by domain'. Below the search bar, there are five search categories: 'Virus Image Search', 'Pulsar Observation Search', 'CASDA Observation Search', 'CASDA Skymap Search', and 'CASDA Cutout Service'. Each category has a brief description of the search criteria. At the bottom, there is a 'Featured collections' section with a video player and two image thumbnails, one of which is labeled 'Thaumatoperi'.



DAP (classic) vs CASDA, what's the diff?

- DAP infrastructure, servers and storage at CSIRO Canberra Data Centre (CDC)
- CASDA infrastructure at Pawsey (servers and storage),
 - But UI, authentication and authorisation, collection/DOI service at CDC, shared with DAP
- DAP (classic) does not provide Astro specific services:
 - E.g. Cone search, cutout service
 - Virtual observatory services (TAP/astroquery/etc.)
 - Data is just served as a collection of files/directories
- DAP (classic) currently limited to 1.5 TB collections, but with move to S3 (imminent!) this limit will be removed



Data Access Portal Deposits

WHAT

- Can deposit data and software
- Any type of file can be stored
- Make sure files and directories are well-named

WHO

- Any CSIRO staff member, Nexus account
- Deposit permission may have to be activated by DAP team
- I have deposited data for external ATCA users, e.g. ATCA Legacy Project CACHMC



Data Access Portal Deposits

HOW

- Prepare the data and metadata
- Deposits are via UI data.csiro.au
- <https://confluence.csiro.au/display/dap/Create+a+Collection>
- File upload from local laptop (for small files) to SFTP (larger files) and `rclone+SFTP` for files on e.g. petrichor
- For very large collections or files Research Data Support team can help
- Select an approver – myself or Matt Whiting for S&A
- Cite the generated DOI, all collections have a DOI



Data Access Portal Deposits

Demo of UI for DAP Classic deposits



CASDA User Derived (L7) Data Deposits

WHAT

- Catalogues (VOtables), images/cubes (FITS) and spectra (FITS)
- value-added ASKAP data
 - But some other older ATNF surveys

WHO

- Any CSIRO staff member, Nexus account
- Deposit permission will need to be activated
- I have been leading and curating these, due to the CASDA requirements
 - Will look to train members of ASKAp SSTs on how to do this



Data Access Portal Deposits

HOW

- Prepare the data and metadata
- Deposit via UI data.csiro.au
- Files need to be at Pawsey, /scratch on Setonix
- Select an approver – myself or Matt Whiting for S&A
- Cite the generated DOI, all collections have a DOI



CASDA User Derived (L7) Data Deposits

Catalogue requirements:

- The catalogue is provided in VOTABLE v1.3 format
- Three params must be provided:
 - "Catalogue Name" – the name which people will use to query the catalogue in TAP
 - "Indexed Fields" – Which fields should have an index (usually RA, Dec)
 - "Principal Fields" – Which fields should be recommended for returning in each query
- All fields/columns are named
- Field names aren't reserved sql/adql words
- Field names are alphanumeric with underscores only



CASDA User Derived (L7) Data Deposits

Catalogue requirements (cont):

- Minimum of id, RA, and Dec fields with the following (UCDs):
 - meta.id;meta.main
 - pos.eq.ra;meta.main
 - pos.eq.dec;meta.main
- Any other Unified Content Descriptors UCDs used are valid according to <https://www.ivoa.net/documents/UCD1+/20230125/index.html> (the latest astropy has an up to date list)
- Descriptions for each field are 255 characters or less
- **We encourage each field to have a description and UCD but do not require it.**



CASDA User Derived (L7) Data Deposits

FITS images/cubes requirements:

- Single HDU support, information read from header of first HDU
- Axes supported are RA, DEC, GLON, GLAT, FREQ, VELO, VOPT, VRAD, STOKES
- Less than 50,000 files total
- For entry in skymap service, each image must have an OBJECT header value which relates to a single column in the target catalogue



CASDA User Derived (L7) Data Deposits

Demo of UI for CASDA Deposit



Summary

- Publish your data at a repository or archive!
 - Data preservation and discovery
- Preparation of the data and metadata is key
 - Follow IAU conventions/recommendations
- CSIRO Data Access Portal (classic)
 - Any data or software
 - Suitable for processed ATCA/ParkeS data
- CASDA
 - ASKAP level 7 or value-added data
 - VO services
- Come talk to me, happy to help!



Thank you

Space and Astronomy

Minh Huynh

Senior Astronomer and ATNF Group Leader

+61 8 6436 8696

Minh.Huynh@csiro.au

Australia's National Science Agency