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

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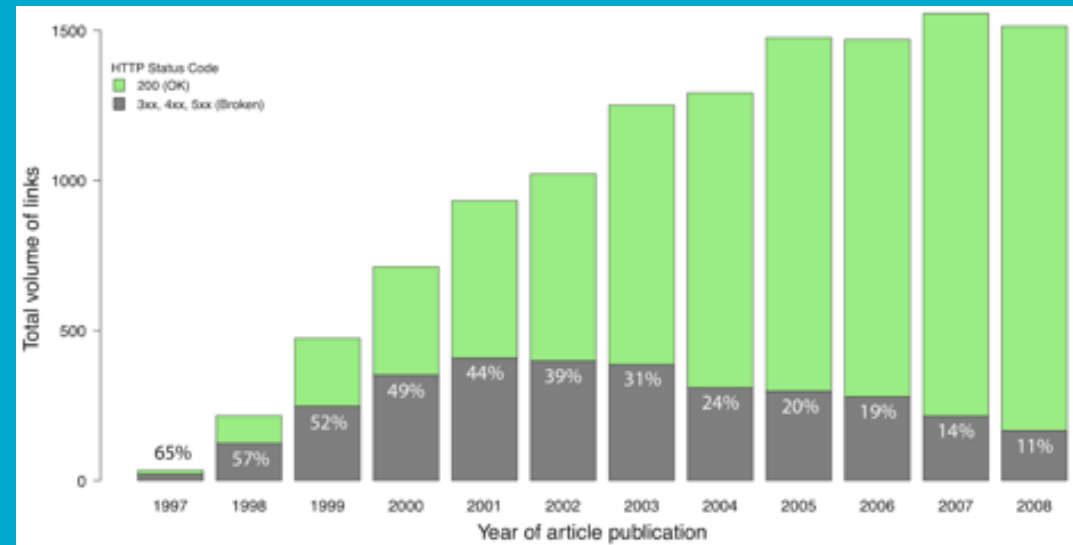
Total volume of external links in AAS articles published between 1997 and 2008, color coded by HTTP status code. Green bars represent accessible links (200), grey bars represent broken links. Analysis done in 2011

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

Notes on nomenclature

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

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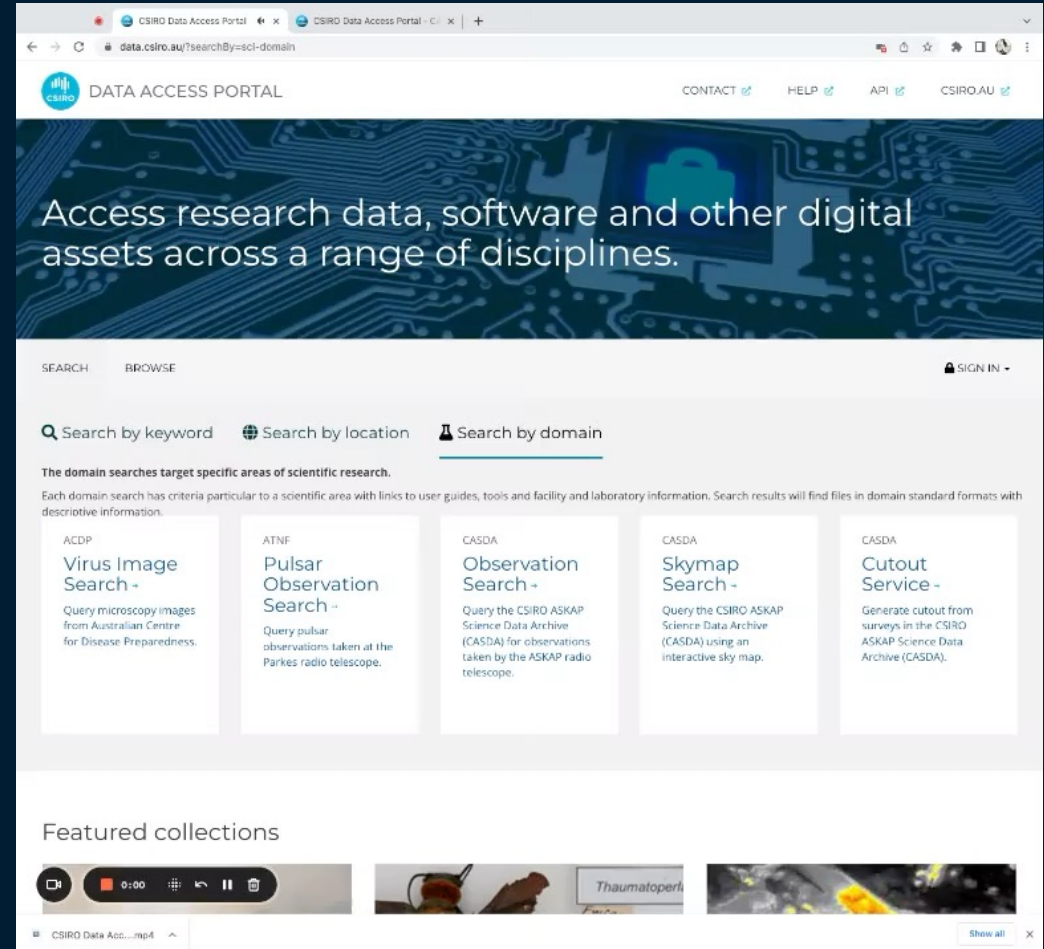
- Publicly released in 2011, many updates since then
- Contains research data (and software) from all CSIRO research areas
- Mission is to provide access to CSIRO's research data, enable reusability, and ensure long-term preservation and persistence.

https://data.csiro.au
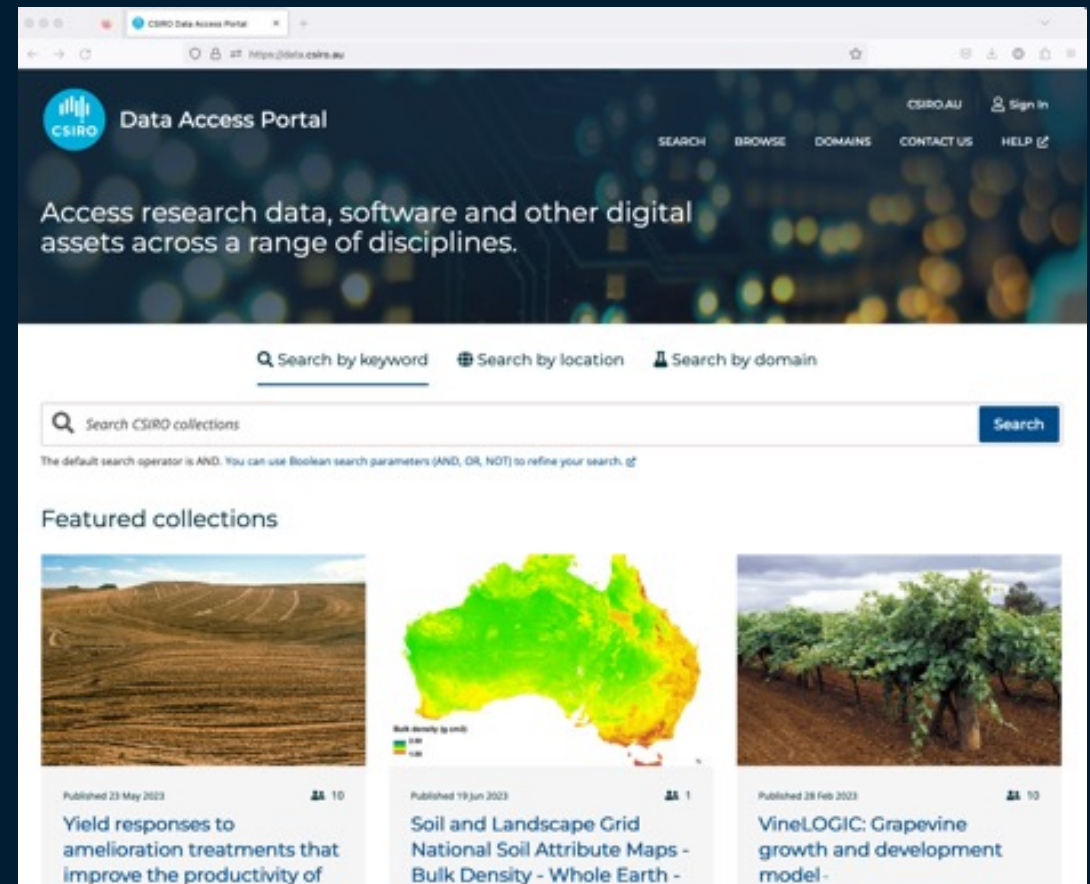
CASDA

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**

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