Publishing Your Data Well

A.K.A: Why can't I just dump a tarball on a website?

data central

Outline

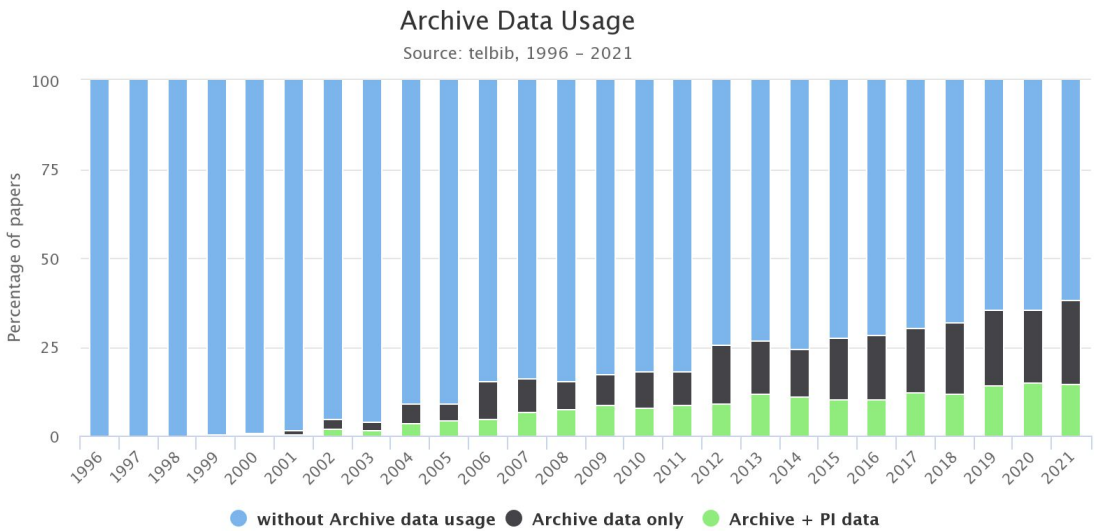
- Why should you publish data
- What do we mean by publishing “well”
- Why you should publish data “well”

data central

Why you should publish your data

- Funding agencies are starting to require it:
  - ARDC now requiring data follow the FAIR data principles—will cover this in the next few slides

- Increasing use of archival/ existing data:
  - ESO Archive Usage
  - Spitzer Publication Statistics
  - Public data gets more citations
- Multi-wavelength/ Multi-messenger astronomy benefits from more datasets



ESO Archive usage from 1996-2021—
https://www.eso.org/sci/php/libraries/telbibstats/archive.php

data central

What do we mean by publishing data “well”

- Data is both FAIR and easy to use
- FAIR:
  - Findable
  - Accessible
  - Interoperable
  - Reusable
- You do not need to worry about your data being FAIR, the Virtual Observatory handles it (almost)
  - The next talk we’ll cover what’s missing, see 2022arXiv220310710O for details
- You do need to think about how to make your data easy to use though
  - We’ll go into the low-level details in the next talk

data central

Why astronomers should care about FAIR data/VO

- The same tools work everywhere:
  - TOPCAT uses the VO to find what's available, and to download and search catalogues
  - DAS uses the VO to query images and catalogues, and to find relevant catalogues
  - Astroquery uses pyvo (which uses the VO) to perform queries and download data
- We can combine datasets to produce new ones
- Funding agencies (e.g. ARDC) pushing/requiring data be FAIR
- BUT setting up and maintaining a VO service is not easy
  - Use an existing VO provider (such as Data Central)

data central

Example: Finding redshifts on VizieR

- Could use well-known catalogues
  - How to discover new catalogues?
- Find all tables which contain a column with the name “z”
  - What about capitalisation?
  - What about “z_best”, “z_fitted” etc.
- Find all tables which contain a column which matches the regex “^z.*”
  - Is the regex correct
  - Regex isn’t easy to use in SQL
- Find all tables which contain a column with the UCD “src.redshift”
  - The VO specifies how to state that a column represents a redshift
  - This is how DAS finds tables with redshift information
- Takeaway: Metadata enables new tools and features

data central

What you should consider to make your data easy to use

- Think about how you want common users to interact with your data
  - If it's a simple type of interaction, a small amount of documentation goes a long way.
  - If it's a complex type of interaction, odds are that there may be tools available to facilitate that. These tools may already have been developed in house by us, or may be widely available as part of the VO and/or astropy ecosystem.
- Don't throw up barriers to intermediate/advanced users
  - Normalised or "tidy" data makes queries both faster and easier
  - Many simple things are easier to work with than a few complex ones
- Easier to use data will be used more widely

data central

What does the VO provide?

- Services for querying
  - TAP: catalogues/tables
  - SCS: cone-search (newer versions allow spatial and temporal queries)
  - SIA: images (plans to build a Simple Data Access Protocol for more complex data)
  - SSA: spectra
- Tools for more complex interactions:
  - SAMP: cross-tool communication
  - MOC: define complex regions—e.g. check for target in survey coverage fast
  - HiPS: explore catalogue/images spatially—see DAS

data central

What does the VO provide?

- New services/tools are being developed all the time, and existing ones adapted/improved
  - Managed by the IVOA: https://www.ivoa.net/
  - Always interested in what astronomers need
  - Has “Interest Groups” to find/assist with specific areas
- Data Central has representatives on two interest groups:
  - Simon (Vice-Chair): Theory
  - Brent (Chair): Time Domain
- Data Central also has broader interaction at the IVOA
  - Present new tools such as DAS and PAWS
  - Suggest improvements/new services: e.g. Simple Data Access
- Tell us what new tools/services you need

data central

What does Data Central provide for data?

- VO Services (see Brent's talk yesterday)
- Web interfaces for browsing/querying/visualising:
  - Schema browser
  - Single Object Viewer (includes visualisations of data)
  - Image Cutout
  - Catalogue Query
  - File Downloads
- Long term documentation space
- Private Data Releases
  - Test/internal team releases
  - Reviewer Access

data central

Why can't I just dump a tarball on my website?

- Not FAIR
  - Can’t be queried
  - Standard tools may not work
- Static—doesn’t gain new features
  - Hard to attract new users of the data
- Can easily disappear
  - Move institution—website disappears
  - University rebranding—website disappears
  - University moves to the cloud—website disappears
  - 2014PLoSO...9j4798P found that nearly 80% of links to personal websites in astronomy papers were dead in 10 years.

data central

Questions?

data central