The Astrophysics Data Access Infrastructure

Peter K. G. Williams (CfA, AAS)
@pkgw • https://newton.cx/~peter/

JSM2020 – 2020 Aug 6

Chris Wiseman, NASA

Astronomical data are worthless.

That is, they have no economic value. No PII either!

This has led to a relatively strong tradition of data sharing. Common model is proprietary access for 1 year, then open access.

Astronomy has generally been a leader in open science, FAIR data, etc.

(Related: few commercial actors in the field. Astronomers all use Macs nowadays since the software heritage is all Unix, not Windows.)

Astronomical data come from space.

Space missions are planned for finite lifetimes, leading agencies to think about post-mission data stewardship.

Combined with *$pace buck$*, this has created a field with a robust set of archive institutions charged with long-term preservation and reuse.

Ground-based facilities do not have the same resourcing, but at least have a model to emulate.

Most astro data fit into a few broad categories.

Many important exceptions, but a few data types dominate:

- Images, and
- Spectra; processed to create
- Catalogs; often analyzed to yield
- Timeseries (“light curves”)

More and more analysis of *cubes* is important (e.g. 2D spatial, 1D spectral) as well as numerical model outputs, trained neural networks, etc.

Astronomy has had strong data format standards.

Common data types, sharing, and archives have all helped astrophysics to enjoy relatively good data standardization.

Most famous being the FITS container format, originally for images but also for tabular data and more. Many standards engineering lessons to learn from its success (nearly 40 years of use!).

Successful standardization is about **much** more than bytes on disk. Need a *data model* that people buy into — e.g. the World Coordinate System (WCS).

Astronomy has led in web-based data access.

Thanks to standard formats and archives, astronomers have been sharing data over the Web for a long time.

- HEASARC at NASA Goddard (1990)
- IRSA associated with NASA infared missions (mid-90's)
- MAST archive associated with Hubble (1997)
- Canadian Astronomy Data Centre (1986)
- ESO Science Archive Facility
- ESAC Science Data Centre

Not only data distribution, but value-add services like LEDA (1983), SIMBAD, and NED.

The “VO” has aimed to unify digital astronomy.

“The *Virtual Observatory* (VO) is the vision that astronomical datasets and other resources should work as a seamless whole.” (ivoa.net)

“Why does each of these online astronomy data archives have its own, different user interface?!?” (me, lots of other people)

Once again, ahead of its time — serious US VO work began ca. 2000.

Both struggles and successes — most archives use VO protocols somewhere, but promises made in 2002 are far from being fulfilled.

These legacy systems are starting to show strain.

The world has caught up — and well-established infrastructure can start to feel more like a hindrance than a help.

Historical archive paradigm: “find your data, download them to your computer”. Rapidly becoming impossible (see: every other talk this session).

Custom data container formats (like FITS) are less and less appealing.

VO protocols are old-fashioned (XML everywhere, SOAP).

Small teams (e.g. simulators) are creating ever-larger, complex data sets but are simply not equipped to provide great FAIR access.

There are still terabytes of data out there that you can download for free!

Data rates are increasing exponentially.

True in many other fields as well, of course. What's unusual about astronomy?

- We’re trying to direct modern data streams into infrastructure that is beloved but was designed 20+ years ago.
- Nature of the data is evolving: growing emphasis on time-domain and multi-messenger astrophysics (see Protopapas on LSST)
- Data-gathering facilities are often extremely remote: either spend a lot on data transfer, or automate front-line analysis
- Simultaneous cultural shift from small teams toward particle-physics-type undertakings.

Archives are becoming “science platforms.”

Virtually every archive sees a future in which they provide computation as well as data — the data are just too big to move.

Everyone is converging on environments based on Jupyter(Lab) and Python.

Challenge 1: we have to rebuild almost all of our data visualization tools from scratch.

Challenge 2: notebooks are not a great way to create software that is reliable and composable.

Astronomy data sets are getting more complex.

And the analysis questions are getting more subtle — science topics like weak lensing and the Epoch of Reionization require exquisite precision.

New tools (formats, protocols, software, services) are necessary to enable the data access and scientific analysis that we want.

We’re likely entering an era that’s a bit more “Wild West” than astronomers are used to — risks, but also rewards.

Astronomers are starting to use “foriegn” tools.

Still a strong “roll your own” / “not invented here” mentality, but more teams are opting for existing solutions.

Especially true for machine learning problems — phew!

But also in data access areas — e.g. LSST alert stream (see Protopapas)

Uptake of cloud computing has been slow.

Cost structure is a major barrier — why pay a monthly bill to Amazon when you can use university and national HPC resources “for free”?

This could change — NSF has considered setting up an intermediary to deal with providers and let researchers approach cloud computing as if it were one more HPC center.

That’s not the only driver. We expect a great deal of science to involve joining billion-row catalogs coming from a variety of missions — only feasible if the underlying data all live in the same place. (Such as one AWS region.)

Here's a summary.

- Astrophysics has a relatively strong tradition of open data sharing, format standardization, and institutional archives.
- There are many online astrophysics archives from which you can download terabytes of data for free.
- The rest of the world has caught up — astronomy's infrastructure is starting to show its age.
- The field is transitioning to the "science platform" era of remote analysis of datasets that are too large to move.
- Astronomers are starting to adopt off-the-shelf technologies for tasks like machine learning; maybe the cloud is next?

Thanks for your attention!

Peter K. G. Williams • @pkgw • https://newton.cx/~peter/

HTML talk info: https://tinyurl.com/htmltalk • *Design credits*: Hakim El Hattab ("white" theme), Julieta Ulanovsky (Montserrat font), Steve Matteson (Open Sans font) • *Tech credits*: git, reveal.js, KaTeX, Firefox developer tools, d3.js, WWT, ThebeLab,

Acknowledgments: this work was supported in part by the American Astronomical Society. We'd all go nuts without ADS and CDS.