Linking Literature to Astronomy Data with DOIs

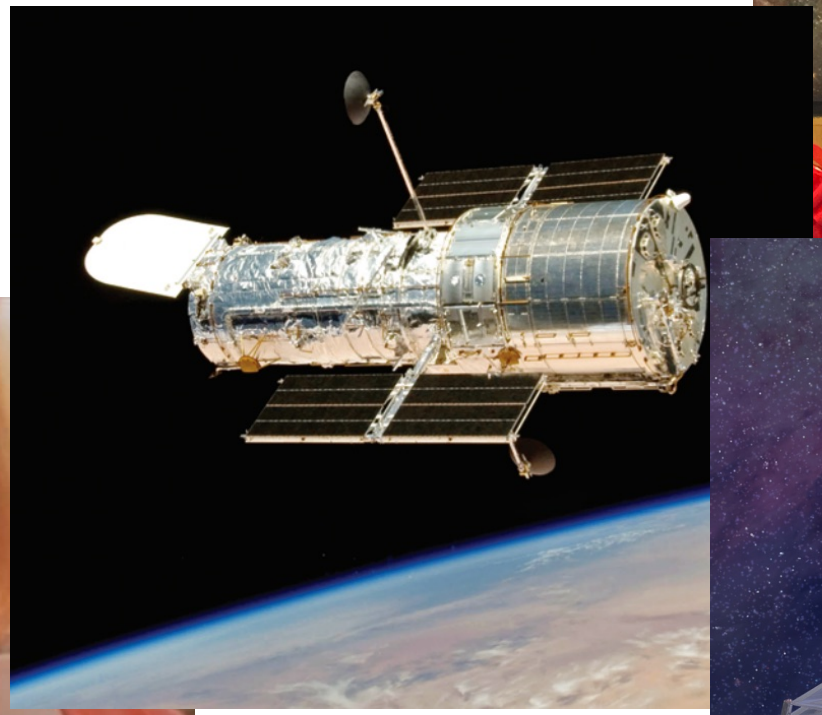
Some Challenges

Sarah Weissman – STScI/MAST

Code4Lib DMV 2017

STScI/MAST – who are we?

- Hubble! JWST!
- … in Baltimore!



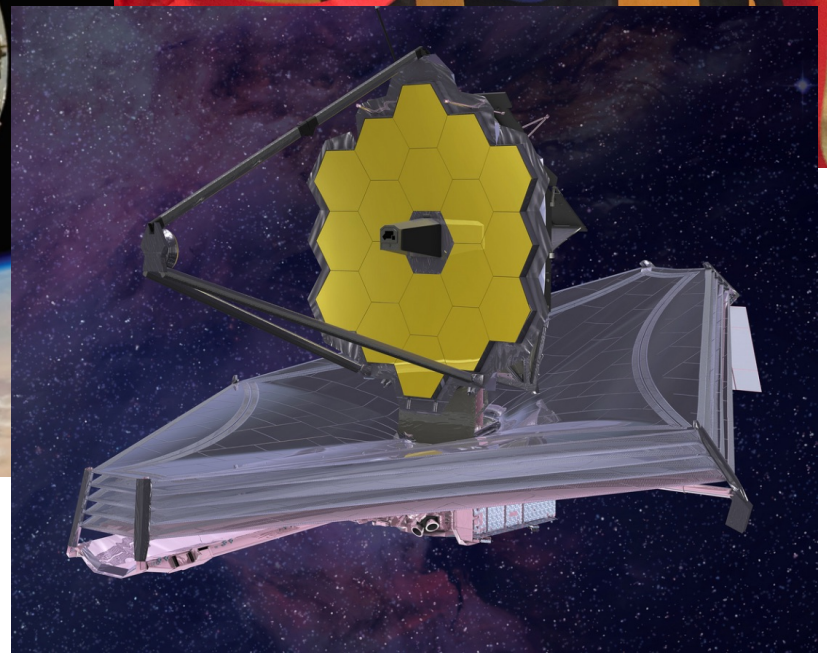


Image credits: HST, NASA 2009 via hubblesite.org; JWST, Northrup Grumman via webbtelescope.org; Divine from Pink Flamingos, via baltimoreorless.com; Mikulski, STScI via YouTube

Data DOIs

- DOI = Digital Object Identifier. It's a permanent link to a digital object.
  - URL + ID + metadata container
- Data DOI demo

fainter magnitudes. As shown in Figure 9, these are also the sources with larger photometric errors, therefore it is not a surprise that these objects were not measured as well as those characterized by higher $Q_{\rm fit}$ values. Sources that have been well fitted by our PSF model are likely stars. If their luminosities have been measured in more than one _flt/_flc image, they have been assigned $\mathtt{f}_{\rm filter} = 1$.

The photometric errors (Figure 9) are determined using the formula $\sigma_{\rm mag} = 1.1\sigma_{\rm flux}/{\rm flux}$, where $\sigma_{\rm flux}$ is the standard deviation of the independent measurements. For the sources detected in just one exposure, it was not possible to measure a photometric error. We assigned to these stars the most probable error for their magnitude and marked them using the flag $\mathtt{f}_{\rm filter} = 2$ if they were characterized by a good quality of the PSF fitting ($Q_{\rm fit} > 0.75$); otherwise we used $\mathtt{f}_{\rm filter} = 5$. Sources with photometric errors smaller than 0.25, but with a poor fit of

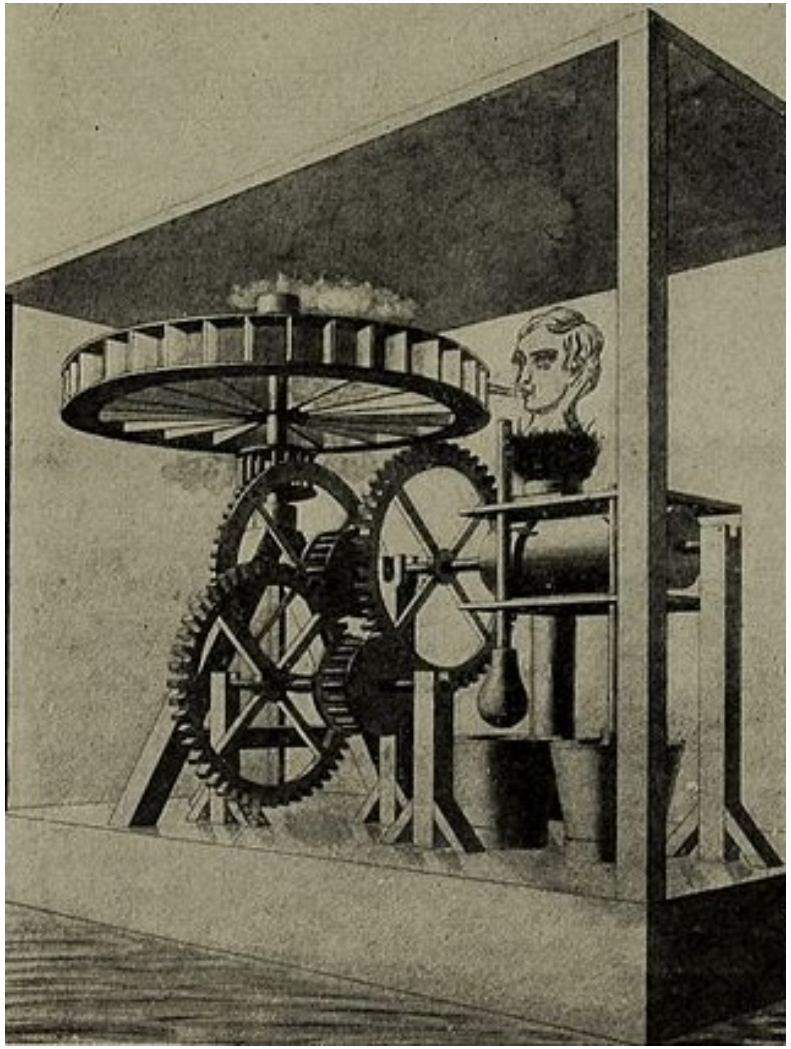
The final catalog is available for download from the Mikulsky Archive for Space Telescopes (MAST) at https://doi.org/10.17909/T9RP4V. For each of the HTTP filter, we also created a larfge mosaic using astrodrizzle.[22] Each mosaic can be downloaded from the url https://doi.org/10.17909/T9RP4V as well.

4.1. Artificial Star Tests

Artificial star tests are a standard procedure to assess the level of completeness and accuracy of a photometric analysis. The tests are performed by inserting stars with known positions and fluxes into the data set, and then repeating the photometric analysis as was used for the real stars. The selection criteria applied to the observed catalog to discard spurious detections are applied to the recovered artificial stars as well.

What are DOIs good for?

- Globally resolvable (kind of like URLs)
- Machine readable (like URLs)
- Persistent links (as long as you update them)
- Usually come packaged with some kind of metadata (if you can find it)
- Not easily bookmarked
- Obscure their destination (like bit.ly)
- Have to be updated
- Meant for machines, not humans (10.7059/T9G0bled33)



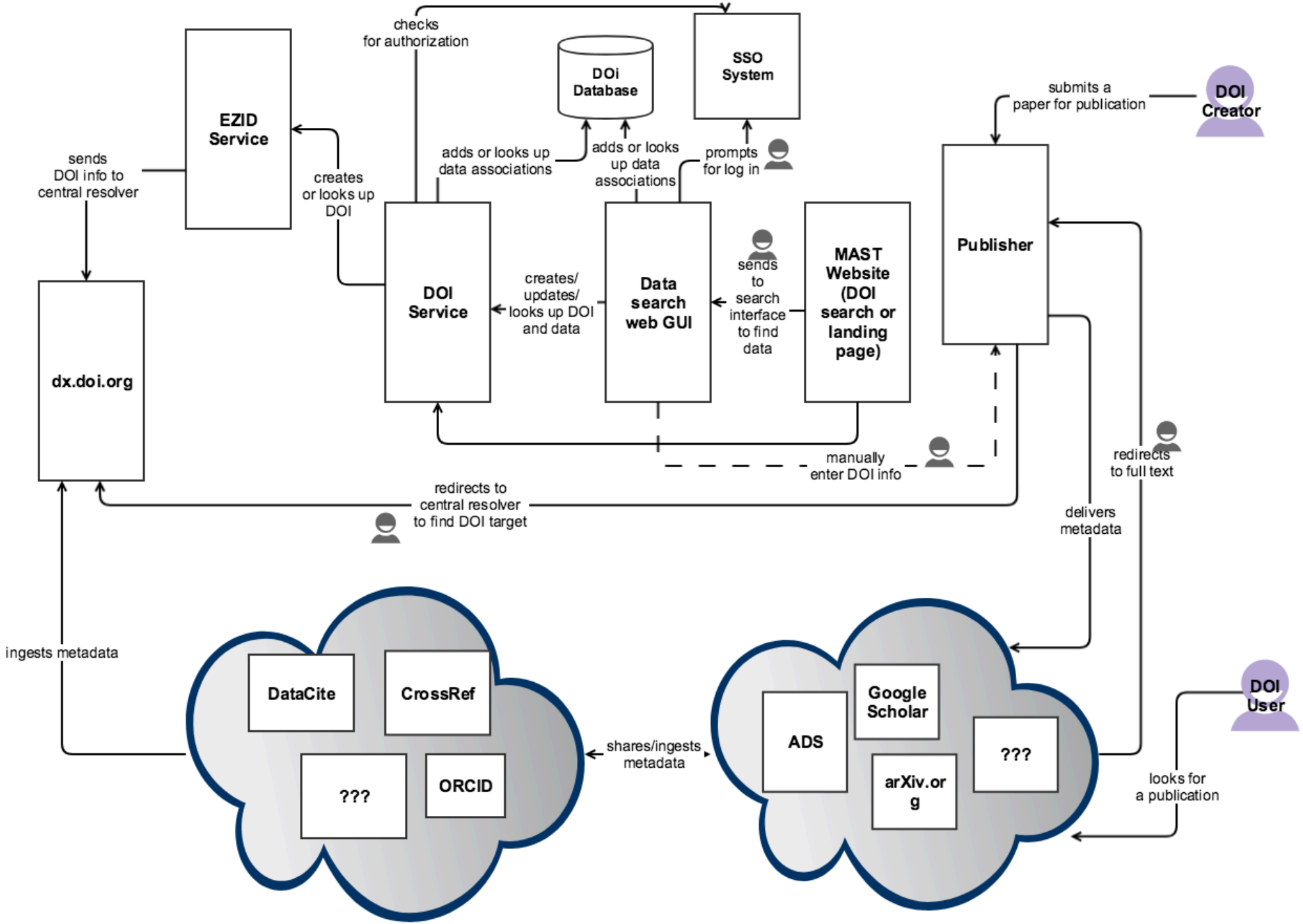
By Williams, Henry Smith, 1863-; Williams, Edward Huntington, 1868-1944, joint author [No restrictions], via Wikimedia Commons

Why Data DOIs

- Allow astronomers to link to their data in a standardized way
  - Convenience
  - Reproducibility, openness
- Make Telescope bibliographies easier
  - Largely a manual process currently
  - Archive planning
  - Justify funding

Data DOIs at MAST

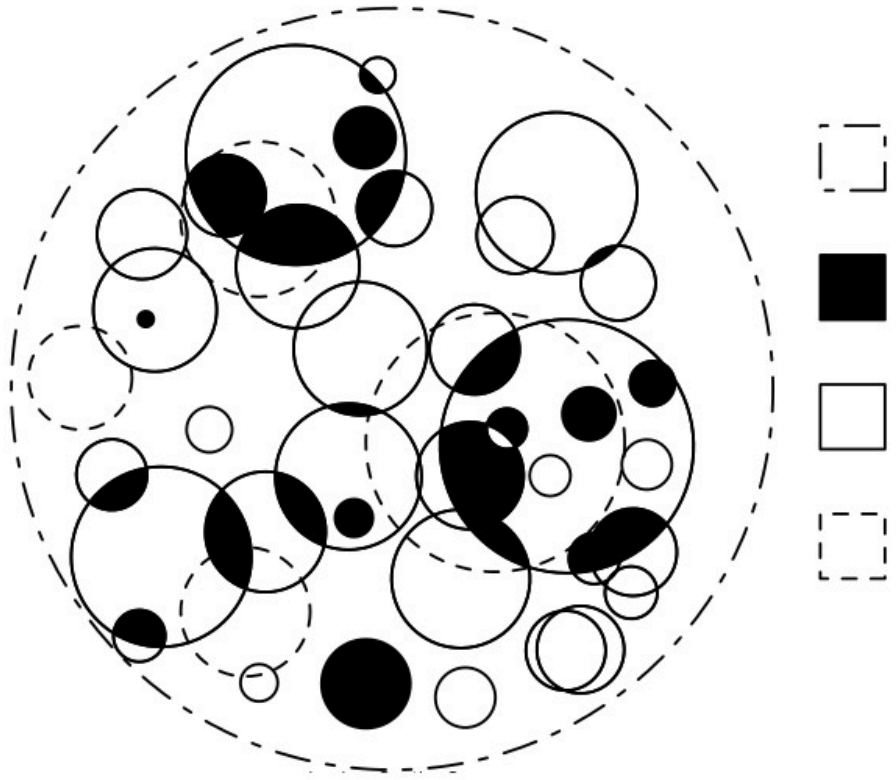
- Collaboration between MAST, AAS Publishing and the STScI Library
- Debuted our DOI service in April 2016, currently in Beta mode
- ~14 STScI authors have created DOIs for publication
- Fall/Winter 2017 - Plan to open service to 12 other institutions.
- Links:
  - http://archive.stsci.edu/doi/search/ (Main DOI entry point)
  - https://mast.stsci.edu/portal/DOI/help (DOI API documentation)


checks
for authorization
DOI
Database
SSO
System
EZID
Service
submits a
paper for publication
DOI
Creator
sends
DOI info to
central resolver
creates
or looks up
DOI
adds or looks up
data associations
adds or looks
up data
associations
prompts
for log in
Publisher
dx.doi.org
DOI
Service
creates/
updates/
looks up DOI
and data
Data
search
web GUI
sends
to
search
interface
to find
data
MAST
Website
(DOI
search or
landing
page)
manually
enter DOI info
redirects
to full text
redirects to
central resolver
to find DOI target
delivers
metadata
ingests metadata
DataCite
CrossRef
???
ORCID
shares/ingests
metadata
ADS
Google
Scholar
arXiv.or
g
???
DOI
User
looks for
a publication

Challenges – Permanence & Uniqueness

- How to make sure that your DOI links keep working?
  - Landing page + service (demo)
  - YALI – yet another level of indirection

- How to link to ”data”?
  - Data is often the not-well-defined glob of stuff (files, database records)
  - What level to link to your data?
    - Observation, data product
  - Once a scientist downloads data, they typically transform it, so it no longer resembles its form in the archive.
  - Had a data model (CAOM), so we used it
    - Even given this, things are messy. ID formats not well defined and not actually unique!



Challenges – Buy in from publishers

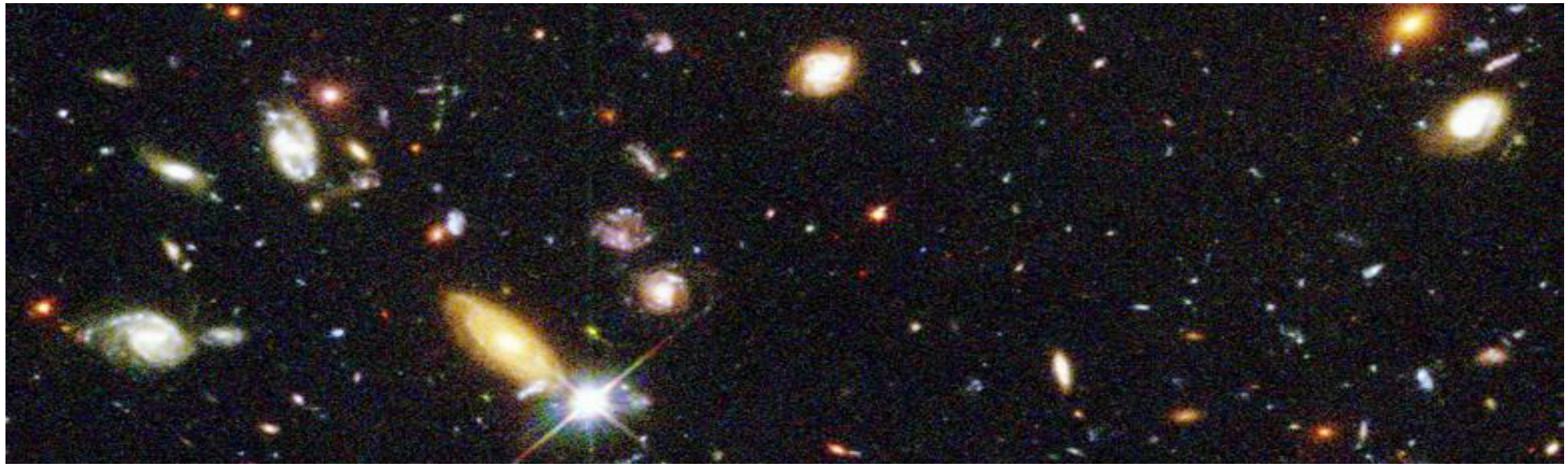
- Luckily we had a good working relationship with AAS Journals and EJ press
- Luckily the world of astronomy wrt data and publishing is relatively open. (E.g. http://adswww.harvard.edu/)
- Publishers aren't building their own software.
- Build relationships with each publisher
- (Publisher has to build up programs with each data center.)

Challenges – Metadata

- Are a number of standards for metadata – DataCite, ERC, DC, CrossRef
- We went with DataCite **BUT** we don’t want data DOIs to be first class citable objects.
  - Ad hoc collections of data that could in theory change
  - Usually when an astronomer publishes a data set there is a paper and THAT should be cited
- DataCite has domain-specific elements (relatedIdentifierType), which makes it hard to use for general purpose metadata.

Challenges – Large Data

- Limitations on our software (Javascript) only allow users to work with so much data at once.
- Large data sets like catalogs can contain millions, even billions of rows, how to efficiently represent any subset of this data?

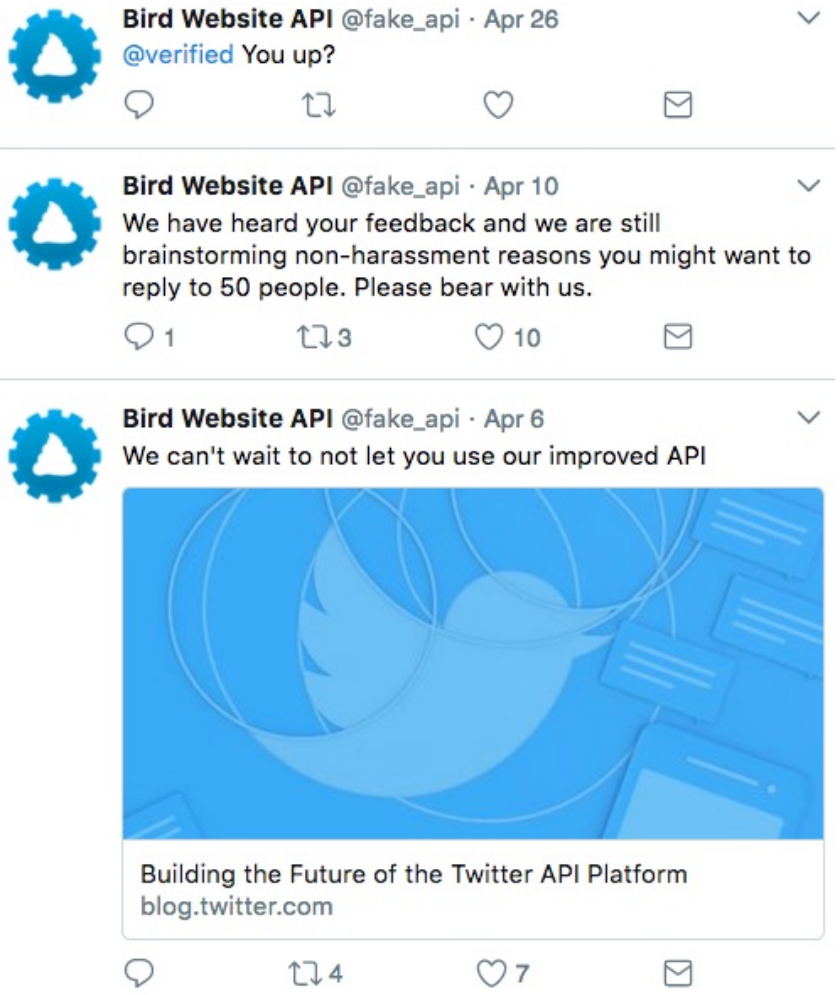


Robert Williams and the Hubble Deep Field Team (STScI) and NASA

Challenges – API

TFW:

- You want your DOI landing page URL to contain the DOI, but it doesn't exist yet.
- Your API is really a thinly veiled proxy for another API.
- Your API is supposed to be generally applicable, but it's actually inextricably linked to this gddmn JavaScript GUI.
- Your identifiers aren't recognized by DataCite so you have to use a customized metadata format.



Future Work

- Get out of Beta. Expand to more users.
- Fully integrate with our data search tool (mast.stsci.edu). Right now we are just a client.
- Provide more links between related DOIs, related literature.
- **Use data tagging to build tools for data enrichment.**

Questions?

- sweissman [at] stsci.edu