EPJ Web of Conferences **186**, 10003 (2018) https://doi.org/10.1051/epjconf/201818610003
LISA VIII

Elevating MAST-Data Publications with Digital Object Identifiers (DOIs)

Jenny Novacescu[1,⋆], *Josh E.G.* Peek[1,⋆⋆], *Sarah* Weissman[1,⋆⋆⋆], *Karen* Levay[1,⋆⋆⋆⋆], *Scott* Fleming[1,†], *Elizabeth* Fraser[1,‡], and *Jonathan R.* Hargis[1,§]

[1] *Space Telescope Science Institute (STScI), 3700 San Martin Drive, Baltimore, MD 21218, USA*

Abstract. The use of digital object identifiers (DOIs) to identify data sets used in original research allows peer reviewers and journal editors to more easily validate research methods and verify results. Fellow astronomers can duplicate results or expand on initial findings when the precise data used in a research project are identified. Precise identification of data may allow archives and observatories to better understand how the community is accessing and combining its data to reach new scientific conclusions. Earlier studies have suggested that papers with linked data are more highly cited in the literature, providing motivation for authors to adopt more stringent and thorough data citation practices.

1 Background

Since late 2015, the Space Telescope Science Institute (STScI) and the American Astronomical Society (AAS) have been working together on a pilot project which provides submitting authors based at STScI a means to assign a digital object identifier (DOI) to the data set(s) referenced in original research articles.

The Space Telescope Science Institute, which is operated by AURA for NASA, was founded in 1982 to oversee the scientific and data archiving operations of the Hubble Space Telescope (HST) and will manage the flight and scientific operations of NASA's latest flagship mission, the James Webb Space Telescope (JWST), scheduled for launch in 2018. The Barbara A. Mikulski Archive for Space Telescopes (MAST) is located at STScI. MAST collects and provides access to astronomical data from over 20 missions, with a historical focus on scientifically related data sets in the optical, ultraviolet, and near-infrared parts of the spectrum. Archived data comes from missions including Hubble, Kepler/K2, and GALEX. TESS and JWST data will also be housed in MAST.

⋆ e-mail: library@stsci.edu ORCID: 0000-0002-8523-015X
⋆⋆ e-mail: jegpeek@stsci.edu
⋆⋆⋆ e-mail: sweissman@stsci.edu
⋆⋆⋆⋆ e-mail: klevay@stsci.edu
† e-mail: fleming@stsci.edu
‡ e-mail: fraser@stsci.edu
§ e-mail: jhargis@stsci.edu

AAS was chosen as the initial partner for this pilot project because a high percentage of papers which cite MAST and specifically HST data are published in AAS journals, including Astronomical Journal (AJ), Astrophysical Journal (ApJ), Astrophysical Journal Supplements (ApJS), and Astrophysical Journal Letters (ApJL). Of the over 15,000 papers identified since 1991 that use HST observational or archival data from MAST, over 55% were published in AJ or ApJ (including ApJS and ApJL). Archival data accounts for nearly two-thirds of all publications related to the Hubble Space Telescope.

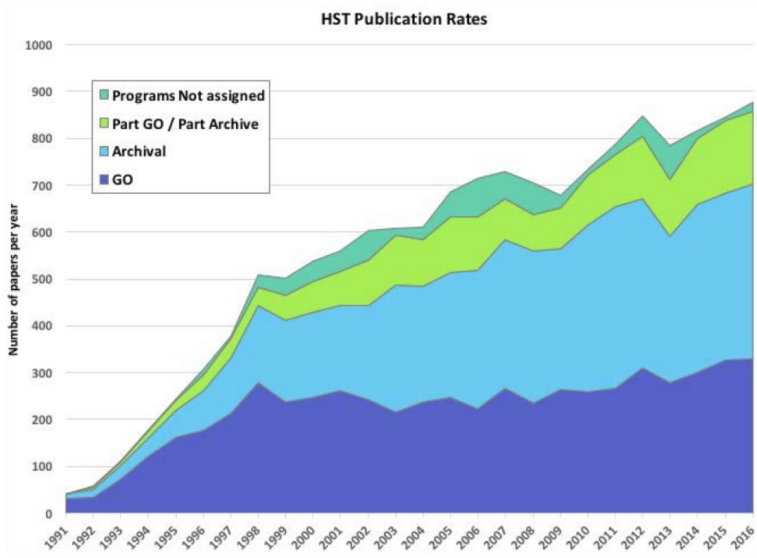

Figure 1. Hubble Space Telescope Publication Rates from 1991-2016, divided by guest observer (GO), archival (AR), combined (GO/AR), and Program not assigned [unidentifiable] data use.

Much of the text of this conference proceeding was derived or quoted in part from a 2017 unpublished, collaborative white paper available in Authorea by the named authors and AAS pilot project contributors. This white paper was drafted in preparation for the expansion of the pilot program and the LISA VIII presentation. Interested readers may contact the authors for the latest version of the working paper, *Digital Object Identifiers for Astronomical Data Archives and Journals: Some Initial Principles Developed at MAST*, which is continually updated.

2 Definitions

Digital object identifier (DOI): the American Psychological Association defines a DOI as "a unique alphanumeric string assigned by a registration agency (the International DOI Foundation) to identify content and provide a persistent link to its location on the Internet".[1] The International DOI Foundation further expands on this definition, noting that "A DOI is a Digital Object Identifier. That is a Digital Identifier of an Object, not an Identifier of a Digital Object... an actionable, interopera-

[1] What is a digital object identifier, or DOI? http://www.apastyle.org/learn/faqs/what-is-doi.aspx

ble, persistent link."[2] For further information on DOI conventions and applications in publishing and science, refer to the *Driven by DOI* brochure.

3 Goals and Objectives of the DOI Project

MAST and the STScI Library have worked collaboratively since the inception of both departments at STScI. In 2015 a DOI Working Group was formed consisting of staff from both departments, along with an associate astronomer with an interest in data citation. The Working Group has since expanded to the newly formed Data Sciences Mission Office (DSMO) at STScI. When the DOI Working Group was first created, MAST and the STScI Library recognized shared concerns with data citation in their work flows and bibliographic indexing efforts.

The pilot group was motivated to implement a DOI-based model due to earlier research by Pepe et al. (2014) [2] which showed static url data links decay over time. In this earlier study, it was found that 44% of data links from a decade earlier were broken.

A partnership with AAS was formed, and the following two goals were identified.

Primary goal: to provide authors who use MAST data a simple and elegant means to identify data analyzed in their original research publications. Doing so:

- Simplifies refereeship and data editing from the journal publisher's perspective
- Allows readers to verify, cite, and reproduce results reported, and build upon findings to produce new results
- Provides an accurate accounting of the types of data retrieved from an archive; this is important for metrics, internal reporting, and funding purposes
- Saves staff time at the bibliographer/library level and increases precision; the author as opposed to the bibliographer is in the best position to identify the data used

Secondary goal: To create a model for best practices in data identification and encourage other astronomical archives and our publishing partners to adopt the use of data DOIs in their own workflows.

Related to this second goal, STScI and AAS are operating on the premise that data identification will become increasingly important with the rise of survey telescopes and proliferation of astrophysical data, as outline by Zhang and Zhao (2015)[3].

4 Principles

The DOI-data set pilot was designed with the following mutual principles in mind for STScI and AAS:

4.1 Journal publishers and archives must actively solicit DOIs for data identification.

The data-hosting archive must provide a simple and elegant tool within its interface if authors are expected to use the service.

[2]Driven by DOI https://www.doi.org/driven_by_doi/DOI_Marketing_Brochure.pdf

4.2 Minimal integration is best. This model allows for interoperability among other publishers/archives in the future.

DOIs are passed between MAST and AAS by simple cut-and-paste into a webform. There is no handoff of data happening behind the scenes. This has the major benefit of easy federation, and no additional standards. The expectation is that in the future AAS Journals will work with other archives to integrate DOIs into articles; similarly MAST will partner with other journals. We could not see a motivation for a more complex system.

4.3 Fixed DOIs vs. custom DOIs should be made available.

Fixed DOIs refer to High-Level Science Products (HLSPs) such as catalogs, surveys, and entire Kepler quarters, which are housed in MAST. In many cases, specific large data sets are used as a whole in a manuscript. Authors have the ability to identify HLSPs using pre-assigned, fixed DOIs. Having a set of fixed DOIs for popular products eliminates the problem of minting multiple DOIs to refer back to the same observations. By assigning a consistent identifier in advance to an HLSP, MAST has given authors a means to provide a persistent link to the data and encourages accurate citation and proper acknowledgment.

In contrast, an author may also need to refer to sets of previously unrelated observations. The MAST DOI Portal tool[3], built by S. Weissman and T. Donaldson, allows users to concatenate observations and make a new DOI at will. Custom DOIs allow the author to aggregate observations and data sets within the entire MAST Portal which might otherwise appear unrelated. Authors are given flexibility to create one DOI for all observations analyzed in a research publication or multiple DOIs for different sets of observations.

4.4 Data DOIs are not first-class citable objects on their own.

While DOIs are often used to link to and identify first-class citable objects, e.g. journal articles, MAST and AAS are in agreement that data DOIs are not first-class citable objects on their own, and thus do not show up in the bibliography.

The logic behind this decision is that an author wishing to return to the data set for future study should cite the paper in which the data set was first identified and analyzed, along with the full context of the data use. Data set DOIs are thought of as "permalinks" for data, and are used to point to existing data in much the same way that simple URLs or DOIs are used to point to journal articles today.

4.5 DOIs refer to the described data set.

Because the intention of the DOI is to describe the data set analyzed in the paper, we do not force the DOI to be forever fixed in the case where the author made some kind of mistake in DOI generation. Our protocol is for MAST to check with the AAS before changing the content of the data set to which the DOI is linked, but we generally believe users should be allowed to edit their DOIs via a mediated process to match the content of their papers.

[3]MAST DOI Portal https://mast.stsci.edu/portal/Mashup/Clients/DOI/DOIPortal.html

5 DOI Minting Process During Paper Submission

In an effort to encourage more observatories and journal publishers to adopt DOI implementation, we are outlining the simple workflow we have established in consultation with the AAS.

5.1 Initiating the form

The author begins paper submission on the eJournal (EJ) Press website to submit to AAS titles. The EJPress submission form asks whether data from the MAST archive was used in the publication.

See further discussion below in *Pilot Outcomes and Lessons Learned* regarding how the question is phrased on the submission form. At this time, only submitting authors whose email domain ends in @stsci.edu are prompted. STScI is currently working with the AAS to expand the domains and invited institutions.

5.2 Determining eligibility

The author specifies whether MAST data was used.

"No" reroutes the author back to finish the paper submission process on EJPress. "Yes" routes the author to an MAST DOI Home page.[4]

5.3 Selecting the DOI type

From the MAST DOI Home page, authors are asked if they used:

1. a collection of specific, curated observations (custom DOI);
2. data from a High-Level Science Product (fixed DOI type);
3. a catalog, e.g., Kepler/GALEX (fixed DOI type); or
4. a large, clear sub-section of a catalog, e.g. a quarter of Kepler long cadence data (fixed DOI type).

Authors who select options 2, 3, or 4 are prompted to select a pre-assigned, fixed DOI(s) from a list. Authors who select option 1 are directed to the custom version of the MAST Portal which allows them to select observations used in their research. A future release of the MAST Portal may allow users access to their download history, but at this time users must re-create their search in order to locate the observations used in their research. As noted, authors have the liberty to mint a single DOI for all observations or a subset of DOIs for different sets of data.

5.4 Metadata assignment

When creating a custom DOI, the author submits basic metadata such as their name, manuscript title, and an optional description of their data. Other metadata, such as date created and data set ID, are auto-assigned.

STScI is investigating future uses and expansion of DOI metadata fields, though most relevant metadata about the individual observations such as instrument and wavelength are already stored in MAST and do not need to be replicated.

[4]MAST DOI Home https://archive.stsci.edu/doi/search/index.html

5.5 Form submission

Once the custom DOI(s) is created, the author is taken back to the EJPress submission form where they cut and paste the DOI(s) and complete the manscript submission. The author also receives an automated email with their information and summary of the DOI metadata.

At this point, the author has completed their end of the process.

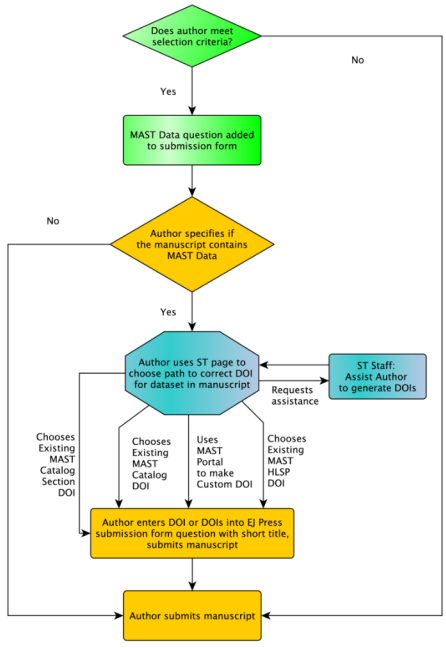

Figure 2. A workflow diagram for an author interacting with the EJPress submission form when submitting to a AAS journal. Green indicates EJPress site functions; yellow indicates author actions on the EJPress site; blue indicates actions on the MAST DOI site.

Another consideration in developing workflows between archives and publishers is the role of the journal publisher and/or data archive in reporting non-compliance. STScI and the AAS came to an agreement that AAS staff would initially track instances of non-compliance so they could report back to STScI. STScI is then responsible to follow up with authors to find out why they elected not to mint a custom DOI or select a fixed DOI, or were unable to do so. Tracking non-compliance at the start of the pilot is essential as it allows your institution to contact authors who were eligible to mint a DOI, but chose not to. Non-compliance also helps the institution find out what part of the process prevented or deterred the author from using the service so the user interface can be enhanced or simplified as needed. Was the initial question about data use unclear, for example? Were instructions confusing? What technical challenges did the submitting author experience when trying to create a DOI(s)? Starting with your own institute allows you to use local staff as a test bed for the DOI service, but you should keep in mind that internal staff may be more savvy at navigating your data archive than the general community.

6 Pilot Outcomes and Lessons Learned

Since the launch of the pilot in early 2016, fifteen STScI first-authored papers used the MAST DOI Portal to identify data sets and observations used. Two of the fifteen authors made use of HLSP pre-assigned DOIs and the remaining thirteen created custom DOIs. One of the thirteen custom DOIs was used to collate over 14,000 rows of data, demonstrating that this process allows for succinct citation of large and complex data sets. With marketing and user-assistance on the part of the local DOI Working Group, the overall compliance rate for eligible papers was 75%. Non-compliance occurred with less than five submissions. In these cases, STScI authors submitted AAS manuscripts that contained MAST data but elected not to use the DOI minting service.

Authors of the earlier papers were confused by the way the initial question was phrased on the EJPress submission form. Our original question was: *Does your manuscript **directly refer to data** from Hubble, Kepler, GALEX, IUE, or other data in MAST?*

Our revised question is now: *Does your manuscript **use or analyze data** from Hubble, Kepler, GALEX, IUE, or other data in MAST?*

Acknowledgments

The authors thank Jill Lagerstrom, former Chief Librarian at the Space Telescope Science Institute, for her ground work on the initial DOI planning, protoyping, and implementation. We wish to also thank our contacts at the American Astronomical Society: Julie Steffen, Director of Publishing; August (Gus) Muench, Journals Data Scientist; and Greg Schwarz, Journals Data Editor. We extend our appreciation to Tom Donaldson and Amanda Marrione in the Archive Sciences branch at STScI for their work developing and testing the DOI minting infrastructure in MAST and Randy Thompson in the Archive Sciences branch for his efforts on web integration.

This research has made use of NASA's Astrophysics Data System (ADS) Bibliographic Services.

References

[1] E.A. Henneken, A. Accomazzi, *Linking to Data: Effect on Citation Rates in Astronomy* in *Astronomical Data Analysis Software and Systems XXI*, edited by P. Ballester, D. Egret, and N.P.F. Lorente (2012), Vol. 461 of Astronomical Society of the Pacific Conference Series, p. 763–766

[2] A. Pepe, A. Goodman, A. Muench et al. , PLOSl One **9**, 104798 (2014)

[3] Y. Zhang, Y. Zhao, Data Science Journal **14**, 11 (2015)