Arfon Smith, Data Science Mission Head – Space Telescope Science Institute, Editor-in-chief – Journal of Open Source Software

Space Telescope Science Institute was founded in 1981 to help build and run the most famous astronomy mission ever, the Hubble Space Telescope. Since 1990 we have operated Hubble on behalf of NASA and serve the global astronomical community who make use of this flagship facility. A key part of our work as the *science operations center* for Hubble and the soon to be launched James Webb Space Telescope (JWST), is to ensure the continued scientific legacy of the missions. Capturing and preserving the data associated with Hubble is the responsibility of the Barbara A. Mikulski Archive for Space Telescopes (MAST) which is the archive for Hubble and more than twenty other mission datasets. MAST currently holds data from 21 missions and surveys and with a data volume of close to 3 petabytes is a major infrastructure support effort in and of itself.

The last decade has seen a dramatic shift towards science being more open and collaborative. Open source software is now the dominant mechanism for developing scientific computing tools, and the ecosystem of software packages available to astronomers to support their research is now very broad and deep. Astronomy-specific packages such as Astropy and their affiliated suite of tools offer a rich set of functionalities for anyone to use, and technologies such as Jupyter notebooks and collaborative platforms such as GitHub have given scientists new ways to share software with their peers.

Like many areas of science, astronomy has also been moving towards collecting large, homogeneous samples of data, suitable for a broad range of different science investigations, and usually with little or no embargo or restrictions on their use. Common datasets encourage sharing between peers, and missions such as Kepler, K2, and TESS have vibrant communities of astronomers developing novel tools and using state of the art data science techniques to analyze mission data, often sharing these products back with the community.

This switch to community-generated, open source software means that staying abreast of the latest and greatest tools, technologies, and techniques for analyzing and interpreting astronomical data is an ongoing and growing challenge. Additionally, many of the biggest scientific opportunities arise when large quantities of archival data are simultaneously analyzed using advanced statistics and data science techniques. These two factors combined create a significant and growing accessibility and knowledge infrastructure challenge for many astronomers as knowing which tool or technique might be best for your problem, and finding sufficient computational resources for carrying out your scientific investigation can often rely upon professional networks, personal connections, and the ability to attend the right conferences or collaboration workshops.

At the same time, it is clear that in the next decade, archival data such as that held at MAST is going to be a critical factor in the ability of the community to characterize progenitor sources for transient sources from instruments such as LSST and LIGO. Classifying and understanding the

transient sky is going to rely upon data collected decades earlier and the value of these datasets is going to rely upon the ability of any member of the community to understand the unique characteristics of particular datasets and execute legacy software to interact with these data.

Further out in the 2030s and 2040s, potential future flagship missions such as LUVOIR[1] will drive a renewed interest in the UV photons collected decades earlier by Hubble. How can MAST preserve software, capture the knowledge held by today's astronomers and preserve it for a future generation?

Essentially MAST, and by extension the astronomical community, face two key knowledge exchange challenges: One the one hand, maximizing the scientific productivity of current and future missions requires the broadest possible fraction of the astronomical community to be using modern, community-built, open source tools and applying modern data science techniques in their research. On the other, MAST must ensure that we continue to preserve the scientific legacy of our existing mission datasets at a time when the number of individuals with first-hand experience of working with these data is declining. In both cases these skills and expertise are not evenly distributed which means scientific opportunities are inevitably being missed.

What are the most urgent research questions to address about KI? Why?

The biggest challenge we face at MAST is what kind of combination of technologies and knowledge support infrastructures are necessary to maximize the scientific return of archival data in the coming decades, especially with a fast-moving ecosystem of tools, technologies, and techniques for generating scientific knowledge. How can MAST preserve software, capture the knowledge held by today's astronomers and preserve it for future generations?

Identify a KI whose survival is under threat.

MAST is not under any immediate threat – NASA funds a small number of archives and continued support of these facilities is considered part of NASA's core mission. More generally though, astronomical data generated by other facilities (e.g. SDSS and soon LSST) does not have funding for long term preservation and curation which is a major risk for future science.

a. What led to these threats? Over what time frame?

These threats have arisen because of the emergence of long-lived datasets that have ongoing value to research communities. In astronomy, these datasets have been generated for generations, but in the last couple of decades, large surveys such as the Sloan Digital Sky Survey (SDSS) and Pan-STARRS have amplified these challenges significantly.

[1] https://asd.gsfc.nasa.gov/luvoir/

b. What actions or changes in circumstances might lead to its survival?

Szalay and Barish[2] make the case for a coordinated strategy for preserving scientific data with some kind of *Data Trust* that would not only sustain the data, but also preserving the *expertise* which is often associated with people working on these projects.

c. What will be gained or lost, by whom, if this KI fails to survive?

In the extreme scenario, where all archival astronomical data were lost, the scientific productivity of new facilities would be substantially reduced. This would be especially bad for facilities focused on the *transient sky* (e.g. LSST, LIGO) where classifying and understanding transient sources often relies upon having historical data available for the progenitors (what was there before).

How do KI spread information? Misinformation? Alone and in combination with other infrastructures?

MAST has a responsibility to both preserve data but also the expertise for working with them. Archive Scientists at MAST engage actively with the communities we support through email, web-based resources, the scientific literature, and in person at conferences. More generally, our library staff curate the scientific literature associated with Hubble and NASA's ADS performs a similar function for the wider astronomical literature. MAST is therefore a component of a broader set of knowledge infrastructures that support the global astronomy community.

[2] https://ui.adsabs.harvard.edu/abs/2019BAAS...51g..16S/abstract