MAST Users Group Annual Meeting Report
December 2016

A lively and constructive meeting of the Mikulski Archive for Space Telescopes (MAST) Users Group (MUG) occurred at STScI on Dec. 15-16, 2016. This year MUG members included Todd Tripp (chair, University of Massachusetts - Amherst), Martha Boyer (STScI), Dan Foreman-Mackey (University of Washington), Steve Howell (NASA-Ames), Knut Olsen (NOAO), Dan Weisz (University of California – Berkeley), and Gail Zasowski (STScI). The MUG was introduced to Arfon Smith, the new director of the Data Science Mission Office, and we had opportunities to have frank conversations with Arfon about a variety of topics of interest and concern. These conversations were very useful for quickly giving the new director a sense of MUG priorities (and vice versa).

This report summarizes the MUG reactions to a variety of topics that were discussed and provides feedback to MAST regarding priorities and concerns from the perspective of this representative group of MAST researchers. The order of this report follows the order of presentations during the meeting. We begin by commending MAST for another successful year characterized by significant progress in a variety of areas (e.g., significant improvement of the MAST Portal for searching the archives and acquiring data, preparation of massive datasets such as PanSTARRS for release to the public, improved calibrations and development of higher-level tools, etc.). That said, clearly much work remains to be done, and we offer both endorsements and reservations in this report to help MAST to make decisions about priorities with input from the broader community.

Todd Tripp is rotating off of the MUG this year. The MUG elected Dan Foreman-Mackey to chair the committee in 2017.

The NASA Astronomical Virtual Observatory Collaboration

The MAST Users Group (MUG) was presented with background on NASA participation in the Astronomical Virtual Observatory (VO) and confronted with a set of questions:

1. Do you know what the VO is?
2. Can you use the VO?
3. Have you used it personally? Do you know others who have used it?
4. Are there additional things the VO should be doing?

The discussion that ensued was freeform in nature and did not exactly unfold as direct responses to these questions, but came close.

The MUG first acknowledged that at its inception, the VO was billed as a facility akin to a telescope that would allow users to easily retrieve a wide array of observations for arbitrary targets. The VO's inability to deliver on this early promise has led to

the broad perception by the community that the VO was a failure. The fact that the original domain of the US VO, us-vo.org, has been taken over by a non-astronomical enterprise adds credence to this perception. However, the MUG also acknowledged that a large number of widely used tools and services use protocols developed by the VO, including MAST, NED, Gaia, CADC's CANFAR effort, Aladin, ds9, and TOPCAT, as well as new efforts like the NOAO Data Lab and LSST's alert system. The strength of the VO is thus not in developing complete software tools, but in providing widely agreed-upon protocols and standards for individual data centers to implement as they develop their data facilities and tools. In technical terms, the protocols developed by the VO tend to reside in the transport layer of software systems, such that systems developed by different sites are able to communicate freely, rather than the application layer, where the core software that users interact with resides. Thus, in the view of the MUG, NASA's continued participation in the VO is very important and a strategically prudent decision that enables the US space-based astronomical community to draw from a wide range effort and expertise as it designs its data archives and platforms.

The MUG did have a couple of recommendations for highlighting the importance of NASA's VO effort. First, it suggested that MAST should identify a way to highlight the VO compatibility in the branding of its services, to make the breadth of the effort more visible to the community. While the term "Virtual Observatory" may be a misnomer for what VO protocols actually provide, the MUG acknowledged that changing the name of the VO is likely not possible. Second, it suggested that MAST draft a memo to describe the extent to which MAST services rely upon VO-developed protocols, to help educate the community about the importance of the effort and suggest areas for further work.

The MAST Newsletter and Forums

After a three-year hiatus, MAST has restarted the MAST newsletter with the intention of frequently issuing short articles to update the community on various topics. With the creation of the Data Science Mission Office and the arrival of Arfon Smith, the first director of this office, the MUG strongly supports this reboot of the newsletter. This will provide MAST with a valuable mechanism for communicating with the astronomical community about the purpose and priorities of the Data Science Mission Office as well as important developments/news from the archive and critical public relations and outreach. For example, during this meeting of the MAST Users Group, we discussed the pros and cons of the Virtual Observatory (VO) initiative, as summarized above. The Virtual Observatory has failed to achieve some of its original objectives and suffers from considerable stigma and misunderstanding in the broader community. Nevertheless, the VO has been quite useful for establishing standards and protocols that have been highly beneficial for the MAST mission. By design, this success of the VO is invisible to most astronomers, but funding is required for the behind-the-scenes work of the VO, which continues to develop and refine a universal framework for science archives. In order support continued funding for the VO, the MUG encourages the archive to

communicate about how and why the VO is important for archival researchers, and a newsletter provides an effective vehicle for this type of communication.

The MUG also discussed the utility of MAST forums for sharing information. In principle, the forum format can be highly effective for answering questions from archival researchers and solving advanced problems that might not be addressed in data handbooks and standard documentation. This communication format has been successful in other disciplines such as computer science/programming. However, the MAST users community is not as large as the computer programming world, and the challenge is to build enough interest in a forum (and awareness that the forum exists in the first place) so that it is useful and worth the effort that MAST would invest in its development. It is not entirely clear how MAST can attain this "critical mass" for new forums, but the newsletter could help to get this effort started, and the MUG thinks that it might be worth a try. However, some members of the MUG stated their frustrations with the meandering flood of information (much of which is not relevant or useful) that sometimes appears in forums. If MAST forums catch on, we encourage the archive to curate these discussions so that the most important information is easy to find. Indeed, if really useful ideas are developed in forum discussions, it might be useful to reproduce this information in relevant documents such as instrument or data handbooks.

MAST Website Redesign

A presentation was made about an entirely new look and new functionality for the MAST website that is currently in preparation. The MUG appreciates this effort to redesign the website appearance and navigational structure in order to provide more straightforward access to both data and documentation. A two-tier approach to presenting documentation, along with a push to standardize the mission pages, should provide MAST users with a more streamlined and internally consistent experience. The MUG was shown two drafts of the front page design and liked many elements of both; members felt that some content, like the mission statement, could be de-emphasized and moved to make room for content and links that a user is more likely to be looking for. In discussing how to decide which content would be most useful for the most users, the MUG wondered if it would be possible to have a customizable sidebar (or similar element) that contained "Favorite" links shown when a user was logged in. These could be links to particular mission pages, datasets, news updates, reference material, or other spaces that users might bookmark during a project. The MUG members also like the concept of the uniform mission icons, though caution should be taken to use sufficiently unique icons.

Metadata Improvements, including Digital Object Identifiers

A lively discussion of options for providing more sophisticated metadata occurred, with an emphasis on the concept and implementation of Digital Object Identifiers (DOIs). DOIs are already widely used in the publishing world, but this tool could also be very useful for archival research. For example, when a paper is published, it

could include a DOI that summarizes the archival information about datasets that were used in the paper and thereby provide future researchers with a convenient and efficient means to retrieve the same data for additional analyses. This could save a lot of time and effort.

The workflow for minting Digital Object Identifiers (DOIs) for archiving and sharing datasets selected in The Portal was demonstrated. The goal of this project is to encourage reproducibility and ease tracking of publications that the use MAST provided datasets. The MUG was impressed by the prototype implementation of this concept and encourages the continued development of the DOI service. So far the MAST DOI interface has been rolled out for STScI employees publishing in the AAS journals and the plan is to enable the interface for more users shortly. One feature that will greatly benefit users of this service (and the community at large) is the "saved search" feature of The Portal. This would allow users to save their queries when they download the data and then mint a DOI for that specific query without having to reconstruct it at a later date.

Some concern about versioning was discussed in the context of DOIs. This is a relevant discussion across all of the MAST products but it is especially important when it comes to DOIs because the explicit role of a DOI is to point to a static object. The current implementation of DOIs for MAST are limited by the fact that most data products are not versioned – updated reductions overwrite all previous versions of the data. This means that a DOI will always resolve to the most recent version of the data instead of the version that was available when the DOI was created or, more specifically, when the data were downloaded. This shortcoming might interfere with the goal of enabling scientific reproducibility. The argument was made by a MAST representative that the costs of archiving all versions of the data would be too great for the added benefit and the MUG is willing to defer to that assessment, but we hope that this decision will be very carefully considered. At the least, we suggest consideration of the inclusion of metadata about the data reduction version on the DOI landing page.

Archiving of Data from the Transiting Exoplanet Survey Satellite (TESS)

The NASA TESS mission promises to be a great boon for general astrophysics in addition to exoplanet science. The primary mission data sets are the ~200,000 two-minute "postage stamps" aimed at exoplanet transit detection. These will be valuable for a wide variety of science applications as well as exoplanet research. However, in addition to the postage stamps, other TESS data products will be of high value to the community. The MUG discussed the value and challenges of archiving some of the additional TESS data including the TESS input catalogue (TIC) and the full-frame images.

The TIC has ~600 million entries and each of these is matched to a number of other MAST served multi-wavelength catalogues, providing the largest and best SED catalogue of the entire sky. The full-frame images, large portions of the sky obtained

with a 30-minute cadence for days to weeks to years, will likely be very interesting for a variety of research purposes. We encourage MAST to provide these data as quickly as possible to the community. We also support the plan to provide "cutouts", i.e., regions of interest (RA, DEC, size) as time slices, perhaps in the same format as K2 data so all the tools developed for K2 photometry can simply be carried over to TESS data.

The MUG also feels that access to the TIC through the MAST data portal would have great utility. This allows the TIC to be not only searched in clever ways but connected (via VO protocols) to numerous other catalogues stored in US and non-US archives.

We applaud the MAST efforts related to TESS and believe funding of such tools is a high priority for the coming years.

The Survey of MAST Users

The MUG heard the results of an email/Surveymonkey survey administered to MAST users regarding the Hubble Legacy Archive (HLA) transition to the MAST Portal and other features of MAST. In response to previous MUG recommendations, this survey was kept to six questions and focused on user opinions. The MUG applauds the effort to directly solicit user feedback on MAST changes, but cautions against putting too much weight on the opinions of a small number of users, who most likely do not represent the user population in general (see related recommendations in the HLA section). The MUG encourages the team to explore ways to increase response rates, such as longer windows to respond and careful timing around proposal deadlines, and whether using checkboxes with pre-written answers would provide a more manageable dataset. But the MUG recognizes that these surveys are each designed for specific needs, and had no significant criticism of the process in general.

Using Gaia to Improve HST Astrometry

The MUG was briefed on ongoing efforts to introduce better absolute astronomy into new and archival HST images. HST is capable of excellent relative astrometry (<0.4 mas), but absolute astrometry is typically substantially worse (0.2" - 1"). The limiting factors in absolute astrometry have been the absolute positions of the guide stars, calibration of the fine guidance sensor position, and calibration of the observing instrument in the focal plane. With the advent of accurate absolute positions from the ESA Gaia Mission, the dominant uncertainty in absolute astrometry is the focal plane solution.

There is concerted effort to improve absolute HST astronomy by matching sources in HST images to those in external catalogs/images, notably Pan-STARRS, which will also be hosted by MAST. This work has benefited from development of the Hubble Source Catalog (HSC), which provides catalogs of (bright) sources in HST images.

Initial absolute astrometry improvements have been achieved by cross-matching HST/HSC catalogs to external, calibrated catalogs and have been propagated into recent/new images in the HLA. These improvements will continue to be propagated into HLA data products, but not into all products accessible in MAST. The MUG has some concern about the inconsistency between these various data products, and more concern that this difference is not made explicitly clear in the MAST/HLA documentation. This issue aside, the MUG views efforts to improve absolute astronomy of HST images as incredibly valuable, and believes that adding absolute astrometry to all HST products (including older data such as WFPC2 and NICMOS) would substantially enhance the legacy value of HST.

The Hubble Legacy Archive and the Hubble Source Catalog

The MUG was pleased to hear the plans to improve and expand the data products available in the HLA, including super mosaics and improved astrometry from the HSC. The current plan is to retire the HLA interface in late Spring or early Summer 2017 due to increasing costs to maintain and support the aging hardware and software. The full functionality of the HLA interface will be absorbed into the MAST portal before this transition. Some respondents to the recent MAST users survey (see above) were displeased with the plan to retire the HLA interface, and the MAST team asked the MUG for input on a possible "softer" transition wherein the HLA is supported for a longer period (around 6 months). The MUG feels that this extension is not necessary and is unlikely to assuage the concerns of users who prefer the current HLA interface. The MUG recommends that MAST retire the HLA interface as planned. However, the MUG urges MAST to defer the HLA decommissioning until after the Cycle 25 HST proposal deadline. Some of the MAST-use statistics presented in this meeting show a clear surge in HLA use in the weeks before HST proposals are due, and we conclude that this is still an important tool for proposal preparation and should be maintained until this need subsides after the deadline. The MUG also recommends that an alert should be placed on the current HLA interface to notify users that the interface will soon be retired, preferably with a link to the portal and/or to instructions on how to search the HLA within the MAST portal. This will ensure that users who bookmark the HLA interface (and therefore bypass the HLA homepage news items), are informed of its retirement. It could also be useful to somehow inform HLA users about how to use the "album view" mode of the MAST Portal, which will provide most of the capability in the HLA interface.

The MAST Subscription Service

A brief presentation was made regarding the development of a new type of "subscription service" to notify archival researchers about a variety of information such as arrival of new data, recalibration of existing data, or the appearance of new publications that make use of observations from a particular program. While MAST and STScI have provided this type of notification in the past, this has been primarily done with an inflexible email notification system. The new service will have a browser-based, self-service model that will enable researchers to have more control

over the type of information they receive, the frequency of notifications, and collaborators that will also receive the updates. While we didn't see a full-blown demonstration of this new effort, the MUG agreed that this sounds like a valuable service, and we look forward to learning more about it as this tool continues to be developed.

HST Instrument Calibration Updates

We heard from the four HST instrument leads concerning the current state and plans for the COS, STIS, ACS, and WPF3. All are performing well considering their on-orbit age.

COS had some serious wavelength calibration issues that were discovered in the near-ultraviolet channel last year, and the COS team has worked hard to implement fixes. The presentation from the COS team demonstrated that this calibration problem has been mostly solved. Tests of the fixes are on-going, but once the dispersion solutions are shown to be optimally calibrated, an alert about the issues and solutions will be sent to the community, and the fixes will be used in reprocessing of COS data.

The COS far-ultraviolet wavelength calibration has also been improved during the last year, but this calibration is still imperfect. The COS team continues to work on improving the wavelength calibration of the far-ultraviolet channel, e.g., by implementing a correction for the "walk" problem as a lookup table. The MUG was very pleased to hear about these efforts and urges the COS team to keep working on the FUV calibration. This is relevant to the MAST mission – one of the foremost priorities of a data archive is to provide the highest quality calibrations and data products, and now is the time to optimize the COS wavelength calibration while the necessary expertise is available at STScI.

STIS, after almost 20 years in space, seems to be working well. The focus is changing, based on focal plane modeling and aging of the instrument, and such changes are being monitored in detail and dealt with. These are important changes to detail, even if small, as they are important for small slits (due to light loss) and can mimic flux changes (variability). The STIS Echelle blaze function has been seen to shift and fixes are being pursued.

ACS, at 15 years old, also looks good. Small increases in dark current and thus zero point calibrations are needed and the current PSF is broader than initially obtained (as compared to "Tiny Tim" models). This causes less encircled energy at a small level. The MUG was pleased to hear about improved quantification of the extended halo of the point-spread function in the solar-blind channel and improvements in ACS drizzling and cosmic-ray clipping procedures.

WFC3 too looks good for its age with only ~1% changes (degradations) in the usual factors that affect CCDs in the hostile environment of low-Earth orbit, e.g., CTE and pixel damage issues. New zero points and flat field images have been produced.

High-Volume Data Acquisition Issues, STScI Science Cloud Evaluation, and SciServer

High-Volume Data Acquisition

Prompted by presentations about data access through the discovery portal, MUG members raised the issue of access to MAST data outside the portal. Currently MAST users are directed to download MAST data onto their local machine via the discovery portal web interface. However, it was and remains unclear whether this is the optimal approach for downloading large amount of data due to interface, security, and efficiency issues. Currently, users must manually add data products to their 'cart' for download onto their local machines. However, for large datasets, manually clicking on items is cumbersome in the limit of large data. A related issue is that large data processing is migrating toward cloud and supercomputing infrastructure, many of which have multi-factor authentication requirements and do not support current MAST data delivery interfaces (e.g., direct download via web interface, ftp, pushing to sftp, curl). Thus, MAST users are first required to obtain files from MAST and then push them to their analysis computers. This hinders efficient acquisition of MAST data, which is a primary purpose of the archive. MAST engineers acknowledge this problem and discussed several solutions with MUG ranging from better 'curl' scripts to the use of Globus, a third party secure file transfer service. MAST engineers agree that Globus may be tractable solution going forward, but also noted that it requires money to use Globus to push data. MUG members appreciated the challenges conveyed by the MAST engineers. MUG advises that improved improved data delivery methods outside the portal be made a higher priority, particularly in light of large datasets that MAST hosts (e.g., Pan Starrs, K2, Galex).

API

As one option for dealing with high-volume data acquisition, the implementation of an application programming interface (API) to enable programmatic access to the MAST data products was discussed. It is clear that this interface is already available as the backend of the Discovery Portal and that the main ingredient that is missing is documentation. The MUG strongly encourages increasing the priority of this documentation task. We recommend that the API be primarily implemented as a well-documented set of URIs instead of a Python library. This interface is generally useful and a Python library can be subsequently developed as a community project (probably as part of the astroquery project) that uses the HTTP interface. There was discussion of the point that the API should implement rate limits (possibly with SSO integration) and stable API versioning. The MUG enthusiastically supports this

continued development and argues that this project would benefit all of the MAST services and especially the portals.

SciServer

An investigation into technology stacks that could be used to allow users to run code on a server close to the data was presented. In particular, the "SciServer" project was considered as an implementation of a cloud compute platform that could be hosted at STScI with direct access to the MAST catalogs and data products through a CASJobs interface and a local filesystem mount. This implementation spins up a virtual machine at MAST and exposes a Jupyter notebook interface that can be used to run arbitrary code in one of the supported languages (Python, etc.). The argument is made that this workflow is crucial for the analysis of the large astronomical datasets where it will be impractical to download all the data to a local machine (Pan-STARRS, WFIRST, etc.). The MUG is cautiously supportive of this investigation but encourages careful consideration of some of the key open questions. Of primary concern are the short- and long-term plans for access to compute resources. For projects with small computational requirements it might be sufficient to expose API endpoints to query the MAST data sources that can be accessed from a local Jupyter notebook. For the more computationally expensive projects, there is no clear picture of the associated costs or time allocation procedures. Furthermore, it is not obvious that the Jupyter notebook is the right development environment for these tasks. It is clear to the MUG that the workflows used for data analysis in astronomy will need to change in the upcoming years but the specific use cases and above questions should be thoroughly considered when allocating resources. We request MAST to continue briefing the MUG in future meetings about this important topic.

SciPortal

The MUG was also briefed about "SciPortal", a MAST initiative to provide new interfaces to the archives tailored for specific science areas.

The first science areas selected for special SciPortals are exoplanets and moving objects. The next examples will be deep fields and gravitational lensing. Assuming this tool is deemed useful, other ideas for specific search cases will be implemented.

For exoplanets, as an example, the user can filter on planet properties, method of detection, method of discovery, etc and be provided with graphical information on the transit light curve (from K2 for example) and table views of archive information on the star and planet. A prototype was shown, and it was noted that the graphs in this sciportal will become interactive in the near future, allowing a user "clickable" mode for real archive data selection (to load and obtain) from MAST (and other) archive holdings.

The exoplanet Sciportal has some common features to the table searches available at NExScI but has the unique ability to select and download ancillary data from archive holdings (e.g., HST images or ROSAT data).

The second sciportal example presented was for moving target searches in MAST archive holdings. The plan is to identify moving targets in all HST and other archive images, building catalogues or specific searches for a user. Future plans include a way to input possible source(s) and identify archive images useful to search for it.

Specific use cases, especially in active community research areas, will make the large general and perhaps daunting general MAST interfaces easier to approach for a user only interested in a specific science subject and source type.

The early development is going well and we look forward to seeing this made available as a public tool, thus receiving feedback and suggestions for improvement as well as other topics for furture Sciportal development.

Updates and Improvements of the General MAST Portal

The new features and improvements of the general MAST Portal were demonstrated for the MUG, including improved capability for previewing spectroscopic data, implementation of PanSTARRS imaging in the portal, and the "album view" mode that will provide an HLA-like interface after the separate HLA interface has been retired. The MUG was enthusiastic about the new capabilities. We had some discussion of how to handle more complicated spectroscopic data such as echelle data from STIS. Ideally, STIS echelle data previews should show spectra with all of the orders merged into a single spectrum. This might look a bit ugly when the full spectrum is shown, but since the portal enables users to zoom in on specific sections of the spectrum, having a single merged spectrum preview would be extremely useful for STIS researchers. Another option (that would likely be even more powerful) would be to provide links to fully coadded STIS data from the Hubble Spectroscopic Legacy Archive (HSLA) in the portal. The MUG realizes that the HSLA has not yet tackled the STIS data, but we were told that MAST intends to include STIS observations in the HSLA eventually, and when this happens, this might provide the best way to show users STIS spectroscopic data.

PanSTARRS

The presentation and demo of the PanSTARRS DR1 data release was a highlight of the MUG meeting. The DR1 release contains all of the stacked images (totaling ~100 TB) and average photometric measurements of 11 billion objects (in a 15 TB database). DR2, planned for the middle of next year, will supplement DR1 with measurements of all objects at all epochs and provide single-epoch images. With these releases, PanSTARRS catalog and image data will quickly dwarf the data volume of all other missions hosted by MAST. The demonstrated image cutout server, catalog access tool, and sky viewer all worked well and provide useful

interfaces to the data. The successful ingestion of this large data volume and the exposure of the data through public search and access tools were accomplished in a very short time, with the final version of the data made available only shortly before the public release. The ability to pull off the release of this enormous dataset is a testament to the long-term investments that the MAST team has made in continued development of robust data storage and access services, The MUG was particularly impressed that the MAST team managed to issue the release during a time when it was also working towards several other significant deadlines.

Closing remarks

The meeting closed with an executive conversation with Arfon Smith about the future. The MUG used this opportunity to share a few thoughts with the new director, including the following:

(1) The MUG has an element of advocacy in its mission, and in the MUG's opinion, MAST should urge STScI to continue to invest resources in optimal calibration of HST instruments. As discussed above, some of the instruments have calibration issues that still require effort, and as JWST ramps up, resources for these final HST calibrations might be diverted for other purposes. We understand that funding can be tight, but once the expertise on HST instruments has dispersed and instrument team members must invest their time in other projects, it will be much harder (possibly impossible) to optimize calibrations and final data products. Of course, delivering the highest quality data is ultimately the most important justification for the existence of MAST.

(2) As astrophysics continues to become an ever bigger example of Big-Data science, the issues discussed above regarding tools for efficiently accessing huge archival databases will become increasingly acute. While the MUG applauds the excellent work that MAST has done to develop the MAST Portal, we felt compelled to discuss with Arfon the fact that any browser-based approach to archival data mining will eventually become inadequate, and therefore MAST should very carefully think about the amount of effort to invest in further refinements of the portal vs. development of tools that are more suitable for research using the vast databases that will be the norm in astronomy within a few years.

(3) For future MUG meetings, the MUG recommends that it be provided with an explicit charge. The charge should include questions and issues for which MAST seeks a recommendation for the MUG, and attention should be given by the presenters to these questions and issues. For presentations which contain significant detailed transfer of information, MAST should consider placing only the most relevant information in slides to be delivered at the meeting, but include a set of backup slides to be read by the MUG before the meeting. The MUG feels that making these changes will help retain focus on the most significant issues facing the MAST user community.