Report of the Astrophysics Archives Program Review for the Astrophysics Division, Science Mission Directorate 10-12 March 2020

Review Panelists: Elisabeth Adams, Jeanne Behnke, Eric Feigelson, Richard Green (Chair), Hannah Jang-Condell, Daniel S. Katz, Kathleen Kraemer, Pamela Marcum, David Schade, Michael Wise

Introduction (based on the NASA Call for Proposals)

Reports from the National Academies (e.g., “Portals to the Universe,” “New Worlds, New Horizons”) have consistently stressed the central role and the growing importance of data archives to astronomy today. Astrophysics data centers have moved beyond simple archives that served as the final repositories of the raw data collected by a mission. These data centers have become places where the data are curated, and places from which high-level data products and data analysis tools are distributed to the science community. The data centers are now moving toward offering science platforms upon which users can integrate and analyze data to obtain new knowledge. During the last three decades, NASA's Great Observatory missions have entered a new era of legacy datasets that are proving to be of inestimable archival value. At the same time, large--scale sky surveys are becoming available across the electromagnetic spectrum.

To meet the challenge of curating the datasets drawn from NASA's astrophysics missions and making those datasets readily available for continued scientific research, NASA supports a number of complementary archives and data centers under the Astrophysics Data Curation and Archival Research (ADCAR) program. These archives include:

- The Mikulski Archive for Space Telescopes (MAST), which curates primarily UV/Optical data
- The High Energy Astrophysics Science Archive Research Center (HEASARC), which curates X-ray, gamma-ray, and legacy cosmic microwave background data
- The Infrared Science Archive (IRSA) for infrared and submillimeter data
- The NASA/IPAC Extragalactic Database (NED), which collates and cross-correlates astronomical data and information on extragalactic objects
- The NASA Exoplanet Archive (NEA), a catalog and data service that collates and cross-correlates astronomical data and information on exoplanets and their host stars

The latter three archives are all co-located as part of NASA's Infrared Processing and Analysis Center (IPAC) at the California Institute of Technology (Caltech). A bibliographic database of astronomical, astrophysical, and physics literature is maintained by the Smithsonian Institution/NASA Astrophysics Data System (ADS).

At NASA's direction, MAST, HEASARC, and IPAC took over joint management and maintenance of the Virtual Astronomical Observatory (VAO) infrastructure and participation in the International Virtual Observatory Alliance in Fiscal Year 2015 (FY15), with HEASARC as the lead institution. At that time, the restructured VAO was renamed the NASA Astronomical Virtual Observatories (NAVO). NAVO was reviewed and extended in 2017, and a portion of each archive's budget is allocated to NAVO tasks.

The Charter and Purpose of the Review

The Astrophysics Archives Programmatic Review is to be held every four or five years. This review conducts an independent evaluation of archive activities. The purpose of the review is to assist NASA in maximizing the overall scientific value of the agency's astrophysics archives and data centers. NASA will use the findings from this review to:

- *Refine its implementation strategy for the archives to achieve astrophysics strategic objectives and meet community requirements*
- *Prioritize tasks and activities for (and within) individual archive centers*
- *Give programmatic direction to the archives for FY 2021 through FY 2024*
- *Issue preliminary direction for FY 2025 (to be reviewed again in 2024)*

The findings of the Archives Programmatic Review will also be used by NASA when planning for a number of significant and important projects.

The 2020 Archives Programmatic Review includes the review of MAST, HEASARC, NEA, ADS, IRSA, and NED. Performance factors for this review include present and future potential for enabling science, technical status, data quality, stewardship, accessibility and dissemination, and future plans and expectations. The Programmatic Review also includes a review of the management and maintenance of NAVO's infrastructure.

Scientific/Technical Merit

Each archive proposed a portfolio of activities and services designed to optimally meet the needs of its user communities. The information provided in a proposal serves as the basis for evaluating the intrinsic scientific and technical merit of each archive's proposed program. Each proposal is evaluated for (a) the scientific and technical merit of the baseline portfolio, (b) the intrinsic scientific and technical merit of any proposed augmentations to the baseline portfolio, independent of the evaluation of the baseline portfolio, and (c) the scientific and technical merit of an archive's proposed portfolio augmentations within the context of that archive's baseline portfolio.

Evaluation Factors (excluding NAVO)

The factors for scientific and technical merit include consideration of the degree to which each archive's proposal:

- ***Supports the science utilization of current data holdings.*** This factor includes consideration of the degree to which the proposed portfolio of data products and services supports the needs of the scientific community, maximizes the scientific return from NASA astrophysics mission legacy datasets, and enhances the full scientific potential of the archive's data holdings. This factor also includes consideration of forward-looking aspects of the proposed portfolio (i.e., new products and activities that will facilitate new or better utilization of current data holdings and expand their scientific potential).
- ***Identifies and ingests new datasets and analysis software.*** This factor includes consideration of the degree to which the proposed portfolio will capitalize on the availability of new datasets and analysis software from both NASA and non-NASA sources. New datasets and tools may include products generated by archive staff, as well as those from sources external to the archive. For cases involving ingestion of datasets from non-NASA-funded sources, this factor will include consideration of the degree to which the non-NASA dataset enhances the scientific potential of NASA-funded data holdings.
- ***Promotes community use of archival NASA Astrophysics data.*** This factor includes consideration of the effectiveness with which the archive engages the scientific community by soliciting and responding to user input, and by providing user support and other capabilities that enhance the user experience. Feedback from user groups and usage metrics for services may be valuable indicators of performance.
- ***Takes full advantage of state-of-the-art data management techniques and processes.*** This factor includes evaluation of each archive's current data management techniques and processes compared to the state-of-the-art. This factor also includes consideration of future plans to 1) address the challenges of Big Data, and 2) to capitalize on the availability of new science platforms such as cloud computing for accomplishing each archive's mission.

Evaluation Factors for NAVO

The factors for scientific and technical merit include the science value provided by NAVO, and the benefit to NASA. Specific aspects include the extent to which the proposed NAVO implementation:

- ***Maximizes the scientific impact of NAVO.*** This factor includes assessment of the effectiveness of the proposed NAVO portfolio for enabling innovative, cross-cutting lines of investigation that would not otherwise be feasible. This factor includes consideration of the likely scientific impact and productivity of NAVO.

- ***Enhances the science return from NASA's archival mission data.*** This factor includes a consideration of the degree by which NAVO increases the scientific potential of NASA's archival data holdings over and above the potential of the individual archives.
- ***Supports the ongoing functionality of NAVO.*** This factor includes consideration of service and maintenance of Virtual Observatory (VO) software at NASA archive centers, maintenance of the VO registry, internal management of NAVO activities, and liaising with the International Virtual Observatory Alliance (IVOA). This factor also includes assessment of the effectiveness of interoperability between MAST, HEASARC, and IPAC to improve multi-wavelength access.
- ***Reflects a vision for the future of NAVO.*** This factor includes consideration of the clarity and overall scientific merit of the archives' vision for the future of NAVO.

The Final Report is provided to Dr. Hashima Hasan (Program Scientist) and Dr. Paul Hertz (Director, Astrophysics Division, Science Mission Directorate).

Review Procedure

Each of the archive centers described above was instructed by NASA to prepare proposals for continued and augmented funding for the period of FY2021-2024. Each center was given guidelines for content and budget presentation. Each resulting proposal described the center's current status including its holdings, services and tools provided, metrics on usage, scientific contributions, and the center activities' relationship to NASA strategic goals, objectives, and research focus. Proposals also presented descriptions of current projects and activities, as well as plans or possibilities for future development over the next five years. Budgets and Full-Time Equivalent (FTE) requirements were presented for both in-guide and over-guide requests. NAVO was not permitted to submit an over-guide request, and HEASARC did not submit an over-guide request.

The review was held March 10-12, 2020. To enhance the effectiveness of the review, several actions were undertaken during a preliminary Phase I. That process began with a kickoff virtual meeting on January 31, just prior to the proposal due date. At that meeting, Dr. Hashima Hasan from NASA Headquarters presented the charge, process, schedule, and review assignments to the panel. The next virtual panel meeting was scheduled for February 20. In advance of that date, the reviewers read and submitted preliminary independent reviews for their assigned proposals. Those independent reviews were merged and made available to the panel through Google Docs one day in advance of the virtual meeting. The panel discussed initial impressions of the proposals emphasizing any major weaknesses, and then formulated questions to the proposal teams for clarification. The panel requested written responses for some questions and posed other questions for verbal response by the proposal teams during the primary review meeting. The proposal teams were given four days to formulate their written responses, which were made available to the reviewers on March 2, 2020. Revised review drafts took these responses into account and were submitted two days in advance of the formal panel meeting.

Given the rapid spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) at that time, the principal review meeting was transformed to a virtual format a few days in advance of the scheduled meeting. Each of the archival centers gave virtual presentations to the Astrophysics Archives Program Review (AAPR) panel during the first two days of the meeting. Each center was represented by 3-4 people, who met with the panel for a scheduled 45-minute period. Each center gave a presentation that covered highlights of its proposal and updates since submission, and that addressed the questions and concerns that the AAPR panel had submitted prior to the principal review meeting. During this period, there was time for questions and discussion with the center personnel. A “chat” document for each proposal was maintained on Google Docs for the panel to keep notes collaboratively and discuss potential questions for the presenters. The AAPR panel wishes to thank all center staff for their diligence in preparing the proposals as submitted, their cooperation in providing detailed responses to questions, and their responsiveness during the discussion.

The AAPR panel met in a brief executive session (including NASA personnel) following each presentation in order to discuss the presentation and identify any further questions to ask the center personnel. For one proposal, the panel submitted a written follow-up question for overnight response. Following the presentations from all centers on the second day of the meeting, the AAPR panel returned to editing the reviews. Both strengths and weaknesses were identified, and over-guide budget requests were reviewed. Both the review drafts and the accompanying “chat” document were maintained on Google Docs for effective interaction among the reviewers during editing.

On the final day of the meeting, the panel members jointly reviewed each report in order to develop a version that reflected the strengths and weaknesses identified by reviewers. Secret ballots were taken for the final rating of each proposal. The panel also discussed a core issue that was applicable to all of the archive centers and provided comments in a separate letter to NASA as discussed below. Throughout the meeting, NASA officials were helpful in providing background information and guidance on the process of the review. The NASA officials were very responsive to questions from the panel.

Outcome of the Review

This programmatic review was not competitive among the proposers. The ratings are therefore for feedback and guidance to both the archives and to NASA regarding the degree to which the evaluation criteria have been met during the current period of performance, as well as the vision and utility of the plans for serving the community during the upcoming period. The panel found that ADS and MAST were at an excellent level of performance and merit. These were very strong proposals that fully responded to the AAPR Call, and the panel identified no significant major weaknesses in these archive centers' proposals or presentations. The panel found that the cluster of activities under IPAC (i.e., IRSA, NED, and NEA) rated Excellent/Very Good with strong effort enabling community science, although with some concerns to address. The panel rated the NAVO activities Very Good, with the implementation of IVOA standards identified as a critical activity to interoperability of archives, although with some unrealized potential. HEASARC was rated Very Good/Good with solid long-standing support of high energy missions and now Cosmic Microwave Background datasets, but could benefit with some updating of relatively static tools.

Collaborative Development of a Science Platform

The review panel had a major concern that spanned a number of the proposals.

MAST and the IPAC-hosted data centers were requested by NASA to submit over-guide requests to develop a science platform that would allow user code to perform server-side analysis on large, complex datasets. Additionally, HEASARC proposed development of its own science platform on the National Science Foundation (NSF)-supported SciServer through its in-guide request. Science platforms are certainly a key component of a vision for the future of astronomical data science. Each archive's proposal spoke to developing a science platform that would "interoperate with the other NASA archives and other institutions." However, the emphasis actually seemed to be on development strengths available locally, with minor focus on full interoperability and collaboration with other archives. Independent development of a science platform by each archive that is intended to support science across multi-mission and multi-archive datasets is not the optimal approach.

Having the archives collaborate on a single, unified science platform under unified funding is a more likely path to true interoperability. Such a project would require a strong leadership component. ***The fundamental science case for Science Platforms is joint analysis of distributed datasets. This goal will not be achieved by discrete platforms developed in isolation at each data center.***

A single science platform:

- Reduces duplication of development effort by each archive
- Reduces overall costs for the NASA Astrophysics program
- Enables users to see the large breadth of NASA holdings
- Enables easier access to multi-wavelength data for research
- Allows users easier access to services and tools

Upon review of the proposals, it is clear that the archives participate in community-developed standards and processes. Each proposal suggests that the teams are adopting some components of the standards as they suit the astrophysics discipline. However, building a science platform for the future requires the full adoption of common open-source policies and single standards. The best realization would be a single platform providing a common set of services for data access, user storage, processing management, and common software elements. Such a platform would focus on infrastructure and services that support users, but leave the provision of high-level analytic software in the hands of the scientists where it belongs. In other words, the platform would provide the infrastructure to share data, software, and environments (whether through Jupyter notebooks, virtual machines, containers, etc.) but allow users to manage their own environments.

A productive initial approach to developing a science platform would be to examine existing non-NASA science platforms for both best practices and the possibility of adapting substantial portions of existing infrastructures. These platforms include the NSF National Optical-Infrared Astronomy Research Laboratory (NOIRLab) Data Laboratory, the Vera Rubin Observatory (Legacy Survey of Space and Time [LSST]) Science Platform, ESASky, European Southern Observatory (ESO) Common Pipeline Library, and the NSF-supported CyVerse. These platforms' existence also highlights the ultimate need to provide inter-platform interfaces.

NAVO and VO protocols are key to interoperability. Full adoption of these protocols would require increased collaboration within NASA and with international partners. This collaboration would lead to interoperable science platforms that use a common architecture and infrastructure.

COMMENTS ON INDIVIDUAL ARCHIVES:

MAST

MAST is one of NASA's major data centers and is responsible for hosting the data from several NASA missions such as Hubble Space Telescope (HST), GALaxy Evolution eXplorer (GALEX), Kepler, and Transiting Exoplanet Survey Satellite (TESS). MAST also hosts ground-based project datasets including the first completed part of Panoramic Survey Telescope And Rapid Response System (Pan-STARRS1). MAST's core mission is data curation and providing the means for the science community to discover and access NASA data. MAST has been a principal player in the development of current science data management practices.

MAST has pushed the boundaries of its mission by creating advanced data products and services that go beyond basic data access. MAST has ingested these assets and other non-NASA mission datasets. These expansions reflect a very progressive approach that MAST brings to its mission. This approach is overwhelmingly positive in the sense that it serves the needs of the science community.

Strengths:

MAST's performance of its core functions of data curation, support for data discovery, and data access for multiple missions is excellent. MAST has excellent science and technical teams that guide the development, operations, and planning for the data center. Publication and usage metrics based on data from Hubble Space Telescope (HST) and other missions provide evidence of this excellence. The datasets and analysis tools from MAST support the production of hundreds of papers per year for both HST and non-HST data. The science cases highlighted in the proposal demonstrate that the archive is providing data that address contemporary science questions and yield high-impact results.

MAST has excellent technical staff. Ongoing operations are solid. Network performance has been improved. The development team is forward-looking and plans to expand and improve microservices, database scalability, cloud use, and Application Programming Interface (API) service effectiveness. The success of these implementations of MAST APIs and client libraries is already evident in their usage; access through MAST APIs account for over half of data downloads. To further support the usage of these APIs, MAST has adopted a style guide to standardize code format and code documentation associated with Jupyter notebooks.

MAST infrastructure has been significantly modernized in order to better serve the community, including storing/serving archives in the cloud. MAST has migrated data to the Amazon Web Services (AWS) cloud-computing environment, thus enabling large computationally-intensive applications such as machine learning that would be very difficult to execute in any other way.

MAST is excellent at development of advanced products and services that enhance the value of the archive holdings. The archive has a progressive approach to developing data discovery tools and interfaces. Users can access the data holdings through both a traditional web interface and new API services, allowing the community members to determine what is best for their science. The plans presented for interface and service development are well considered and present a feasible roadmap. Within the baseline budget is a modernization of the “MAST Classic” search engine, which will address some significant shortcomings including inability to integrate with other MAST services to allow for cross-mission queries. The planned upgrade is called Unified Search. This capability will allow form-based interfaces (a need of the community), will be fast, and will enable cross-mission searches.

As a response to community requests, MAST has proposed plans to provide tools to explore time-domain aspects of MAST holdings as well as moving targets (e.g., high proper motion stars and solar system objects). The baseline efforts include the development of new synthetic data sets (i.e., model grids of exoplanet atmospheres observed in transmission and emission, and simulated galaxy images). These data will aid in the interpretation of data from James Webb Space Telescope (JWST), HST, and the Wide Field Infrared Survey Telescope (WFIRST).

MAST has invested significantly in the creation of advanced data products like the Hubble Legacy Archive (HLA) and High-Level Science Products (HLSPs) such as the Hubble Source Catalog that create new and valuable data content for the science community. The Hubble Advanced Products initiative will carry this work into the future. Advanced data products from the community and, for example, from Astrophysics Data Analysis Program (ADAP)-supported research add further value to both the original mission and the grant program. The TESS cutout service is useful and innovative.

MAST has demonstrated a strong commitment to IVOA work and VO work in general. Efforts (often in collaboration with NAVO) include implementation of services based on VO standards as well as contributing effort at IVOA and elsewhere toward development of new and better standards. MAST has actively promoted the adoption of the Common Archive Observation Model (CAOM) by other NASA data centers, co-sponsoring two workshops.

The archive has expanded services to address the evolving needs of the scientific community. Examples include the updated new website, a new technical blog hosted by MAST Labs, and a new newsletter that was restructured to focus on concise scientific content with an electronic format suitable for viewing across several different types of devices including smartphones. Included in proposed future work are planned online live tutorials describing how to use specific MAST tools and mission data products. Unique to this plan is a focus on more senior astronomers who finished their graduate work before the era of big data, and who may benefit most from such tutorials.

Outreach efforts are well-planned and substantial. MAST has an active presence at American Astronomical Society (AAS) meetings in order to engage the community. The help desk is responsive to queries and suggestions, maintains a knowledge base, and maintains mission/database-specific documentation. The archive has demonstrated an efficient responsiveness to community needs through feedback from the MAST Users Group (MUG). The MUG has made specific recommendations such as the need to focus on the time domain and moving objects, which are proposed projects that MAST will be incorporating in the near-term to expand the science utilization of the archived data.

Given that Pan-STARRS is very heavily used and is the most cited mission after HST, the Pan-STARRS Disk Refresh over-guide is well justified. Although ground-based, Pan-STARRS is used as an important supporting resource for NASA missions. A highlighted science case required Pan-STARRS data to characterize the host galaxy of the neutron star merger detected in gravitational waves. Among other things, Pan-STARRS provides astrometric and photometric reference stars significantly fainter than Gaia. These reference stars are essential to the update and use of the Guide Star Catalog. The small-field instruments on HST do not have sufficient Gaia reference stars, and the Gaia grid will saturate for most JWST observing modes. The relatively high angular resolution of the Pan-STARRS data also provides a critical discriminant for the origin of variability in TESS, Kepler, and K2 time-series holdings.

Weaknesses:

Becoming too dependent on commercial cloud services and not using a NASA-approved vendor and contract could pose a risk in terms of future budgets, data access, and data security. Getting three years of free cloud services is actually a bit alarming in the absence of understanding the terms involved in a longer-term commitment. The strategy for cloud computing in the future is not well described.

See the general comments above about the proposed approach to science platform development. The MAST science platform is called the "Integrated Knowledge Engine" in the MAST proposal.

HEASARC

HEASARC is NASA's archive for High-Energy Astrophysics (HEA) and Cosmic Microwave Background (CMB) data, supporting the broad science goals of NASA's Physics of the Cosmos theme. HEASARC provides vital scientific infrastructure to these research communities by standardizing science and calibration data formats and analysis programs, providing open access to NASA data, and implementing powerful archive interfaces. Over the next five years, HEASARC will ingest observations from 21 operating HEA and CMB missions and experiments, while serving data from over 50 archival missions to the community. HEASARC continues to provide a reusable mission-independent framework for reducing, analyzing, archiving, and distributing data in these areas. HEASARC also provides services for all of NASA astrophysical sciences as well as leadership of the archive and VO coordination.

Strengths:

HEASARC hosts the archives of several dozen NASA missions and experiments in HEA and CMB studies. The datasets are complex, and HEASARC provides sophisticated software for access and analysis of these datasets. This core mission is performed well. HEASARC currently ingests data flows from twelve current HEA space missions (six operated by non-U.S. agencies) and seven CMB ground-based experiments. HEASARC successfully provides these dynamic archives to the United States space astronomy community. The archives are also available to the world-wide research community through use of IVOA standard protocols. Science usage indicated by data downloads and publications (around 1600 papers/year) is strong at a constant level. Two million links are available between HEASARC datasets and these datasets' use in ADS bibliographic entries.

Longstanding HEASARC tools are used for NASA missions from the Great Observatory level down to CubeSats and suborbital programs. These tools also serve European Space Agency (ESA) and Japan Aerospace Exploration Agency (JAXA) missions. Some software packages related to Flexible Image Transport System (FITS) formatted data such as the C-Flexible Image Transport System Input/Output (CFITSIO) Library, Flexible Image Transport System TOOLS (FTOOLS), FITS Tools, and fv strongly serve the wider multi-wavelength astronomical community. There are also other useful multi-wavelength tools relating to celestial sphere (e.g., Skyview, Maki) and spectra (e.g., Profit). The full integration of the Gamma-ray Coordinates Network (Transient Astronomy Network) service into HEASARC for rapid dissemination of transient events is important for the growing field of multi-messenger astronomy.

HEASARC plays a leading role in collaborating with data centers NASA-wide and world-wide in order to ensure interoperability for multi-mission and multi-wavelength science. Leadership roles include the NASA Astronomical Virtual Observatory, International Astronomical Union (IAU) CFITSIO standard, NASA's Astronomical Data Centers Coordination Committee, and Astronomy Datacenters Executive Committee (ADEC).

HEASARC establishes excellent relationships with new missions, even those missions in early stages of development. HEASARC assists these missions in organizing data products, calibration databases, and documentation.

HEASARC is currently negotiating with the German space agency to archive and serve to the community (the German-owned) half of the data from the extended ROentgen Survey with an Imaging Telescope Array (eROSITA) survey experiment on the Russian-German (RG) Spektr-RG mission that was launched in 2019. This archive will be a tremendous asset for multi-wavelength surveys as well as X-ray astronomical studies with its unprecedentedly large samples of stars, galaxies, galaxy clusters, and active galactic nuclei. A similar arrangement with the Russian agency for the other half of eROSITA data is strongly encouraged.

The expected beginning of the X-ray Imaging and Spectroscopy Mission (XRISM) archive after launch in 2022 will provide uniquely high-resolution X-ray spectra of hot plasmas using a high-technology cryogenic quantum detector. Making these spectra available through HEASARC should lead to major advances in understanding HEA processes in many astrophysical environments.

Significant achievements have been made since the 2015 Programmatic Review, including standardized Python interfaces to NAVO and stand-alone HEASARC software products. These achievements promote community use of these products beyond traditional HEA researchers.

The Center interacts effectively with its principal scientific constituencies through the HEA and Legacy Archive for Microwave Background Data Analysis (LAMBDA) Users' Groups.

Weaknesses:

HEASARC exhibits no capabilities for users to contribute diverse methods, software, and high-level science products. This limits HEASARC's ability to take advantage of the wide expertise of its research community. The Center views itself as the software creator and the community primarily as users. Mechanisms (e.g., those provided through GitHub) are needed to encourage and support the import of user contributions, transitioning (in part) to an open-source, community-based development system. This approach was recommended by the 2015 Programmatic Review, but it has not yet been implemented. In the diagram of HEASARC's role as a bridge between missions and researchers (Figure 2.1 in the HEASARC proposal), a new arrow from the researcher community to the HEASARC is needed. Community members could contribute material to newsletters and blogs, as well as informal contributions in Facebook groups (for example) to communicate improvements relating to HEASARC services.

Few plans are provided to improve functionality of the HEA and CMB data analysis capabilities. Many of the software tools and underlying methodologies have been nearly static for years. An example of recent research methods outside the scope of those current methodologies is the use of nonparametric procedures (e.g., hierarchical density-based clustering, Voronoi tessellations) to objectively define extended emission regions for HEA spectral analysis. Another example is use of various mathematical techniques (e.g., Hierarchical Equal Area isoLatitude Pixalization [HEALPix] wavelet decomposition, filtering on a frame of spherical needlets, Independent Component Analysis, Morphological Component Analysis) for discrimination of cosmic and foreground CMB emission. The notable exception is the proposed pursuit of Intensity Mapping for CMB and other cosmological fields.

No clear response is given to the 2015 review recommendation for “strategic planning in response to flat budget and loss of FTEs.”

While HEASARC performs well in promulgating its datasets to the traditional HEA and CMB research communities, there is little evident effort to broaden the scope to multi-wavelength datasets, particularly through the vast holdings of the International Virtual Observatory. Facilitating use of HEASARC's resources can operate in both directions (i.e., encouraging its traditional communities to use outside datasets, and encouraging outside communities to use HEASARC datasets).

The commitment to hosting HEASARC software on the NSF-supported SciServer effort at Johns Hopkins University (JHU) may have been persuasive a few years ago, but this hosting arrangement is not clearly a strong long-term solution. First, this commitment separates

HEASARC from similar Science Platform environments now under development by the MAST and IRSA archive centers. Close collaboration on a joint science platform by all major NASA archive centers would be highly advantageous to the user community. Second, it is not clear that JHU will continue to develop SciServer in the future or continuously provide server-side analytics for HEASARC data analysis. Third, the need for closely-integrated data and science platforms is not urgent for these relatively small datasets. Multi-wavelength integration may be more important.

Longstanding software systems for HEA science analysis have statistical methodologies that are below modern standards for data analysis. The X-ray spectral fitting package XSPEC provides only simplistic tools for a complex problem (i.e., nested high-dimensional nonlinear regression models with Poisson distributed data). XSPEC does not incorporate advances in likelihood-based fitting and model selection, high-dimensional optimization, and visualization. The statistical procedures and graphics of XSPEC, which is used in ~300 publications per year, should be reviewed by expert outsiders and raised to professional standards.

While basic data formatting and training on Jupyter notebooks for Machine Learning and Artificial Intelligence (ML/AI) is well-motivated, an intensive and enduring purely in-house ML/AI effort may not be warranted. The choice of hiring an in-house scientist for ML/AI for activities beyond an initial period was not clearly motivated in the proposal or the team's responses to questions. Many public domain ML/AI software systems are in widespread use, and the software development specific for HEA was not clearly justified. This individual scientist could usefully prepare datasets for ML/AI analysis, produce tutorials to guide inexpert HEA astronomers, and publish a few applications using HEASARC datasets. However, most science results based on ML/AI must emerge from the wider research community. HEASARC could involve the research community in new ways through this scientist. One possibility is designing stimulating Data Challenges on important problems for the science and methodology community.

NEA

NEA has the joint goal of supporting NASA exoplanet missions and supporting the broader exoplanet community. This joint goal is addressed by acting as a single repository for a wide range of exoplanet data. NEA consists of two separate but related services: the Exoplanet Archive and the Exoplanet Follow-up Observing Program (ExoFOP). These services promote scientific collaborations and support follow-up projects, thus enhancing the value of NASA mission-based datasets in this rapidly growing field.

The Exoplanet Archive contains the most comprehensive archive of exoplanet properties available, providing a critical resource for planet population-level analyses. This archive is the repository for the processed data for several NASA exoplanet missions including Kepler, K2, and TESS. The Exoplanet Archive provides a single-point access platform for information on exoplanets and their characteristics, as well as for the exoplanets' host stars.

While the Exoplanet Archive is public and open to all, ExoFOP is specifically intended for registered members to share data in order to enable science that is directly related to planets discovered by Kepler, K2, and TESS. The ExoFOP also served as a key tool for Kepler in prior incarnations.

Strengths:

As the repository for the processed data for several NASA exoplanet missions such as Kepler, K2, and TESS, NEA provides an important single-point access platform for information on exoplanets and their characteristics, as well as for the exoplanets' host stars. Several other important large-scale projects also have high-level science data hosted on the Exoplanet Archive. These projects include the Kilodegree Extremely Little Telescope (KELT) transit survey, several microlensing surveys (e.g., United Kingdom InfraRed Telescope [UKIRT], Microlensing Observations in Astrophysics [MOA]), and the Frontier Development Lab’s (FDL’s) Python Active Mediators Object System (PyAMOS) atmospheric models. The number of confirmed exoplanets is expected to increase by a factor of ten during the next decade, and NEA is making plans to deal with the continued onslaught. NEA has upgraded the set of planet hierarchies, for example, although in some cases (e.g., moons) these hierarchies are still only theoretical.

The last time NEA was reviewed in 2015, there were several other competing exoplanet databases (e.g., Open Exoplanet Database, exoplanets.org), but most of these databases appear to have ceased updating a few years ago. The only other competing database currently running is exoplanet.eu (i.e., The Exoplanet Encyclopedia), but this database does not have the same functionalities as NEA. At this point, NEA has definitely surpassed its competition in terms of completeness of planets and breadth of data offered. The direct linkages from the catalogs of all planets to data on (for example) MAST, ADS, and in particular the ExoFOP make this a highly useful portal for exoplanet science.

There are many different ways to access the data, with several different tables containing various kinds of datasets. Users can easily create plots of exoplanet properties (e.g., mass versus period, radius versus mass) to assist in demographic studies. For individual planet pages, links to papers are made available. These links are valuable because different authors sometimes come to different conclusions about planet properties.

As more data are acquired on objects, properties like ephemerides will need to be updated. Planet parameters are linked to the papers in which they are published. The database has been restructured to reflect the fact that new planets are constantly being discovered, revised, refuted, and salvaged. The inclusion of models to extract planet properties from light curves and Radial Velocity (RV) data (e.g., lightkurve, exoplanet orbital fitting program EXOFAST) is a strength. There is currently support for microlensing data, including light curves and fitting parameters. This support will be important for the WFIRST. Adding python Lightcurve Identification and Microlensing Analysis (pyLIMA) to produce model microlensing light curves is also a strength.

The datasets and analysis tools from the Exoplanet Archive contribute to a few hundred papers per year. This number is growing at a reasonable pace and is likely an underestimate since not all papers fully acknowledge data sources.

NEA ingests not just exoplanets discovered or studied by Kepler, K2, and TESS, but also contains data on all confirmed exoplanets. The Exoplanet Archive will ingest results from ongoing and upcoming missions that will produce exoplanet data, including TESS and the joint NASA-NSF RV initiative. There are at least 10 current and future missions that are planned for ingestion during the next five years. These ingests include data from both discovery and characterization missions, and also from synthetic models.

The ExoFOP is one of the greatest success stories of collaborative science in a highly competitive field. Its continual development has made exoplanet discovery in the K2/TESS era much more open than during the Kepler days when data were more closely guarded. ExoFOP provides an important venue for researchers to share data products and results, and to organize follow-up projects for future observations. This is an important service that supports community efforts to characterize exoplanets and their host stars, enabling data sharing and new collaborations that might otherwise not occur.

NEA has a good plan for improving access to its datasets, including new access protocols. NEA has implemented the use of APIs to facilitate use of the archive. The archive has developed new "tables" that summarize the data and parameters for exoplanet systems. These tables should be helpful in getting an overview of a system and accessing the underlying data. There are plans to produce Python-wrapped APIs that will enable users to better query the increasingly complex database. There are also plans for tutorials to demonstrate varied patterns of use.

NEA is investigating the use of cloud computing such as Amazon Web Services. Possible uses include generating periodograms, fitting, and running EXOFAST that may be too slow in house, especially if these capabilities are improved to allow for API access. These capabilities are currently only useful for demonstration and education on single datasets, and are not suitable for any sort of large project.

Weaknesses:

The many and varied parts of the NEA are not sufficiently well integrated. Some of these issues are discussed in the proposal, but integration is a key feature of NEA that would benefit from improvement during the next five years. Some current examples include:

- There are places where longer photometric time series are stored, but there is no link to the periodogram calculator to search for transiting signals.
- The periodogram is linked from a single transit curve where it's not useful, and the periodogram documentation does not sufficiently discuss the caveats that ought to be associated with the results.
- The transmission spectra for HD 189733 b are not linked from the planet page for HD 189733. The user has to click through to the "Planet System Overview (alpha)."

Relatedly, some effort is required in order to find some of the referenced tables (e.g., spectroscopy tables, microlensing tables).

There are many planned upgrades with new features and data. Because these upgrades will only increase the interconnectedness and possible synergies to be leveraged across data sets, it is important to maintain a strategic vision about how to present these features and data as usefully as possible for the many diverse audiences. The increased information density (which is good) on many pages can already make it visually overwhelming to locate required information. This is true even for basic queries such as "is this planet confirmed?" Some attention to User Interface (UI) usability and User Experience (UX) is needed. This problem will only get worse as more types of data are added in the future. One possibility for addressing UI and UX needs could be inviting novice users to join the User Panel and to experiment with using the web interface in order to recommend improvements.

Ingest of data relies on several FTEs of effort to regularly read the literature. Judgment calls are made on which papers are relevant and which planet parameters to adopt. These judgment calls could result in inaccuracies or inconsistencies.

ADS

ADS provides services that are utilized by virtually every astronomer on the planet. ADS plays a vital role in tracking the literature, along with citations and references. ADS has been in operation for 27 years, during which time the number of both papers and scientists has grown substantially. This growth is expected to continue. ADS services are unique to NASA and the community; there are no other similar astronomy systems to compare with them. ADS recently completed a major upgrade to its features, including a new user interface and full-text searching. Additional substantive upgrades are planned, such as improved back-end ingestion software and concept-mapping for articles.

Strengths:

ADS is an indispensable tool for astronomers to find references in the literature, whether searching by direct citations or searching to discover relevant papers. Moreover, the increasing integration with other archives and services (e.g., NED, NEA, Set of Identifications, Measurements, Bibliography for Astronomical Data [SIMBAD], Open Researcher and Contributor ID [ORCID], Digital Object Identifier [DOI]) facilitated by ADS' recently-developed API provides additional ways to serve data to various astronomical communities.

ADS is unique in science in the completeness of the scholarly data that it provides; this provides advantages to astronomy that do not exist in most other fields. ADS supports advances in scholarly communication including indexing, citations, publishing, search, and scientific analysis. Compared to other services, ADS is strikingly advanced in its features, its low cost per article, low cost per users served, and its vision for the future. ADS' filters (e.g., object names and all their synonyms), cross-references, citations, links to both the main article and arXiv versions, DOIs, and reference export (just to name a few) are unique. ADS also has advantages based on the disciplinary knowledge and expertise of its curators, unlike Google Scholar and more general (and more automated) alternatives. This expertise makes ADS more responsive to the ongoing needs and specific requests of its user community. ADS is the most relied-upon source of citations to astronomy literature for working astronomers, compared to Google Scholar or similar commercial services. ADS is therefore extremely valuable for scientific research. In contrast, usage of Google Scholar for astronomy results seems to be dominated by the general public and students.

Datasets are largely ingested by pulling in publications from the various journals tracked by ADS. Coverage is carefully maintained for core "astrophysics" topics, and less so for adjacent fields such as planetary science, geophysics, optics, and high-energy physics. What constitutes the "core" of astronomy has increased considerably in the past decade. For example, ADS broadened exoplanet-related content in response to user feedback.

Completed and planned modernization efforts offer many new options for accessing the existing data, as well as providing new connections and ways of discovering data. A major overhaul of the legacy (Classic) system was successfully completed, making ADS a much more powerful and modernized tool. The new set of search terms allows for highly sophisticated searches of the data. The ADS intends to keep up with the changing complexity of astronomical publications both by linking to other data archives and by archiving a variety of "non-traditional digital objects, each of which represents particular aspects of the research being done, such as proposals, presentations, software or data products."

The team understands the challenges that ADS will face in the future with respect to the increasing number of papers and other products, and new ways to use them more fully. The discussion of new technologies to implement in planning for the future (i.e., concept-based indexing, natural language processing, and cloud services) shows that ADS is properly exploring future trends and tasks that will be required to execute its core mission going forward. The planned concept-searching project sounds very useful for improving discovery of relevant material, as does natural language processing for searches. The team identifies important plans for collaborative development of software and notebooks, and for general work to support the external developer community. The software and overall architecture work ADS has done and plans to do is sound.

The use of ADS as a dataset in itself for research on the process of science (i.e., searching for connections between authors and concepts) including data visualization tools is a highly worthy ancillary goal.

Existing outreach is varied through presence at conferences, tutorials on YouTube, social media (i.e., Twitter, Facebook), and online support forums. The ADS Users Group meets annually and writes a summary report. This group has been providing useful guidance on features and directions for new development.

Moving to an agile workspace where no single person is key to the success of any ADS component is admirable and should be the standard for all archives, particularly in light of difficulties filling some positions and with turnover in a hot job market. The ADS also provided a description of processes for including user feedback into its sprints. This practice is an excellent way of ensuring that new development meets or exceeds user expectation.

The over-guide activities are strong. These activities include metadata enrichment and text-mining efforts. Activities also include front-end refactoring in order to improve the design of the user interface, as well as developing new functionality to support search, personalizations, and notifications.

Weaknesses:

The panel is concerned about the difficulty of finding personnel, particularly the UI designer and UX specialist. Given the tight job market and salaries being substantially lower than industry (based on a salary scale set by Smithsonian requirements), a serious effort is required to attract additional personnel. Advertising such jobs more widely and finding ways to attract interest outside of normal astronomical recruitment pipelines are likely to be productive approaches. The panel appreciates that the team has evolved in its thinking about what sort of experience is required and how to advertise for candidates with such experience.

IRSA

The IPAC's IRSA is the repository for the data from most of NASA's InfraRed (IR) and submillimeter (submm) missions, as well as several important non-NASA datasets in that wavelength regime. The missions that IRSA supports generate both data-intensive, all-sky surveys (e.g., Two Micron All Sky Survey [2MASS], Wide-field Infrared Survey Explorer [WISE], Near-Earth Object Wide-Field Infrared Survey Explorer [NEOWISE], Infrared Astronomical Satellite [IRAS], Akari, Planck) and more focused, user-directed projects (e.g., Spitzer, Herschel, Stratospheric Observatory For Infrared Astronomy [SOFIA], InfraRed Telescope Facility [IRTF]). Current active missions include SOFIA and IRTF, while upcoming missions such as Spectro-Photometer for the History of the Universe, Epoch of Reionization, and Ices Explorer (SPHEREx), Euclid, and Near-Earth Object (NEO) Surveyor will be generating data in the near future.

The core mission of IRSA is curating the data from these missions and providing the means to the science community to easily discover and access the large datasets it manages. IRSA needs to manage large archives (currently about 3 PB of imaging data and over 200 billion rows of catalogs), as well as ingest substantial new datasets from ongoing and upcoming missions. In support of these goals, IRSA has contributed to the development of current science data management practices. IRSA has created a variety of data discovery and visualization tools as well as advanced data products that go beyond basic data access. This strong initiative produces overwhelmingly positive results that serve the needs of the science community. IRSA is accomplished at carrying out its core functions. The publication rates for the missions IRSA hosts provide evidence of this excellence.

Strengths:

IRSA's performance of its core functions of data curation, data discovery, and data access is extremely strong. IRSA provides the community with vital access to NASA data and complementary non-NASA missions, as well as numerous tools to support scientific use of its data collections and services. IRSA hosts data collections from numerous past and current missions. IRSA hosts 3 PB of imaging data as well as very large databases. IRSA will be ingesting large datasets from upcoming NASA and non-NASA missions that will expand its holdings. IRSA has excellent science and technical teams to guide the development, operations, and planning for the data center.

A number of web-based tools are available for browsing and searching for data (e.g., data discovery, image viewers, catalog search tools, time series tool, etc.). The Firefly web interface recently developed at IPAC has become a useful archive interface that is now in use at all three archives hosted by IPAC (i.e., IRSA, NED, and NEA). The shared expertise between the three archives reduces redundancy and improves lessons learned. APIs are also available as alternative means for accessing the data holdings. Despite the large size of the databases, information is readily accessible for researchers. Consequently, the information is heavily used.

Publication and usage metrics are the principal indicators of success, and are both very good for IRSA. Refereed publications based on IRSA data reached 1800 in 2019. This number is likely an underestimate since not all papers acknowledge data sources, and because some papers probably acknowledge mirror sites rather than IRSA itself. More than half of NASA's ADAP grants use an IRSA dataset. Since ADAP is one of the primary avenues for using NASA data beyond the original mission(s), this is a vital resource for getting the most return on investment for NASA funds and taxpayer dollars.

IRSA regularly ingests a healthy variety of new datasets and makes these datasets available to the astronomical community. These new datasets come from both ongoing and upcoming NASA missions (e.g., the SOFIA instrument suite, the SPHEREx data cubes, the mosaic and coronagraphic image data from the WFIRST, as well as data from the European Space Agency [ESA]-led Euclid mission). These data sources also include highly relevant, complementary, ground-based and non-NASA projects such as the Zwicky Transient Facility (ZTF) or Gaia. IRSA ingests contributed datasets and software resulting from ADAP programs. IRSA will also host simulated images for use with cosmology survey missions such as Euclid, WFIRST, and LSST, although the latter survey data will be hosted on NED. There seems to be a clear path overall to increasing the value of the archive collections with new data and analysis software.

IRSA has an active presence at American Astronomical Society (AAS) meetings in order to engage the community. The help desk is responsive to queries and suggestions, fielding about 220 questions annually. IRSA maintains a knowledge base as well as mission-specific and database-specific documentation. Examples of activities at AAS meetings include workshops on how to use Python to access archives. Nearly 100 video tutorials are available online, representing a unique and impressive outreach effort among the NASA astrophysics data archives. The IRSA User Panel meets regularly (roughly every six months). A User Survey is conducted “every few years,” and the reports are available on the IRSA website.

IRSA employs standard data management techniques for its core functions, similar to those used in the other data centers. IRSA supplements these techniques with interesting, specialized tools for data discovery and visualization of cross-mission datasets. The techniques and processes seem very good overall. Cloud services are being considered for data archiving and computing services for data analysis. A new approach such as this is necessary, since IRSA is outgrowing its current relational database (Oracle) and will soon need another solution. Cloud services could support new “big data” science, which should be enabled by the increasingly large datasets becoming available.

IRSA’s first over-guide request is to support closeout activities for two missions, Spitzer and WISE/NEOWISE. Since both of these missions ran longer than anticipated as of the time of the last review, previous closeout funding was needed for operations. The requested augmentation would modernize the somewhat outdated software those datasets use, and would ensure that the data continue to be accessible as new toolkits are developed. An augmentation to fund the utility and maintainability of data access for these missions is entirely reasonable, based upon the proposal's statement that the level of in-guide allocation is fully utilized for transfer of all data and documentation from the two missions to IRSA and for continued operation of existing archive services.

Weaknesses:

The transition from Interactive Data Language (IDL) to Python necessitates some caution. While much of the community is moving toward Python, other languages may provide unique capabilities (e.g., such as the advanced statistical modules in R) that may be worth incorporating into the data analysis toolset. Other packages or languages that have not yet become trendy may also provide valuable capabilities. As an open-source environment, Python is intrinsically less stable than previous analysis tools have been. Python's flexibility and the diverse incarnations of its packages are part of its strength and appeal, but these attributes mean that user support for the new tools that are developed by IRSA for use by the community may need significantly more support than the new tools that are in older languages for the next several years.

The proposal provided insufficient detail on the data management practices and how state-of-the-art these practices are. There was also limited discussion on plans for improvements, such as database performance or basic computing infrastructure. The proposal mentioned that better data management solutions are needed (e.g., outgrowing Oracle). The proposal and IRSA team did not provide an adequate description of how this would be accomplished, even in light of the additional response to the panel query.

The second over-guide request is for development of a science platform. Panel concerns about that overall issue are discussed above.

NED

NED supports NASA astrophysics by collecting panchromatic extragalactic quantities derived from NASA astrophysics missions, space-based and ground-based photometric and redshift surveys, and the literature. NED then standardizes those measurements to a common framework. Accurate cross-identification of objects is one critical task performed by NED through its Name Resolver server, which is also extensively used by the other NASA mission archives and global astronomical services. The proposed work will preserve and enhance this service by upgrading to cloud services and enabling state-of-the-art pattern-recognition capabilities for mining published text. This upgrade will help automate the currently manually-intensive task of identifying publications with NED-relevant data. Planned machine-learning algorithms using the IRSA science platform and applied to large datasets will enable the identification of new classifications of astrophysical objects for study. The team continues to develop APIs for increased user access.

Strengths:

NED continues to provide an irreplaceable service that the extragalactic community relies upon, serving as a “one-stop shop” for connecting measurements across NASA's mission archives and for providing standardized panchromatic galaxy parameters. This service enhances, for example, community utilization of NASA observatory facilities and the construction of meaningful samples for conducting specific studies. Without NED, the community would be spending substantially more time consolidating such information and would likely be doing so without the same level of “completeness.” Such a lack of NED's service would probably result in substantial loss of science. The proposed activities will continue providing this valuable service to the community.

NED supports NASA astrophysics by providing a Name Resolver server that is heavily used by the other NASA mission archives and global astronomical services.

NED proposes a probabilistic-based alternative to single-best-match reporting of objects in multi-survey cross-matching, which informs the user of the reliability of the match. This information increases the robustness of any sample-selection process or comparisons using NED-based data. Cross-matching methods are challenged by the proliferation of multi-instrument and multi-band surveys having different sensitivities and resolutions, particularly with the new generation of sensitive telescopes where galaxy blending is common. Various statistical approaches to this blending problem have been suggested without a clear best choice. NED staff are aware of the issue, and propose to provide matching probabilities to multiple counterparts where appropriate. This more informative approach is an essential improvement to a single-best-match procedure.

An inherent characteristic of NED is that it continuously extracts datasets from the literature and datasets as they are published, synthesizing and ingesting them into the archive. In addition to maintaining this capability, NED proposes to include future data from JWST, Euclid, and SPHEREx. These inclusions will continue the relevancy of the NED archive to current research.

While the technology is still in its infancy and has not been tested on the whole archive, NED's consideration of machine-learning techniques to automate the identification of publications with NED-relevant data could be a critical factor in NED's ability to ingest increasing amounts of data as the literature grows. These techniques could also serve as an effective mitigation for the risks inherent to continued reliance upon manually performing this task.

An Ambassador program was established to help foster community engagement for major research collaborations and for training of students. The program encourages the following of best practices when publishing data, such that those data are readily ingested by NED. This program was a recommendation of the NED Users Group.

The Level 5 Knowledgebase materials are substantially utilized as astronomy course materials at universities, in outreach programs such as the NASA/IPAC Teacher Archive Research Program, and by citizen science programs such as Galaxy Zoo. This use expands awareness of NASA missions and supports science outreach.

The archive management has demonstrated forward-looking planning in order to keep the archive relevant as the era of Big Data rapidly evolves. Although in their infancy, “pilot studies” initiated by the NED team to test viability and scalability include cloud-based computing environments and a probabilistic algorithm (i.e., Match Expert, or MatchEx). This algorithm substantially increases cross-matching and data synthesis by using a local background source density in parameter space to estimate the likelihood for a match, minimize the false-positive match rate, and then estimate match completeness. These approaches have made possible the ingestion of extremely large datasets such as the Spitzer Source List (42 million entries), Two Micron All Sky Survey (2MASS) Point Source Catalog (471 million entries), and the ALL Wide-field Infrared Survey Explorer (ALLWISE) Source Catalog (747 million entries). Such ingestion is a feat that would be much more challenging without inclusion of these new computational methodologies.

The proposed over-guide includes the ingestion of some large space-based and ground-based datasets and surveys covering a wide wavelength range (e.g., Sloan Digital Sky Survey [SDSS] IV, Dark Energy Survey, Evolutionary Map of the Universe, Gaia, Dark Energy Spectroscopic Instrument, the Legacy Survey of Space and Time). These added datasets would significantly enhance the NED science provisions. The new additions would be synergistic with existing NED databases in several cases, providing new parameters that were not previously available for the community.

Weaknesses:

The time-intensive, manual process currently used to collect data from the community and ensure format compatibility for ingestion into the archive is unsustainable. The current process sometimes involves one-on-one interaction with individual authors in order to ensure that they provide their data tables in a compliant machine-readable format. This inefficient approach will become untenable as the literature grows. Close consultation with ADS (and perhaps Centre de Données astronomiques de Strasbourg [CDS]) staff on automated procedures for identification and ingestion of the extragalactic literature is a promising alternative approach. The current extremely high standards of completeness would need to be reduced to a more feasible level for greatly expanded information flow in a constraint of roughly constant resources.

The proposed over-guide includes an expansion of the number of journals routinely data-mined. This expansion within NED appears to represent duplicative effort; while machine learning (or even searching for extragalactic keywords) could assist NED with ingestion of data sets within those papers, ADS is developing potentially relevant machine-learning techniques.

The proposal to extend NED capabilities to serve time-domain astronomy carries the risk of being infeasible to the extent that it is intended to provide standardized or derived data products from federated datasets. This is of concern given the various factors that can complicate the interpretation of a light curve "harvested" from a paper. These factors include incompatible spectral bands, different durations versus instantaneous observations, complicated error structures (e.g., systemic, Gaussian, Poissonian), spurious or instrumental outliers, reliability flags, upper limits with different definitions, incorrect matching due to crowding, etc.

The over-guide requested would increase the staff over the current ~12 FTEs by 50% (adding 5 to 6 FTEs) in Fiscal Years 2022-2023. The NED proposal did not describe sufficiently detailed plans for the out-years that justify increasing the staff by this amount.

Part of the over-guide request is for development of a science platform. Panel concerns about that overall issue are discussed above.

NAVO

The NAVO project is a joint project of the major NASA astrophysics archives HEASARC at Goddard Space Flight Center (GSFC), MAST at the Space Telescope Science Institute (STScI), and IRSA and NED at IPAC. The NAVO project works with the IVOA to develop the interfaces necessary to support machine-based access to NASA's rich collection of archived data, then works with the NASA archives to deliver these interfaces and tools that implement them. These VO interfaces are compliant with the IVOA data access protocols and have become the most common way that the global astronomy community accesses data in the NASA archives. As such, the work of NAVO represents crucial infrastructure for the successful scientific exploitation of NASA's archival data.

NAVO plays a leading role within the IVOA in developing these standards. These standards underpin the infrastructure of astronomy world-wide, and are especially crucial for data-intensive projects such as Gaia, the Vera C. Rubin Observatory, and Euclid. NAVO supports these services by performing the following:

- Maintaining the query infrastructure developed by NAVO's predecessor, the Virtual Astronomical Observatory (VAO)
- Measuring and optimizing the performance of database services
- Developing standardized Python interfaces for community use
- Executing a program of professional outreach

Ensuring the discoverability, accessibility, and interoperability of NASA archival data is the driving principle behind the development of the VO tools and interfaces that the NAVO project maintains.

This underlying VO infrastructure and the role of the NAVO project will become increasingly important as the scientific use of archival data becomes more complex, more distributed, and more multi-wavelength. The NAVO project is developing new tools and capabilities to adapt to the changing data analysis needs of the community. NAVO will provide services with performance that will scale as datasets grow, as platforms evolve, and as usage patterns become more complex. These services include:

- Supporting and extending the implementation of a standard set of access protocols to NASA archival data
- Enabling science-based data discovery by establishing uniform metadata discovery of science holdings
- Understanding the need for new standards for efficient interoperation of major new science platforms where users can bring their analysis tools to the data
- Developing standardized Python libraries to access NASA data
- Providing outreach through workshops and its web site
- Continuing to lead within the IVOA to develop appropriate standards

The NAVO project can ensure that the needs of the NASA science community are reflected in these new platforms and ensure that the community benefits from the substantial opportunities and efficiencies the VO brings.

Strengths:

NAVO's development and support of VO standards clearly is essential in allowing the archives to work together in delivering common APIs for data access. NAVO has successfully made great progress in implementing VO standards across the NASA data centers. NAVO's partnership with the archives is a major strength, as is the specific collaborative work done with the archives to prepare for new datasets.

NAVO has given the NASA data centers a leadership role in the IVOA, and thereby ensures IVOA developments remain responsive to the needs of the NASA science community. The implementation of IVOA standards is very positive for data accessibility, and this implementation contributes to global interoperability of NASA data. NAVO work with IVOA and leadership in IVOA has been generally successful. This effort includes the very strong presence that NASA data centers have at IVOA. This presence includes 10 people as Chairs or Vice-Chairs and five people on the Committee for Science Priorities. If those people could be funded at a higher level in order to make yet stronger contributions to IVOA, the ultimate benefit would be interoperability among NASA-supported and other key international archives.

The development of standard Python interfaces is an excellent way to deliver core NAVO functionality to a wide segment of potential users of NASA archive data. These libraries provide an effective means to demonstrate and deliver VO functionality to a wide cross-section of the community. The libraries enable software development that can ultimately be independent of the specific host of the data to be analyzed.

NAVO's implementation of standards enabling improved access to NASA archival data should directly enhance the rate and amount of science return from that data. Adding uniform services for new missions as they become available is an important core function of the NAVO project that directly increases the science returned from NASA archives.

The registry service, related metadata, and the CAOM and Unified Content Descriptors (UCD) work are strongly needed in order to support the ongoing functionality of NAVO. The distribution of effort across the core VO operational functions (e.g., registry, mission interfaces, metadata frameworks) is both essential and appropriately prioritized. NAVO operates a United States-based full VO registry service housed at MAST, which has been successfully operating and providing data to all NASA archives. This service appears to have a solid infrastructure framework that is supported by MAST. NAVO has several plans for improving data systems for the next five years. The team mentioned a lot of cross-archive metadata cleanup work for the next few years, which is an important effort not to be undervalued.

The outreach activities that NAVO has performed, and those that are planned, are a good means of interacting with large parts of the current and future VO user community. These activities represent an excellent investment to increase the use of NAVO developments.

The review panel felt that the NAVO project is doing a good job overall in delivering crucial infrastructure for the NASA archives and maintaining a leadership role for the US in the VO community globally, all with fairly modest resources.

Weaknesses:

Despite the good progress being made by the NAVO project, the review panel notes several concerns. First and foremost, the panel finds that the overall mandate for NAVO to lead in the definition and implementation of VO-related interfaces and tools is not being realized to the extent necessary across the various archives. While collaboration and communication amongst the archives is clearly visible, the strength of NAVO to make real decisions about design and implementation that are then deployed across the archive centers is not clear. The various archive centers continue to develop their own interfaces and tools. Although these interfaces and tools are no doubt developed for well-motivated reasons, these efforts appear to duplicate NAVO efforts in some cases. More generally, NAVO appears to have a passive approach to working with the archives as opposed to taking a coordinating role. In the current approach, NAVO waits for the individual archives to define the activities for which NAVO support is required. This approach invites the possibility of duplication and inefficiency. As a cross-cutting activity by definition and given NAVO's perspective on developments in the international community, a more effective outcome would be achieved if NAVO were to develop a future vision for the necessary tools and interfaces and then work with the various archives to implement that vision.

The development of Python-based VO tooling is one obvious area that could benefit from tighter coordination. Although this work is (or could potentially be) highly relevant for various software development efforts across the NASA archives, it was not obvious that these efforts are well-coordinated. With the current mode of collaboration, the incorporation of VO functionality into new software elements developed at the various NASA archive centers seems to be uneven. Similarly, the existing IVOA standards are not implemented uniformly across the NASA data centers (e.g., DataLink is operational only at HEASARC). This non-uniformity seems to reflect a lack of coordination in addition to divergent priorities for implementation.

The review panel is concerned that the level of available resources could be adversely impacting the NAVO project in several areas. As a general statement, the NAVO effort consists of 7 FTEs spread over four participants (i.e., NED, IRSA, HEASARC, MAST), which are then divided into 18 Work Breakdown Structure (WBS) elements. This division therefore leaves only fractional people to support a variety of major efforts among the NASA archives. This distributed development model can be difficult to manage. Without a strong set of processes, this model leads to limited development in the NAVO systems and also makes the creation and execution of a coordinated strategy more difficult.

The NAVO project has defined the following core activities:

- Maintaining existing services
- Adding new missions and data collections
- Developing new software tools and interfaces
- Participating in the national and international VO collaboration
- Executing a public outreach and training effort

Although the review panel feels that the NAVO project is making progress in all of these areas, the level of available resources seems fairly minimal for these activities and allows only limited efforts for some of them. Additional resources would strengthen these activities and increase their impacts on the archives. Likewise, a stronger leadership role for NAVO in interface and tool development across the NASA archives would clearly require additional resources. The review panel believes such leadership would benefit the archives and their users.

This issue of limited resources would be further exacerbated if the NAVO project were to assume any substantive role in the design and development of a NASA Science Platform. As stated in the proposal, NAVO currently sees its role in the Science Platform discussion as limited to assessing how existing standards must be adjusted and which new standards must be developed in order to support the disparate efforts at the various archive centers. However, even the stated ambition to work with the centers to develop a set of NASA-wide requirements for remote machine access to science platforms will be difficult to achieve with the current level of resources. Further, the review panel feels that a number of options for the development of a unified NASA Science Platform effort will require an increased role for NAVO. Such a role will represent extra work for which additional resources will be necessary.

Finally, a clear agenda for NAVO participation in the IVOA efforts was not made obvious. It was not clear what priority areas NAVO would like to see IVOA adopt or develop. It was not clear when NAVO leads and when it follows. IVOA is very diverse and has a democratic approach to the development of standards, which makes it slow to develop standards. This situation leads to challenges in the scientific direction of IVOA's activities. The greatest benefit to science from NASA data would arise from NAVO having a clear plan to vigorously encourage the continued evolution of IVOA toward being a more efficient organization that is focused on the most important scientific and technical issues of the time.