Report of the Astrophysics Archives Review for the Astrophysics Division, Science Mission Directorate 12-14 March 2024

Review Panelists: Bruno Altieri (ESA), Michael Blanton (NYU), Eric Burns (LSU), Justin Finke (NRL), Peter Gao (Carnegie), Richard Green (U Arizona, Chair), Mario Juric (U Washington), Dana Ostrenga (NASA/GSFC), Luca Rizzi (NSF/AST), Yue Shen (U Illinois)
Executive Secretary: Rebecca Levy (U Arizona)

Introduction (from the NASA Call for Proposals)

Reports from the National Academies (e.g. the 2007 "Portals to the Universe", and the Astro2020 Decadal Survey "Pathways to Discovery") 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 to become data centers where the data are curated and high-level data products and data analysis tools are distributed to the science community. NASA's Great Observatory missions have entered a new era of datasets that are proving of inestimable archival value. At the same time, large-scale sky surveys are becoming available across the electromagnetic spectrum.

- NASA Astrophysics supports a number of complementary archives and data centers to meet the challenge of curating the datasets from NASA's astrophysics missions and making those datasets readily available for continued scientific research, according to the FAIR principles (Findability, Accessibility, Interoperability, and Reuse) for scientific 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), primarily for infrared and submillimeter data;
- the Mikulski Archive for Space Telescopes (MAST) which curates primarily UV/Optical data;
- the NASA Extragalactic Database (NED), that collates and cross-correlates published astronomical data and information on extragalactic objects; and
- 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.
- A literature-centric database covering astronomy, astrophysics and additional Science Mission Directorate (SMD) disciplines maintained by NASA's Astrophysics Data System (ADS). ADS indexes bibliographic content, supplemented by other scholarly resources such as published catalogs, cited software and some data products, and also hosts full-text articles from the historical astronomical literature. As a part of its Open Source Science initiative, in 2018 NASA SMD conducted a Data Workshop: proceedings are available as "NASA SMD Maximizing the Scientific Return of NASA Data". Following that event, in 2020 SMD decided that the bibliographic database, ADS, should be expanded to include other disciplines within the Science Mission Directorate.
- The NASA Astronomical Virtual Observatories (NAVO), jointly managed/operated by HEASARC, MAST, and by IRSA and NED at IPAC, coordinates the efforts of NASA

astronomy archives in providing comprehensive and consistent access to NASA's astronomical data through interfaces that follow standards set by the International Virtual Observatory Alliance.
- A major finding of the last review conducted in 2020, was that NASA should build a single science platform that would enable a joint analysis of data sets across the Astrophysics archives. NASA has started the development of such a platform.

The Charter and Purpose of the Review

The Astrophysics Archives Review (AAR) provides an independent evaluation of archive activities to assist NASA in maximizing the overall scientific value of the agency's Astrophysics archives and data centers. NASA will use the Review findings 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 2025 through FY 2029;
- Issue preliminary direction for FY 2030 (to be reviewed again in 2029).

The Astrophysics Archives Review will include a review of the management and maintenance of the infrastructure of NAVO and of plans for the Astrophysics Science Platform. HEASARC, MAST, and IPAC are directed to include within their individual proposals those NAVO activities which are specific to their Archives. NAVO activities cutting across all the Archives, interaction with the user community, process adopted for prioritizing cross cutting activities, interaction and coordination with other VO efforts through the International Virtual Observatory Alliance (IVOA), overall management of the NAVO infrastructure, and a vision for the future for NAVO should be included in a separate NAVO proposal. Activities across all the Archives for the development of the Science Platform should be included in a separate Science Platform proposal.

Evaluation Factors (excluding ADS, NAVO, and Science Platform)

The factors for scientific/technical merit include consideration of the degree to which each archive's proposed work over the period FY 2025 – FY 2030:

1.) Supports the science utilization of the archive's data holdings.
2.) Identifies and ingests new datasets and analysis software as appropriate.
3.) Promotes community use of archival NASA Astrophysics data.
4.) Takes advantage of state-of-the-art data management techniques and processes.

Evaluation Factors for ADS and its extension to SMD science data

This proposal should include specifics of the ongoing and proposed expansion of coverage to SMD science areas beyond astrophysics, and plans for engagement with those science communities. The factors for scientific/technical merit include consideration of the degree to which, over the period FY 2025 – FY 2030, ADS and its proposed extension:

1.) Supports the science return from NASA SMD missions and research activities.

2.) Identifies and ingests appropriate new publications and datasets to better serve SMD's scientific community.
3.) Promotes community use of NASA scientific information.
4.) Takes advantage of state-of-the-art data management techniques and processes.

Evaluation Factors for NAVO

The factors for scientific/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 in the period FY 2025 – FY 2030:

1.) Maximizes the scientific impact of NAVO.
2.) Enhances the science return from NASA's archival mission data.
3.) Supports the ongoing functionality of NAVO.
4.) Reflects a vision for the future of NAVO.

Evaluation Factors for Science Platform

The factors for scientific/technical merit include the science value provided by a Science Platform, and the benefit to NASA. Specific aspects include the extent to which the proposed Scientific Platform implementation over the period FY 2025 – FY 2030:

1.) Defines an effective structure for management and organization.
2.) Enables users to search for desired data in the NASA Astrophysics archives, whether they are held in the cloud or on-premises at one of the three mission data archives, and to access these data for computations on the platform.
3.) Enables users to make data accessible to a process on the platform when those data are not already proximate in the cloud.
4.) Allows users to perform server-side analysis on datasets held by the NASA Astrophysics archives or in other widely-used data archives held in the cloud proximate to the platform, or on datasets belonging to the users themselves.
5.) Provides access to pre-installed astrophysics software packages based on open source science-driven code, together with examples for analysis of datasets held by the Astrophysics Archives, as notebooks and containers that users can modify and edit for use on the platform or elsewhere.
6.) Allows multiple users to collaborate on the platform by sharing data products and notebooks to view, edit, and run code.
7.) Operates in a manner consistent with the Astrophysics mission data archives.
8.) Takes full advantage of state-of-the-art data management techniques and processes.

The Final Report is provided to Dr. Hashima Hasan, Program Scientist, Dr. Linda Sparke, Program Scientist, Dr Eric Smith, Associate Director for Research, Astrophysics Division and Dr. Mark Clampin, Director, Astrophysics Division, SMD.

Review Procedure

Each of the archive centers and projects above was instructed by NASA to prepare proposals for continued funding for the period FY 2025-2030 and given guidelines for content and budget presentation. Each proposal described the centers' current status, including its holdings, services and tools provided, metrics on usage, scientific contributions, and relation 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 5 years. Budgets and FTE requirements were presented for both in-guide and over-guide budget requests, particularly given the loss in purchasing power with the budget envelope provided.

The review was held March 12-14, 2024. 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 February 13. At that meeting, Dr. Hashima Hasan from NASA Headquarters presented the charge, process, and schedule to the panel, along with their review assignments. The next more extensive virtual panel meeting was scheduled for February 28. In advance of that date, the reviewers read and submitted preliminary independent reviews for their assigned proposals. Each proposal was reviewed by one primary and two secondary reviewers. The independent review reports were merged and made available to the panel through Google Docs in advance of the virtual meeting. The panel discussed initial impressions of the proposals, particularly major weaknesses, then formulated questions to the proposal teams for clarification. The panel requested written responses for some questions and posed others for response during the primary review meeting. The proposal teams were given six days to formulate their written responses, which were made available to the reviewers on March 11. Revised review drafts taking the responses into account were prepared for the formal panel meeting.

The panelists met in person with the NASA program officers, Hashima Hasan and Linda Sparke. The archival centers met virtually with the panel. Because of the thorough proposals and answers to the first round of questions, the interaction with the proposal teams was through discussion rather than presentation format. Each center was represented by multiple people, who met for a 45-minute period scheduled with the panel. The AAR 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.

Following each presentation, the AAR panel met in a brief executive session (including NASA personnel), to discuss the presentation and identify any further questions it wished to ask of the center personnel. Following the presentations from all centers, on the third day of the meeting, the AAR panel returned to editing the reviews, identifying both strengths and weaknesses, and reviewing over-guide budget requests. Both the review drafts and the Executive Secretary's extensive notes were maintained on Google Docs for effective interaction among the reviewers.

On the final day of the meeting, the panel members jointly discussed each report, and secret ballots were taken for the final rating of each proposal. Before and throughout the meeting, NASA officials were helpful in providing background information and guidance on the process of the review and were very responsive to questions from the panel.

Outcome of the Review

This programmatic review was not a competition among the proposers. The ratings are therefore for feedback and guidance to both the archives and to NASA on the degree to which the evaluation criteria have been met during the current period of performance and the vision and utility of the plans for serving the community during the upcoming period. The panel found the archives and services to be in a mature and stable state, providing critical services to the astronomy community. They rated MAST, IRSA, NEA, ADS and NAVO at Excellent / Very Good level of performance and merit. These were strong proposals that fully responded to the AAR Call, and contained several major strengths and few major weaknesses. The panel found that the longer-serving archives, HEASARC and NED, rated Very Good, with strong effort enabling community science, although with some concerns to address about modernization and scalability, as elaborated below.

In addition to the established programs, the panel was very pleased to see NASA's positive response to the recommendation of the previous AAR review in establishing a joint effort for development of a Science Platform. The discussion with the team leads revealed a positive collaboration among the three flight mission data archives, HEASARC, IRSA and MAST, in approaching the development. The rating of Very Good for this project just getting underway reflects the cooperative and promising start, while noting some concerns about scope and support of a wide user base.

All the programs noted the impact of the loss of purchasing power from the budget envelope provided for the period of performance.

Collaborative Development of a Science Platform

NASA's response to the AAR review recommendation in 2020 has been to support the development of the Fornax Initiative, a cloud-based system that brings together data, software and computing into a cohesive system focused on providing scientists with modern tools to interact with and analyze data. The proposal is to extend the deployment of the prototype first to a small set of vetted users, and then to a larger set of unvetted users. Additionally, the proposal will support the development of new functionalities and focus on user driven science workflows. Besides the cloud deployment, the platform will be installable by end-users to run on their local systems. The proposed architecture consists of three main components: the science components, made of the astrophysics elements necessary to enable science in the cloud and including tools such as Python notebooks; the science console, a web-based application that handles the users login to access the cloud computing, data storage and data analysis; and the science support system, a program of engagement with the astronomical community that also includes a help desk.

Strengths

The management plan is coherent and reflects a thoughtful division of effort. The project has already delivered a working prototype using a distributed management structure and there is no reason to doubt that they could continue to use this structure effectively for the future. The individual archives have in-depth knowledge of their specific data holdings, the scientific

workflows, and their user communities: based on this, they are ideally positioned to develop the scientific components. Some of the core infrastructure is assigned to a central group of developers, with the right expertise and mandate to carry out the task. The leads within the program management office and at the archives are excellent.

Fornax uses a modular model with multiple points of entry and relies on current standard tools for science platforms/gateways of this sort. As a result, the project will significantly improve the availability of modern, effective tools to enable users to search NASA Astrophysics archives, both for data hosted on the cloud and on premises. Cloud data will be accessed within Jupyter notebooks using standard tools such as PyVO and Astroquery. Data on premises can be imported into the cloud. The alignment of the work performed by NAVO, the archives and by Fornax is an effective way to make sure that proper standards, libraries, APIs (Application Programming Interfaces), and tools are developed to enable users to search for desired data in all the NASA archives, both on the cloud and on premises. The collaboration and sharing functionalities, the inclusion of visualization tools, computing packages, specialized astronomical software, and partitioning capabilities will make this a powerful platform.

The plan to provide access to pre-installed astrophysics software packages is well articulated and complete. It includes the idea of providing examples of analysis of datasets, notebooks and containers, and the ability to tune the environment as well. The current and future focus of the example notebooks is on cross-mission science, which will be greatly beneficial to users who approach a science platform for the first time. This is among the strongest elements of this proposal and one that highlights the importance of the collaboration between archives to produce solutions that are carefully tailored to the needs of the diverse communities of scientists served by the different NASA centers.

The proposal presents compelling plans to adopt a fully Open Deployment model, which includes the principles of Open Science and Infrastructure as Code. This is a strong foundation to make sure that the deliverables continue to take advantage of state of the art data management techniques. The API, storage, and compute system uses state of the art, and well-proven, techniques, architectures, and specific tools.

The idea proposed in the overguide is to develop a joint-archive, multi-wavelength NASA-Rubin catalog. This is a valuable undertaking which is supported by the community (see Astro2020). The proposal makes a strong case for why this overguide should be granted and it would be a huge achievement, saving numerous users immense amounts of time. It will require a strong scientific lead to guide design decisions, priorities, and requirements for success.

Weaknesses

The development of the prototype system is evidence of an effective team. Moving forward, this project has the potential of becoming much more complex, growing in scale, scope, cost, and ambition. The panel believes that the current management structure will soon be inadequate and will require a clearer definition of management procedures. Specifically: the Project Scientist does not appear to have sufficient direct reins to supervise the distributed FTEs; the process to select technical leads or systems engineers at individual archives is not clear and does not specify how candidates are identified, who can decide to assign them specific responsibilities, or how the

archives coordinate on these. The 2 FTEs assigned to the Support Service effort seems insufficient given the scope of Fornax, particularly if that effort is split among 4-5 individuals.

While searching for NASA data both on the cloud and on premises is well described and realistic, a significant limitation will be imposed by the restrictions on upload/download rates. The solution suggested to overcome these limitations is a local installation. While this is certainly feasible and supported by the panel as an option, this is not likely to be a realistic solution for any but a small number of well-resourced institutions with good IT support.

There is an expressed concern about the lack of described data governance with the expectation of having on-premise systems with parallel cloud infrastructure. This is a critical element to ensure that the data is secure, available, and usable. There is an identified risk with data stewardship and potential ballooning resource requirements if the governance is not well outlined.

The proposal does not describe using APIs to access other sources of astronomical data outside the NASA archives, beyond what is possible through PyVO, astroquery, or other tools that could be installed in a container. Likely expensive non-proximate queries would be throttled, which would limit access to Rubin scale data. The proposal states that "understanding how to successfully operate the Fornax system within a fixed budget will be a critical area of work". All these resources will be subject to quotas. This could result in specific large data sets being excluded. An example could be the Rubin dataset. For this specific case, the proposal suggests an overguide aimed at providing access to Rubin catalogs. **In summary, the proposal describes a comprehensive plan to provide access to private storage and downloads, but these features might be significantly limited by their cost.** While the activities described in this proposal are a sensible and required first step, the proposal does not suggest strategies to address possible future scenarios that include a ballooning demand for access to the planned services.

There is a growing world of large astronomical data sets outside of NASA, and any development of science platforms should set up the right hooks that other agencies and data generators can connect to. The current proposal does not describe methods to interact with data outside of NASA other than the possibility of uploading the data to the available cloud storage. This initiative would be ideally positioned to plan for a future where multiple holdings are stored in common formats (or accessed via common abstraction layer) and multiple platforms have access to them. Eventually, this could also lead to a level of convergence on emerging technology and solutions.

The proposal lacks clarity regarding whether the available resources are sufficient to accommodate the storage and computational requirements for a typical user load, and there is no precise estimation of what constitutes this nominal load. The proposed solution for scaling capacity appears to involve transferring the system for deployment onto user-provided solutions, whether on cloud platforms or on-premises. This approach seems to stem primarily from concerns about the cost model and uncertainties surrounding the future of the NASA cloud infrastructure.

As mentioned in other items above, this team is uniquely positioned to provide NASA with a long term strategy for the deployment of the Fornax platform as the main access point for

astrophysical data, but this is not currently in the proposal. Such a strategy should contain different scenarios and estimates of the costs, so that NASA could start the necessary planning process.

There is some concern that the archives in general see Fornax as just another service, and not a higher-level strategic goal and as the future of each archive.

COMMENTS ON INDIVIDUAL ARCHIVES:

MAST

MAST will maintain, operate, and enhance its several-petabyte holdings of several major space-based NASA missions (e.g., HST, JWST, TESS) and a number of additional UV through IR missions such as GALEX as well as ground-based surveys such as Pan-STARRS. MAST is the default platform to archive these data and provide services to the community for accessing data products at different levels (from raw data to High Level Science Products (HLSPs)), performing science analyses and conducting educational outreach with various portals and web tools. Over the years, MAST has provided indispensable services to the astronomical community to work on these data, with an impressive record of impact and long-term stability. The proposal objectives are to improve the infrastructure and capabilities of providing continued services on exist data sets while expanding to accommodate new mission data (Roman, the extended TESS mission, new HLSPs and SDSS data, etc.), to develop new technologies such as scalable databases and cloud computing to meet the challenges of increasingly more diverse data sets and science needs, to ingest and chaperon the growing set of targeted data products and HLSPs, and finally, to develop the tools (such as enhanced search interfaces with large language models) and provide tutorials and workshops to further facilitate the search and science analysis functions of MAST and to lower the barriers to astronomical research for all.

Strengths:

MAST is currently holding a large number of past and current NASA missions (25 in total; over 4PB data) and will continue to expand its holdings with upcoming missions (e.g., Roman). MAST has been run consistently over the past few decades and seen its impact (in terms of publications and citations) grow considerably in recent years. This trend is predicted to grow further in light of JWST, TESS and other upcoming high-profile NASA missions.

MAST holds a diverse range of data sets covering UV through infrared. While most of the holdings are from space telescopes, MAST has the capability and initiative to ingest data from large ground-based surveys (e.g., Pan-STARRS) to enrich the science emerging from NASA facilities. With this array of data sets, MAST provides the science community a wide range of research opportunities, with a uniform interface and flexible APIs to query and retrieve these data. Hosting community-contributed high level science products and non-NASA data is an important service that greatly multiplies MAST's impact.

MAST follows guidelines for open and FAIR (Findable, Accessible, Interoperable, and Reusable) data, and develops a reliable infrastructure to enable long-term stability of services, deploys new technologies and interfaces to improve the user experience, and interacts with other astronomical archives and the science community to enhance the overall value of archived data.

Looking forward, MAST is improving archive infrastructure (e.g., with more cloud services) and developing more tools to improve the overall user experience, to better visualize current data holdings, to develop more science platforms for analyses beyond data retrieval, and to implement new technologies (e.g., AI) in query portals.

MAST is the standard archive for the expanded TESS mission (through 2030), the Roman Space Telescope, and JWST. These three NASA missions will rely on MAST to deliver their data products to the community. MAST has the infrastructure, and is currently developing more tools to fully exploit data from these high-profile missions.

Beyond these major NASA missions, MAST will ingest new data from ground-based programs (e.g., the SDSS legacy data), as well as community-contributed HLSPs. The standardized cross-mission, metadata format used by all MAST holdings is applied to these data ensuring that all of these products can be found easily and accessed according to FAIR principles. Having such products available and served next to MAST products will have a synergistic effect, and further reduce the barrier to entry for cross-dataset (or cross-data product) analyses by the MAST community.

The proposal does a good job of showing the enabling power of MAST for science publications and in its data download and query statistics. Generally MAST seems to be engaging in the right practices with a variety of means for feedback and educational opportunities: the MAST Users Group; the help desk; AAS presentations; Jupyter notebook tutorials; workshops; accessibility initiatives.

MAST utilizes state-of-the-art data management techniques and processes, and will continue to implement new technologies for future development. For example, MAST is advancing into cloud services in a deliberate way and seems to have some clear wins to show for it (e.g. the cutout services). Other examples include modern software development cycle (including source code management, unit testing, continuous integration and deployment), API-first development approach, large-scale databases such as SQLServer and (presently in evaluation stages) Greenplum.

The overguide budget proposes to close out the old technology stack in the Discovery portal, to provide unified data search and catalog services for long-term planning. This would require shifting all remaining services to the new search forms, and could reduce the long-term maintenance cost and also the opportunity cost of the community being slowed down by the outdated interface. The panel considers this a convincing argument.

Minor Weaknesses:

The proposed initiatives for the next six years all appear valuable, but it’s been difficult to

understand how they were selected. Future proposals could benefit from more detailed explanation behind the prioritization process and how it responds to MAST's long-term vision and strategy.

The proposed AI (Artificial Intelligence) enhancement (e.g., with Large Language Models) seemingly overlaps with a higher level effort at STScI and at other NASA archives and institutions. It is unclear from the proposal on potential collaboration (or justification for redundancy) with other archives or the community to advance these AI tools.

The proposal doesn't discuss interoperability with other (non-NASA) large survey datasets expected in the near future, most notably the Rubin Observatory's LSST. While hosting a copy of such a dataset is clearly out of scope, the proposal does not discuss the scenario to facilitate or simplify some cross-dataset analysis use-cases. The cloud hosting and science platform projects could be especially advantageous in enabling these.

HEASARC

The HEASARC is the NASA archive for high-energy astrophysics and cosmic microwave background (CMB) data, being well aligned with the Physics of the Cosmos program. The HEASARC proposes to maintain its existing archives, continue ingestion from a dozen active missions, and start the data archiving and software dissemination for forthcoming missions which fall under its purview, including the SMEX mission COSI, as well as smaller missions like the cubesat BurstCube and the StarBurst Pioneer. The HEASARC proposes to enhance its existing services through general modernization (e.g. adoption of modern package managers and support for Python), updating of old code, updating existing user interfaces, visualization software, and through data curation for AI/MI purposes. LAMBDA, the archive focused on Cosmic Microwave Background (CMB) data, will continue to operate as the main hub of the CMB community, will ingest data from new balloon missions, the partner mission LiteBIRD, and non-NASA sources (e.g. Simons Observatory), and develop standard tools for current and forthcoming data. They propose to update their website, formally take over maintenance of the General Coordinates Network, and continue to enhance interoperability through the VO standards and through partnering with the Fornax project.

Strengths:
The proposed data products are likely to support the needs of much of the relevant community, and satisfy NASA directives on Open Science. The software the archive maintains is already at the core of the X-ray and CMB community, which will benefit from the proposed modernization effort. Upcoming new missions will utilize HEASARC to store their data, and there are no obvious technological challenges in supporting them. Increased utilization of Python and C++ are commendable. Continued virtualization of the hardware infrastructure will ensure future flexibility.

The HEASARC has an established track record in FAIR principles, particularly with respect to leadership in engagement with other archives through its contributions to the Astronomy Data Centers Collaborative Committee (ADCCC), the Astrophysics Data Centers Executive Council (ADEC), and the modern VO / NAVO / IVOA programs. They have broadly adopted and implemented VO standards in the data they house, as well as their value-added products such as Xamin and SkyView. These inter-archive efforts are critical to modern astronomy.

The HEASARC began supporting the General Coordinates Network (GCN) prior to the recent ISFM supported upgrade, and is proposing to maintain the modernized GCN beginning in FY26. This is a foundational piece of the modern time-domain and multimessenger ecosystem and is a key part of the NASA response to the Decadal priority in this area.

HEASARC also maintains a large software stack (HEASoft) for data analysis, in particular for high-energy data. Notable examples include the xspec, cfitsio, etc. These data analysis tools are widely used by the science community, and HEASARC staff actively contribute to the development and maintenance of this software stack (though some are developed by third party). They play a similarly important role for setting standards for, and providing access to, calibration of high energy data.

The HEASARC houses data from a few dozen missions with data reaching back more than 50 years, is actively ingesting data from a dozen HEA missions. HEASARC provides robust hosting for data from these various and differing missions and provides software to analyze their data. The community will strongly benefit from the additions of new datasets and tools to the portfolio. HEASARC robustly houses the data, services it with reliability, and does a great job despite the vast and complex datasets under its domain.

LAMBDA is the de facto CMB archive and is producing software of key importance to the community, ingesting data from 8 active missions with expected updates to approximately 10 more missions during the proposed period of performance. Further, it is successfully including data from missions across agencies, those that are privately funded, and international instruments. These include key Decadal recommendations external to NASA, including CMB-S4. The breadth and completeness of this archive, across national and agency boundaries, are well suited to meet the needs of this community, and to act as a key access point for scientists in other disciplines.

The HEASARC is actively engaged with the community through workshops and demonstrations at major US astronomy meetings every year. They routinely conduct user surveys and engage with a users group designed for continuous feedback with experts from the external community. They have recently refreshed the Xamin interface for a better user experience. In addition, HEASoft/webtools is an indispensable tool for the broad high-energy astrophysics community for data analyses and proposal preparation. The organic structure of data access and data analysis tools at HEASARC strongly promotes community use of archival data.

HEASARC data contributes to the publication of nearly 2000 papers per year for missions with data at HEASARC. It is an exceptionally productive scientific resource. Further, it is a highly efficient investment by the metric of publications per dollar (as compared to other NASA costs,

like Guest Observer programs). The HEASARC is widely used by the community, as supported by the usage statistics summarized in Figure 2.3 as well as the overall increasing usage trend across all metrics reported. This is direct evidence of the success of promoting the community use of archival HEASARC data. Furthermore, making data available within the Fornax science platform is likely to dramatically reduce the barrier to entry and further increase the user community.

The HEASARC maintains a massive dataset, serves large quantities of data to both the high energy and CMB communities, and does so with high reliability. The utilization of virtualized infrastructure and remote backup infrastructure, delivery of software through Docker and package managers, upgrades to the web interface layer, and future integration with Fornax ensure the HEASARC contents and products are available to ever-growing demands on the archives. Additionally, the adoption of the Kafka broker and general modernization of GCN demonstrates use of state-of-the-art techniques, and is well received by the community.

The existing HEASARC SciServer is well utilized and particularly cheap. The experience gained here is useful for Fornax. The planned support for SciServer given the cost effective nature until Fornax is ready is well motivated.

Weaknesses:

The HEASARC software infrastructure is largely complete for the study of X-ray point sources, but is incomplete for some missions which fall under the purview of HEASARC, which includes many of the recent and forthcoming missions (IXPE, COSI, StarBurst, BlackCat, AMS, PUEO, TigerISS, XL-Calibur). This negatively affects science utilization of active and forthcoming missions. The HEASARC products generally work well for analysis of single spectra from X-ray point sources. However, a growing portion of the missions housed by HEASARC do not fit into this area. These missions appropriately develop their own software to ensure full use of their scientific area. The proposal does not detail plans on how and when HEASARC will take over this software as missions end, how HEASARC will guide users between mission-specific toolkits and the more limited mission-specific HEASoft plugins, how mission-specific software will be supported in cloud environments including Fornax, how HEASARC will make Python APIs easily accessible for the community (as other archives are focusing on), before the missions end.

For example, distinct from HEASARC deliverables, IXPE has their own analysis software, which is necessary for taking full advantage of the complex data and capabilities of IXPE (i.e., allowing *spatially resolved* spectropolarimetric analysis, important for disentangling the physics which underlie supernova remnants and local jets), which stands in contrast to the statement in the proposal that IXPE has fully adopted HEASARC software. Additionally, the IXPE team contribution to HEASoft may mislead the community to believe this implementation is complete. While the proposed contains plans related to this weakness, it is requested as an overguide, while it should be a key priority of the archive.

The proposal did not provide a clear plan on building the new UI with critical input from professional subject matter experts in the area of user experience (UX) design, nor on the use of focus group feedback. It is unclear whether the Goddard expertise here is sufficient. Without such

guidance and investment it is unlikely that the interface will be intuitive. For example, relatively mundane choices -- from the number and placement of buttons, to what to offer and where on the screen, can make a significant difference in terms of barrier to entry especially for novice users. The XAMIN UI has been improved, but such limitations remain.

The HEASARC does not have major Python APIs and corresponding documentation usable by the community and the proposal fails to discuss plans to implement them. The emphasis on the XAMIN UI places HEASARC at odds with the broad field focus on Python. In response, numerous members of the community have had to develop their own Python APIs for their specific needs. Recently the community has developed broader Python APIs, such as the Gamma-ray Data Tools, which are being adopted by the community, and are not supported by the HEASARC. The lack of development of Python APIs by the HEASARC and lack of engagement and support for community efforts has continued a barrier to entry in the field and precluded alignment of high-energy data access with broadly adopted methods.

IRSA

The NASA/Infrared Processing and Analysis Center (IPAC) Infrared Science Archive (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 (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, 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 and 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 is the key archive of infrared and submillimeter data for various NASA missions as well as NASA partners/agencies since the IRAS mission. IRSA curates the data of all these missions to the science community, for queries, visualization, retrieval and scientific exploitation following the guidelines of NASA's Open Source Science Initiative. IRSA intends to archive a growing set

of data from new missions in the next years, by storing them in the cloud with the Fornax science platform and contribute to the open science initiative.
As we are moving to an era where most of the research and publications are archival in nature, IRSA is extremely well positioned to maximize the scientific return of NASA.

In the next 5 years, IRSA's holding volume will increase by over an order of magnitude, as the data of new survey missions will be ingested, especially Euclid (in close collaboration with ESA) SPHEREx and NEO surveyor. IRSA presents a clear plan to ingest and curate these very large sets of data. IRSA will continue to ingest in parallel data from IRTF and NEOWISE as well as several community-contributed data and simulated data. IRSA integrates mission data into interfaces that serve data from multiple missions. Most of these tools are based on IVOA standards and the Firefly toolkit. This makes the mission-specific web application easy to maintain and improve.

IRSA is dedicated to supporting research with a very proactive approach to support users. The community outreach includes workshops where NAVO notebooks are demonstrated, regular newsletter as well as continuously updated video tutorials. IRSA's helpdesk is also very appreciated for its fast reaction time to answer questions by mission experts. IRSA get also advises and recommendations from a User Panel as well as user feedback from user surveys.

IRSA is evolving its archive to adapt to the latest tools offered by the International Virtual Observatory Alliance (IVOA), in which it is significantly involved too, through the NAVO collaboration. It has adapted recently the Common Archive Observation Model, with a rich set of standard metadata describing observations. IRSA also adopted ObsCore that ensures increased interoperability with other data centers and the IVOA Multi-Order Coverage (MOC), which enable fast identification of subset of data sets that have relevant spatial coverage. It is also planned to update this tool to include temporal information, using IVOA Space Time MOCs (STMOCs), hence at the forefront of archive technology.

The open-source Firefly implementation of the IRSA viewer is used by NED and NEA and has also been adopted by the Rubin science platform. It could be used by other archives, and can be used in Python layers, including as JupyterLab extension to make it a more user-friendly experience as it can be used to visualize a large variety of astronomical data products.

In order to host large catalogs in the cloud, IRSA is also working on ways to make it easier to mine large catalogs by serving them in Parquet format, a cloud-friendly and analysis-ready table format used in many domains outside of astronomy. IRSA is contributing to a joint effort to define a community standard (currently named HiPSCat) for organizing the rows within Parquet files to enable efficient cross-matching of large catalogs from different surveys, a critical step for a large fraction of science investigations. All new survey catalogs could be released in HiPSCat format in the cloud.

IRSA will therefore be in an excellent position to develop joint-archive multi-wavelength catalogs hosted in the Fornax system or even joint-pixel analysis Euclid-Roman-Rubin in a further stage. Such cross-mission science products were advocated by the Astro 2020 Decadal Survey.

The NASA Sky initiative, a user-friendly web application to support discovery of public data from all NASA missions is welcome. Analogous to the popular ESAsky but making use of the latest IVOA standards. The work plan is realistic with current and planned technologies: ObsTAP implemented in all archives by FY25 to query observations as well as IVOA Multi-Order Coverage (MOC) standard.

Minor Weaknesses:

The design and the user experience of the archive user interfaces (e.g. Firefly), while powerful, are difficult for new users to handle. Continuously engaging UI/UX experts to simplify and modernize the user interface look and feel may reduce the barrier to entry, speed up user work, and maximize the overall usability of the archive.

The over-guide proposal for super-resolution WISE images is seen more as a science proposal and not directly in the scope of curating WISE data. Enhancing the resolution of astronomical images with diffusion models by training the models on Spitzer/IRAC 3.6 and 4.5 microns images is a promising technique but there are other techniques. And the positive outcome of such a research proposal is not granted. This project is planned to start once the WISE finishes to beat the confusion noise. However it is recommended to first make a full coadd as the unWISE (https://unwise.me/) project has done up to year 7 and made available in the legacy survey (https://www.legacysurvey.org/)

NAVO

NAVO's goal is to maximize the return from NASA's astrophysics science mission portfolio, by enabling seamless combination of archival data across NASA, opening up new science opportunities. Strategically, NAVO promotes "FAIR" principles to make data Findable, Accessible, Interoperable, and Reusable, with a major focus on aligning data practices with the standards of the International Virtual Observatory Alliance (IVOA). NAVO functions as a collaboration between the HEASARC, IRSA, MAST, and NED archives. Its activities have included bringing data access interfaces to international standards, providing a VO registry, participating in IVOA leadership, developing open source Python tools (primarily PyVO), and training the astronomical community. This proposal aims to extend the application of open science principles in astronomy, to enable cross-archive science, and to enable science from cloud-based holdings. Each archive provides about 10-15% of its budget towards NAVO goals. It is led by a PI and a Project Scientist, who coordinate regular meetings with the NAVO contacts at each archive and deeper technical meetings among the archives. There are currently about 7 FTE associated with NAVO, decreasing to about 6 FTE at the end of the proposal period.

Strengths:

NAVO's activities are integral to the functioning of the participating archives and to their future plans, and are broadly used by developers and end-users in the astronomical community. Services backed by NAVO technology are queried around 50–100 million times per year. These services support the archive's user interfaces but also allow direct API queries through protocols like Simple Image Access, Table Access Protocol, and others. NAVO produces reference documentation and tutorials that enable the broad use of these tools. Its participation in IVOA represents a strong U.S. contribution to international astronomy, enhances the impact of NASA missions, and strengthens the development of modern standards. These achievements demonstrate that NAVO is meeting its high-level goals towards archive interoperability.

This proposal promises to sustain and expand upon NAVO's previous work. NAVO will continue to participate in IVOA, promote implementation of IVOA standards in modern software, refresh infrastructure, conduct workshops, and develop documentation. It will complete the implementation of ObsCore across the archives to standardize observation metadata, implement STMOC in the NAVO Registry to improve findability with spatial and temporal information, work on updates to IVOA standards for cloud-friendliness, and incorporate simulated data into data models. These activities all strengthen the backbone upon which all other archive activities increasingly rely.

The NAVO team works in a collaborative and cooperative manner, with an evident rapport and common vision among its leadership. This collaboration has been successful at building ties between the archives which potentially will pay high dividends in the future with an increasingly strong relationship between the archives.

The NAVO strategy and tactics are squarely in support of Open Science principles and of increasing the scientific potential of the archives in practice. NAVO's activities are demonstrably motivated by the FAIR principles that promote this increase in access and maintaining and developing them are the prerequisites for a sustainable archive system. NAVO is addressing new technological opportunities and challenges, for example adapting to cloud-based data.

Enabling increased programmatic access to the NASA archives, individually and in combination, has the potential to broaden participation in large data set astronomy beyond those individuals and institutions who have the resources to host significant portions of the full data sets.

The proposal provides a clear and compelling description of the activities required for the continued functionality of NAVO, in all of its components including maintenance of the PyVO software, the NAVO Registry, management, and its leadership in IVOA. The planned development of new standards (such as ObsCore) and the focus on the standards needed by cloud systems will make it easier for users to access data across different archives, including HEASARC, MAST, IRSA, and NED.

The vision behind NAVO is to provide a productive basis of cooperation between archives to increase science end-users' ability to conduct cross-archive science with the archives. It supports this vision through both the technical plan discussed above and by providing a cooperative and productive theater of engagement for archive managers, engineers, and scientists.

The overguide elements of the proposal were as follows, which we list in decreasing order of priority: increased PyVO functionality; monitoring and validation of the NAVO Registry services; and additional workshops and documentation. The proposed PyVO functionality is highly important to the development of Fornax and the support of the new, incoming, enormous data sets, including large new spectroscopic data sets. The monitoring and validation services are important to maintaining the usefulness of the NAVO Registry and the reputation of NAVO within the framework of IVOA. The workshops and documentation would add value given the limited resources available for documentation in NAVO. There may be ways to incorporate some of these activities at a lower level within the in-guide budget.

Minor Weaknesses:

At present, NAVO's direct customer for its software products is the community of astronomical archives, who in turn serve end users. End users can and do use products like PyVO more-or-less directly as well, and NAVO has a desire to serve these end users directly as well. However, there is a lack of "user guide" level documentation, and not enough outreach and training workshops to saturate the potential user base. The lack of documentation reduces accessibility and increases the likelihood of accidental misuse of data leading to incorrect scientific conclusions.

This prioritization of customer class and documentation is made in the face of limits in the available resources, not because of a lack of understanding from the NAVO team.

NAVO and the Fornax Initiative both have the potential of moving NASA to the forefront of one of the likely future and pressing needs of the community: making most astronomical data accessible in a unified fashion. Although it is clear that the NAVO and Fornax teams are overlapping and in communication, it is not clear from this proposal how well integrated their development plans and times are.

The management of NAVO is conducted as a cooperative arrangement that utilizes relatively small fractions of individuals' time, with NAVO tasks working around the archives' overall workflow. Due to the complexity of each archive's schedule, it is not possible to plan forward as one would for a normal project. This state of affairs complicates its implementation and means that achieving fixed goals on a fixed timeline is unlikely.

NASA Exoplanet Archive (NEA)

The NASA Exoplanet Archive (NEA) and the Exoplanet Follow-up Observation Program (ExoFOP) support research and mission planning by the exoplanet community and NASA missions. The NEA and ExoFOP are an integrated activity. The primary goal of NEA is to provide the scientific community with a complete and accurate accounting of exoplanetary systems published by NASA missions and by the community in the refereed literature. The NEA creates visualization and analysis tools to enable the easy extraction and exploration of data for analysis and for planning future observations and provides access through modern data access protocols. The primary goal of ExoFOP is to provide the exoplanet community with a venue for coordination and sharing of follow-up and precursor data for exoplanets, their host stars, and

stars that might eventually be targets for future planet searches. The primary objectives of the NEA for the 2025-2030 period are: continued support for the community and NASA's exoplanet program; keeping pace and ingesting relevant new information; and continuing to modernize access to the data.

Strengths:

The ease with which information can be obtained from the NEA and visualized onsite has led to its near dominance in the exoplanet field as the premiere repository of information on individual exoplanets and exoplanet systems that are used by exoplanet scientists worldwide. The proposal demonstrates this in a number of ways: (1) the citations to NEA far surpasses other exoplanet databases around the world by a factor of 3 or more in the last several years; (2) the number of hits to NEA increased by an order of magnitude before the JWST Cycle 2 deadline, showing the community's reliance on its services; and (3) the number of papers submitted to arXiv that cite NEA has steadily increased since 2010 in concert with the expansion of the exoplanet field.

The NEA maximizes the scientific return from NASA astrophysics missions by being the official repository of exoplanet parameter sets stemming from several such missions, including Kepler/K2, TESS, JWST, ASTERIA, Spitzer, and Keck. This allows for holistic understanding of exoplanetary systems and double checking of exoplanet data by amassing information from a variety of sources across observational platforms and wavelengths. In addition, in support of Open Science, NEA will provide NAVO-compliant, TAP based servers, in preparation of being science-platform ready.

ExoFOP has become an integral part of the Kepler/K2 legacy and TESS mission, as demonstrated by the rapid increase in papers with time that refer to ExoFOP and use its resources, as well as a glowing mention in the 2020 Astrophysics Archive Programmatic Review report.

To make ingesting new data more efficient, the NEA is using a new machine learning classifier that has already seen success at NED to identify relevant papers for ingestion. This method sees relatively low false negatives/positives (roughly 1%) and still ultimately relies on human staff to recognize the relevant papers so that no data sets fall through the cracks. To make this effort even more efficient, the NEA is planning to offer templates for standardized data ingestion to users and journals so that NEA staff do not need to look through the individual papers to find the preferred parameter sets.

The NEA will capitalize on the availability of new datasets by ingesting exoplanet data from publications stemming from future NASA (e.g. Roman, HWO) and non-NASA (e.g. PLATO, GAIA) missions and facilities. Specialized tables will be created for microlensing and astrometry, which will see increased use with future Roman and GAIA data releases. These new tables would add to the holistic understanding of new and current exoplanetary systems.

As recommended by the NEA user panel, the NEA will allow API access to planetary system overview content and tools in the next five years, enabling users to query the NEA directly from their own workspaces, e.g., python notebooks, as well as link to the tools. This will allow the

user base to bypass the online platform and directly interact with the data contained within the NEA in their workflows.

The addition of a fourth NEA scientific staff member with expertise complementing the current three members and appropriate for the arrival of Gaia and Roman data is well motivated by the proposal and the case for this hire is well described.

Minor Weaknesses:
The proposal notes that upcoming challenges are not due to data volume, but data complexity. However, while the proposal demonstrates that the NEA is aware of this issue, it does not demonstrate that NEA can successfully deal with such complexity within the time limit imposed by the upcoming mission timelines.

For the atmospheric spectroscopy tables, the proposal does not discuss which data reduction is ultimately chosen to be featured in the NEA in the case multiple are presented in a paper, nor which atmospheric models will be used for comparison purposes and why.

The proposal does not supply sufficient detail on the timeline and milestones of technological developments. Even though the NEA is a “living-breathing archive” that advances with the needs of its users, the development of the archive may be affected negatively without a vision of the path forward or a plan that stretches over multiple years.

Astrophysics Data System (ADS)

The Astrophysics Data System (ADS) is used by nearly all astronomers to search the literature, and sometimes for other tasks such as finding astronomical data. It is a vital resource for the astronomical community. The proposal is primarily to expand ADS to the other fields in NASA's Science Mission Directorate (SMD), namely, (roughly in order of maturity) Planetary Science (PS), Heliophysics (HP), Earth Science (ES), and Biological and Physical Sciences (BPS). The new repository will be called Science Explorer or SciX. Expansion of ADS to these other fields is at NASA's request. The team structure will evolve to include discipline-specific project scientists, paving the expansion to new communities and reflecting a broader vision of fostering interdisciplinary communication, collaboration, and research. The depository will make use of recent advances in AI/LLM (Large Language Model), and include knowledge graphs and other features to expand discoverability and accessibility. The proposal includes funding for engagement with the relevant scientific communities to obtain feedback and advertise the new service.

Strengths

ADS is *the* way that astrophysicists and astronomers access scientific publications. Essentially every researcher in this field uses it. It generally works very well, and it is crucial that it continues ingesting and maintaining the astrophysics literature.

SciX will be an improvement over the way researchers in those fields currently find publications using, e.g., Google Scholar or Web of Science (WoS). SciX has a better publication vetting

process, better citation metrics, higher quality metadata, better search functionality, and, unlike WoS, does not require a paid subscription.

SciX will enable significant science return through the ease with which users can reach a tremendous compilation of NASA SMD knowledge. The proposal demonstrated its utility for facilitating interdisciplinary research.

Astrophysics will be at the center of SciX. ADS will be largely unchanged, and the proposal makes it clear that the expansion to SciX will leave existing crucial features of the current ADS intact. One individual will have the job of ensuring that they do not "break" ADS.

The baseline scenario will allow near-full development and ingestion of publications and data in the most developed fields (AP, PS, HS).

ADS is essential to the astrophysics community's access to NASA-funded astrophysical research, as well as astrophysical research funded by other agencies both domestic and foreign. It allows users to quickly access all of the astrophysical literature published in most, if not all, astrophysical journals, and most relevant related fields.

The proposal includes a number of well thought-out plans to advertise and promote SciX to the wider scientific community, including hiring a community engagement coordinator, presence at conferences, social media, videos, and a SciX Ambassadors program. It will take in information from the scientific community through feedback from individuals, analysis of usage logs, and an advisory group modeled after the ADS Users Group. Many (but not all) of these can be implemented under the baseline scenario.

The SciX hired a UI/UX designer to completely overhaul the interface and made modifications in such a way that an interface design can be tailored to support specific needs in each of the SMD disciplines. It provides enhanced features that improve the user experience and promote cross-discipline discoverability.

SciX has expanded its team to include discipline project scientists to increase the discoverability of the literature and data by establishing science-based priorities and implementation plans in collaboration with the existing data centers, such as the Science Discovery Engine, to prevent duplication of efforts.

The proposal includes updated ingestion pipelines with newer Apache Kafka technology, most of which can be completed in 5 years under the baseline funding scenario. Increasing size of data/documents being searched will require updated architectural changes that can be realized under the baseline funding scenario.

SciX will use state-of-the-art machine learning and artificial intelligence (ML/AI) techniques to mine texts and enable searches that go beyond matching words in titles/abstracts or author names. Taxonomies will be added (e.g. UAT) such that searches of concepts will also be possible, thereby connecting multiple disciplines through their shared ideas. The example of the

NER metadata enrichment process for recognizing planetary features shows promise. ML/AI will also allow for automatic classification of papers and data sets into their specific collection.

Existing efforts are supporting an enrichment of the metadata called Named Entity Recognition (NER) on the SciXBrain system which can be applied to a wide variety of metadata information and lead to the creation of additional training data to recognize entities within these metadata categories. This is done in a free and open environment which promotes re-use and increases the scientific return on the development.

The new data ingest pipeline in development has improved the efficiency and flexibility to adapt to the new types of information and data ingested. It is more robust to handle the increased workflow. The proposal suggests improving some of the processes to allow for automation and reduce the amount of manual intervention, in turn reducing costs in the long run.

Additional data management techniques are continually being used to enhance the richness of the data in the system.

Fully funding the proposed augmentation would allow the proposal to meet most of its goals--expanding SciX to fully ingest most SMD publications--in 5 years. Failure to fund it at this level will delay implementation and ingestion of publications for fields where it is currently less developed (ES and BPS). The proposed path to improve discoverability through AI/ML will heavily rely on having the expertise and computing infrastructure to support it. Fully funding the proposed augmentation would allow a greater ability to recruit and retain the necessary personnel. Fully funding the proposed augmentation will allow for greater public outreach activities and increase the ability to get crucial feedback from the scientific communities SciX is designed to serve. Between 10% and 40% of the user support activities (liaising, collaboration, user outreach, conference attendance, maintaining user documentation) planned under the augmented funding scenario will not be implemented under the baseline scenario. The augmented scenario will allow funding for greater automation of text mining, metadata normalization, record classification, and updating the ranking algorithm. These will be only minimally advanced in the baseline funding scenario.

Minor Weakness:

An advisory board meeting once per year and other forms of community feedback discussed in the proposal may not provide sufficient feedback from the relevant scientific communities, especially in the early phases.

NED

The NASA/IPAC Extragalactic Database (NED) aims to be a complete catalog of known extragalactic objects. It is an invaluable resource for researchers wishing to quickly find, in one place, vetted information (such as its redshift, data taken, publications) on an individual object or a sample of objects. The baseline proposal outlines three elements for the next phase of NED: enhancing support for TDAMM (Time-domain and Multi-Messenger Astronomy) through tools

useful for follow-up of gravitational wave transients, continuing the ingestion, curation and cross-referencing of data from journal articles (at a reduced scale, limited by budget), and the ingestion of data from large NASA and non-NASA datasets. The over-guide proposal includes funding for ingestion of data from MNRAS (only done on a best-effort basis in the baseline proposal), further enhancements and tools for time-domain and multi-messenger astronomy, and the development of an AI tool to vet matches of objects with sources in NED.

Strengths:

A provider of critical infrastructure: NED is a long-standing service that has been in operation for over 30 years. The archive has been highly successful and broadly relied-upon by the community. Measured by API requests, NED is frequently utilized, with about ~3 API requests per second. It provides services (such as the NED Name Resolver) that have become an indispensable element of worldwide astronomical infrastructure. Because of this role, it is critical NED's highly queried infrastructure services continue without major disruption.

Increasing support for Time-Domain and Multi-Messenger Astrophysics (TDAMM): The proposal's focus on expanding services to support time-domain astronomy and providing metadata and links to time series data in external archives is timely and appropriate. This is a specific area where curated, high-quality, information for (mainly bright) sources is invaluable in making follow-up decisions. NED's new gravitational wave (GW) follow-up service uniquely leverages its holdings to provide critical targeting information in searches of electromagnetic GW counterparts.

Comprehensive Data Fusion, including from papers: NED provides a valuable service in fusing and making readily-available data across the electromagnetic spectrum from NASA missions and science publications (papers). The ingestion of data from papers is especially valuable and unique; it allows NED to be a "one-stop-shop" destination for data on specific objects. This drives a case for prioritization (and overguide funding) to maintain this aspect.

Data Processing and AI Integration: Plans to enhance data processing efficiencies and the integration of AI and machine learning for data extraction and cross-matching are forward-looking. These efforts are necessary to handle the growing data volumes and complexity, thereby continuing to make NED viable as the community grows.

There is an explicit plan for the ingestion of major upcoming large datasets, especially those bringing significant redshift information. These include Euclid, SPHEREx, Roman, and ULTRASAT, as well as non-NASA datasets such as PS1, DESI and ultimately LSST. These would strengthen the utility of NED (with the caveat mentioned earlier about possibly focusing or bright/near subsets of these catalogs). Furthermore, a proposed development of a web-based UI to enable authors to contribute data files to NED with an initial set of data validation checks will help increase the rate of ingestion of data.

AI (LLM) tools will be used to increase the rate of ingestion of data from the literature. The team has already executed a small pilot in this area, and are confident that such tools can be deployed successfully. As we see the journal data ingestion as a critical and unique NED service, we strongly endorse this aspect of the proposal.

The archive engages the community through the users group, surveys, and presence at topical meetings. The community appears engaged and strongly supports the service. The proposal includes a NED Users Committee, which will help obtain feedback on NED from the community and allow them to make improvements. The proposal includes funding for outreach at conferences and through a NED ambassadors program. The team is increasing community outreach and user support through the hosted workshops and Python notebooks, as well as through "ambassadors" to help with community communication.

The proposal includes funding for updating equipment, scaling up capacity, cyber security, and disaster planning. The proposal includes AI assisted data extraction from the literature, improvements in the data integration pipeline, and streamline capturing of data and metadata.

The proposal includes work for additional containerization for future deployments in the cloud, and enhancing services available through APIs. They have an established workflow and it is being improved to be more efficient in the data management processes.

Over-guide funding will allow ingesting of data from MNRAS, quicker development of APIs and sample Python notebooks, more user support, and provide tools for more advanced searches useful to users. The over-guide proposal includes funding to enhance multi-messenger astronomy, including a tool that allows one to find time-domain data, and extending the gravitational wave follow-up service to facilitate rapid follow-up of neutrino alerts and gamma-ray bursts, supernovae, and other transients. It will also fund an application of AI to vet the results of automated matching of sources with objects in NED; and a service to link public data at other places to NED.

We strongly endorse the need for all these activities. We also think continued data ingestion from papers (incl. AI development) and enhancing the TDAMM support (especially in the area of supporting GW follow-up) should have been prioritized within the in-guide plan, over the extensive large dataset imports.

Weaknesses:

A clearly defined guiding principle on what is the scope of the data NED wishes to ingest, and *the science that NED can uniquely* support, is missing from the proposal. Where can NED make the most impact in the decade to come, and how do proposed activities serve to maximize that opportunity?

Organizing NED's activities around supporting TDAMM could provide such a "guiding light". Rather than compiling a master dataset of all information about all extragalactic objects, a comprehensive database of all objects within 1 Gpc (or to some appropriate apparent magnitude cut) may provide invaluable benefits to TDAMM while responding to budgetary realities. It may reduce the technical demands, and allow NED to add even more information for that particular subset of objects. This includes continuing to comprehensively incorporate object data from the literature, something we see as a critical and unique service that only NED provides.

Presented plans to fully ingest a sequence of increasingly larger datasets may not be a way to maximize end-user value (we discuss some alternatives further below). For large statistical

analyses, upcoming science platforms solutions (with on-the-fly or pre-computed join tables) are likely to realize greater efficiencies than a fused database (and one with a complex selection function). The increase in scale also makes it technically difficult to continue offering a comprehensive, merged, human-vetted, per-object database for every single object (even if only a subset of columns is ingested). The cost/benefit ratio of such a pre-joined dataset is not clear: how often is one likely to inquire for all details of a barely resolved 26th magnitude galaxy? It's not clear that comprehensiveness should be NED's priority.

The proposal did not provide sufficiently detailed and fine-grained measures of the utilization of specific services that NED provides. The usage of NED (as measured by citations) has grown consistently, peaking in the mid-2010s at ~750 citations/year), with a small but consistent decline over the past five years to present-day ~600/yr. Some of that decline may be due to forgetfulness of authors to appropriately cite the archive. But it is also likely that, with the growth of science with large datasets and increased joint-dataset-analysis capabilities at other archives, at least a part of this decline reflects a true reduction in usage. This is difficult to discern without more detailed metrics.

The baseline proposal only includes funds to ingest data from the most important papers from MNRAS. This is a unique aspect of NED data holdings, one without a comparable replacement world-wide.

The proposal does not include ingestion of data from some other important journals such as Science, Publications of the Astronomical Society of the Pacific, and Publications of the Astronomical Society of Japan, Nature Astronomy, and others.

Maximally front-loading the development and deployment of AI may increase productivity earlier and help mitigate the reductions otherwise required by the flat budget. The presented plans assume a long timeline (2026-27) for the deployment of AI-driven agents for literature data ingestion. Given the recent improvements in LLMs, it is possible commercially available fine-tuned AI agents (GPT4, Claude3, and others) could take on this job earlier.

The team's software development methodology and tooling does not appear to be fully leveraging present-day best practices. Version control is in SVN, and no plans have been discussed in the proposal to migrate to git or github.

Key software (MatchEx) is not open source and available to the user community. For large cross-matched dataset to be useful for statistical analyses, understanding the performance (selection function) of the cross-matching algorithm is crucial. Access to the code is required for such use-cases.

Required database architectures may be more complex than anticipated. The team is presently piggybacking on IRSA's efforts to study the technical options for the large datasets. But given the difference in structure and access patterns of IRSA vs NED databases (single-mission homogeneity vs multi-mission heterogeneity, write-once vs. frequent update, etc.), it's possible that a substantial additional effort may be needed to find a workable solution.

Improvements

The proposed scope could roughly be thought of as consisting of four parts: 1. core “low level” services/APIs (such as the NED Name Resolver, and the existing database), 2. TDAMM support services (such as the GW Follow-up tool), 3. regular ingestion of data from the literature, and 4. ingestion of new large datasets. The ordering here reflects the panel’s suggested prioritization of the four components.

Complete ingestion of all objects from new large datasets should not be a goal. To thrive in the era of large datasets, data lakes, and science platforms, it will be critical for NED to focus on activities and services adding unique value that cannot be found or easily replicated elsewhere. We believe these lie in supporting TDAMM use-cases, and aiming for comprehensive (ingest data from more datasets, more papers) coverage of a smaller subset of objects most valuable to TDAMM use cases (e.g., a 1 Gpc or V<23 sample).

Exhaustive cross-matching of large datasets is generally valuable, but with outputs in the form “join tables” at various archives rather than transformed databases hosted at NED. Because of that, we recommend the NED team to work on open-sourcing their MatchEx algorithm and code, and make it executable and available on the NASA Science Platform.