Astrophysics Archives Review 2024
Proposal submitted by
The NASA Astrophysics Data System Project
Smithsonian Astrophysical Observatory

NASA Science Explorer

Creating an Innovative, Multi-Disciplinary Information Nexus

Alberto Accomazzi, Michael J. Kurtz, Edwin Henneken, Kelly Lockhart,
Jennifer Lynn Bartlett, Stephanie Jarmak, Lisa Kewley
Center for Astrophysics | Harvard & Smithsonian

February 6, 2024

Executive Summary

The NASA Science Explorer (SciX) will be a permanent component of open science infrastructure. At NASA's request, SciX builds on the 30 year legacy of the Astrophysics Data System (ADS) by expanding it to encompass all scholarly disciplines supported by the NASA Science Mission Directorate. Including Astrophysics, Planetary Science, Heliophysics, Earth Science, and Biological and Physical Sciences marks a significant expansion in scope and collaborative potential. This evolution signifies not just content growth but a reimagining of what a scientific repository can be—a comprehensive nexus for open science and the discoverability of NASA-funded research.

The strategic shift from ADS to SciX is marked by a significant change in team structure to include discipline-specific project scientists, paving the expansion to new communities. This evolution reflects a broader vision of fostering interdisciplinary communication, collaboration, and research, and collapsing silos between scientific domains. By doing so, SciX embodies the principles of open science, making NASA-funded research more discoverable and usable across various scientific communities. The disciplinary expansion is supported by the technological advancements in the rapidly evolving, transformative field of text-based Artificial Intelligence (AI). SciX will have all the relevant research text and connected metadata, enabling services combining language models and knowledge graphs to support greater discoverability and accessibility to this content. Combining expert curation with sophisticated, responsible AI into the project underscores a commitment to maintaining the highest standards of trust and relevance in scientific information dissemination.

This document presents first a discussion of the general motivation, philosophy, and vision which drives the project. Following that, we will present some significant achievements and advancements since the 2020 Astrophysics Archive Review, along with the additional work we have done to expand from ADS to SciX. In the next five years, we aim to enlarge and enrich our collections and collaborations while also advancing our technical capabilities to unify NASA science discovery on a single platform. However, the realization of SciX's full potential is contingent upon strategic investment and support. The risks of underfunding the project during its critical expansion phase carry implications not just for the project but for NASA SMD's open science ecosystem and for the global scientific community that depends on it. It is imperative that the necessary resources are allocated to SciX, ensuring its position as a cornerstone of open science and a model for future scientific collaboration and discovery.

SciX stands as a pivotal project for the future of open science, embodying a critical leap forward in how scientific research is accessed, shared, and utilized. Its success is essential for enhancing the discoverability of NASA-funded research, fostering global scientific collaboration, and ensuring the continued advancement of knowledge across disciplines. The transformation from ADS to SciX, with a focus on discoverability, access, and interoperability, marks a significant evolution of NASA's efforts to promote interdisciplinary communication, collaboration, and research.

Table of Contents

1. SciX Mission

The NASA Science Explorer[1] (SciX) is a transformative digital library that fosters open, interdisciplinary science across astronomy, earth science, heliophysics, planetary science, physics, and biological and physical sciences while substantially increasing overall research efficiency of these disciplines. SciX will connect NASA data repositories, increasing the science return on investments in individual archives. It is the natural extension of the NASA Astrophysics Data System[2] (ADS).

For more than a quarter of a century, the ADS has been the primary means by which astrophysicists search and discover their scientific literature. During this entire period, the average astronomer has used the ADS daily. Now, it is highly unlikely that a single astronomy graduate student anywhere in the world does not know how to use ADS. In less than a decade, we will say the same about SciX and earth science graduate students.

The underlying vision of SciX is that the concentrated knowledge of science can be found in the written literature of science. SciX empowers scientists to maximize their use of this knowledge, both by direct, intelligent search and by linking documents to associated information sources. Because scientific literature is the best discovery tool for data, SciX will be the most powerful discovery tool for all of NASA-funded data through its interconnected literature AND through connecting literature to data. Thirty years of ADS growth demonstrates the power of this approach (Kurtz et al 2005).

With steady NASA support, SciX is continuously improving on the foundation NASA built through ADS. SciX has already made substantial enhancements to the depth and quality of the information in its databases. We are constantly modernizing the system, making use of new technologies and techniques as they are developed. Importantly, the corresponding increase in staff ensures that the unexpected loss of a key individual is not catastrophic.

ADS, the SciX precursor, is now a widely used, stable and trusted component of the infrastructure of science. Its reliable functioning undergirds a substantial portion of astrophysics research. From its inception, 30 years ago, ADS has been a leader in providing open access to scientific research, following NASA tradition (Kurtz, 2020). SciX will continue to be an essential, reliable, trusted, and transparent partner in open science research.

1.1 Motivation

Since its inception, the ADS has been a transformative project in the professional lives of astronomers worldwide. Its universal adoption in astronomy has been followed by an increasing level of adoption in related space science disciplines, in particular heliophysics, astroparticle physics, plasma physics, and planetary science. ADS has *all* the content required for conducting research in the fields it covers, advanced search capabilities, authoritative metrics, and discipline-focused features that maximize their research productivity.

[1] https://scixplorer.org
[2] https://ui.adsabs.harvard.edu

Over time, ADS increased its coverage of research fields connected to astronomy to improve the discovery of related scholarly content. This augmented selection allows astronomers to see connections beyond their discipline and promotes the efficient dissemination of new ideas. An astronomer can easily find and read papers on machine learning that are cited by astrophysics articles, while an instrumentalist can evaluate new technologies being used in space science research. In an era of expanding interdisciplinary research, ADS has fostered the discovery of content across research fields rather than let its capabilities remain siloed in a well-defined, well-curated discipline.

ADS users and advisory committees have consistently requested full inclusion of emerging interdisciplinary fields, such as multi-messenger astronomy and exoplanet research. ADS with NASA support responded willingly to such requests. However, ADS coverage of heliophysics and planetary science was at a best-effort level until 2020.

In October 2020, as part of its strategic goal to support open, interdisciplinary science, NASA requested the ADS team to develop a plan to extend ADS's services to all the scientific disciplines supported by the Science Mission Directorate (SMD). ADS would no longer primarily serve Astrophysics (AP) but become THE digital library for Planetary Science (PS), Heliophysics (HP), Earth Science (ES), and NASA-funded research in Biological and Physical Sciences (BPS). ADS responded with an initial budget estimate in May 2021, and, following budget guidance from NASA, submitted a proposal for the NASA Science Explorer (SciX22) in May 2022. After approval, direct funding for the project commenced in January 2023.

With these changes, Astrophysics Data System does not adequately describe the breadth of our collections nor our commitment to serving the full range of SMD disciplines; furthermore, it does not reflect the cross-disciplinary efforts we are facilitating. To celebrate these changes and to appeal to new user communities, we choose a new name, NASA Science Explorer: SciX, and with it, a new website: https://SciXplorer.org.

1.2 Vision

SciX is the manifestation of ADS at the NASA SMD level. Its expanding coverage of multiple disciplines not only enhances the quality of research in the new disciplines but also encourages collaborations across different scientific domains, leading to innovative solutions and discoveries.

The goal of the SciX project is to build and operate in perpetuity an open, state-of-the-art, authoritative, and trusted information nexus for the SMD disciplines. Our purpose is to enable scientists to do more and better science. Ultimately, SciX will be a digital library where:

1. All discipline-specific research content is aggregated, connected, and indexed for each of the SMD divisions;
2. Relevant taxonomies are used to capture the knowledge and semantics of the subject disciplines;
3. Curation and machine learning-based text mining and enrichment are combined on a platform that scales without sacrificing accuracy and flexibility;

4. Digital collections are enriched with links to other research objects, including open access and restricted items equally, such as data sets, software, notebooks, and funding information;
5. Discipline-specific capabilities and analytic services are exposed to the relevant research communities and made available for interdisciplinary use;
6. Discoverability and access to NASA-funded research artifacts and derived data products are available to all from a public search portal;
7. New and existing initiatives are developed and supported in collaboration with NASA and other research organizations.

After three decades serving the astronomy community, both worldwide and within the context of NASA SMD, SciX's goal is to serve the other NASA SMD disciplines at the same level of service we have provided to astronomy. To accomplish this, the following must be in place:

- All necessary content (refereed literature, non-refereed literature, preprints, theses), see Section 3.1
- All necessary data, appropriately linked/indexed (both from NASA Data Centers and non-NASA sources), see Section 3.2
- Bibliographies (maintained by NASA centers), see Sections 2.2.3 and 3.2.3
- Open Access and licensing information, see Sections 2.2.3 and 3.4
- Authorship, Affiliation, and Open Researcher and Contributor ID (ORCID) information, see Section 3.2.2

Accomplishing these goals requires an amount of time and effort that goes beyond the five-year development plan at the baseline funding level provided by NASA. However, in section 4 we present a plan that meets these goals at an enhanced budget level.

SciX is more than a good investment in NASA science. It is a strategic move aligned with the evolving landscape of scientific research. It harnesses technological advancements to promote interdisciplinary studies and improve data accessibility. In the current digital era, the scientific community emphasizes making scientific data and literature accessible to a wider audience, including researchers from underfunded institutions, educators, policymakers, and the public. With its broad scope, SciX serves as a more inclusive platform, providing access to a richer and more diverse set of resources. This inclusivity is crucial for fostering a global scientific community where knowledge is freely shared and where collaborations are not hindered by access barriers.

Recent technological advancements, particularly in artificial intelligence (AI) and machine learning (ML), offer huge opportunities for enhancing scholarly systems. SciX intends to use these technologies to develop even more sophisticated search algorithms, data analysis tools, and personalized recommendation systems. With these capabilities, SciX is not just a discovery platform, it is also an advanced research tool. By authorizing SciX, NASA is providing critical investment in these technologies, which will lead to more innovative uses of data and more efficient research methodologies. Through SciX and its commitment to modeling Open Science, the development of these technologies happens in the open, ensuring researchers can trust the underlying models.

1.3 Relevance to NASA's Mission

The National Aeronautics and Space Act of 1958 established NASA and charged it to *"provide for the widest practicable and appropriate dissemination of information concerning [...] its activities and the results thereof."* This very simple principle has led the agency to adopt liberal data access policies that have promoted a wealth of research in earth and space sciences since its inception. The establishment of the SMD in 2004 marked a significant step in NASA's commitment to scientific research, emphasizing the importance of a unified approach to understanding our planet, our solar system, and the universe beyond. Meeting this challenge requires not only access to the data, but support for the research ecosystem that surrounds it.

The development of a well-curated, interdisciplinary digital library for earth and space sciences aligns well with NASA's commitment to advancing the frontiers of knowledge. As such, SciX serves as the single, centralized nexus for the scientific literature, significantly streamlining the research process. One key feature is **its seamless ability to interconnect documents and data sets across disciplines through shared or related concepts, as well as co-authorship, co-citation, and co-readership networks**. In this regard, SciX is a key component for achieving a unified functioning cross-disciplinary ecosystem, with, i.e., researchers from different disciplines sharing the same set of sophisticated discovery and authoring tools available from a common source access through a single Application Programming Interface (API).

SciX democratizes access to scientific knowledge, serving as a key component of NASA's Transform to Open Science (TOPS) initiative. By providing free access to the entirety of its high-quality research database, and by linking to open access versions of articles in its collection, SciX supports a global community of researchers, students, and educators. This accessibility is particularly important for individuals and institutions with limited resources, as it provides them with the same level of discoverability as more well-funded entities. This democratization not only promotes equity in scientific research but also enhances the global impact of NASA's work. By facilitating wider dissemination and utilization of research findings, SciX extends the benefits of NASA's research beyond the immediate scientific community, potentially inspiring and informing public interest and education in earth and space sciences.

SciX extends the science return from NASA SMD missions and research activities by making this science return, in the form of scholarly publications, discoverable in easily accessible and intuitive ways. In particular, SciX increases discoverability by:

- Normalizing affiliation information, see Section 2.1.5
- Incorporating and integrating existing NASA bibliographies, see Section 2.2.3
- Identifying and integrating links to NASA data products and software, Section 3.2.3
- Identifying and integrating links to OA versions of NASA publications, see Sections 1.4 and 3.2.3
- Identifying and integrating funding information, see Section 3.1.3

In support of these goals, the SciX team must also:

- Ensure SciX receives data of the highest quality for all core journals
- Implement workflows to mine information to augment publisher-provided data

- Establish communications and collaborations with NASA missions, groups and centers for exchanging data, linking their products, and providing access to their resources

1.4 Towards FAIR Open Access

SciX is a critical component of NASA's strategic initiative for open science (NASA SMD, 2019), which follows the Force11[3] FAIR principles (Wilkenson, et al 2016): Findable, Accessible, Interoperable, Reusable. NASA has long been a leader in open science (Kurtz, 2020). SciX inherits its open lineage from ADS: its mission promotes open science and its operations follow open science principles. The senior project scientist was an invited speaker at both the meetings that led to the founding of Force11[4] and was a member of its original board of advisors, when the FAIR principles were developed.

The **(F)indability of research documents is the core SciX mission**. Key to finding research documents in the SMD disciplines is and will be their presence in SciX. Table 1 shows the current levels of completeness of its literature holdings, a direct measure of the findability of these documents. Likewise, SciX provides, and will provide, substantial resources for the AIR principles as well.

While the primary objective of SciX is to enhance the findability of research documents, findability naturally leads to enhanced **(A)ccessibility of the literature and associated products to which SciX links**. These include data and software with source code (where available), with accessibility dependent on whether the product is open-source or the individual users' unique access capabilities. All the metadata in SciX are open to and directly accessible by anyone on earth. ADS pioneered co-indexing and linking articles to both the journal version and the green Open Access (OA) version on arXiv (Henneken et al 2007). SciX now has collaborations with the ESS Open Archive[5], EarthArXiv[6], and EGUsphere[7] to extend its green OA service. In addition, we collaborate with the publishers, Crossref, and CHORUS to recognize OA versions held by the publishers. Of 154K refereed articles currently in SciX that acknowledge NASA support, 122K (80%) have known OA versions to which SciX points so anyone can access them. In comparison, PubSpace[8] contains just over 32K records, not all of which have OA links available.

In addition to providing the full metadata and access, within the limits of copyright law, to the full text of scholarly documents, SciX provides links to many external sources of information. Through these links, researchers access related information in mission archives, observatory archives, data centers, software repositories, and other information providers. Currently, approximately 66% of refereed papers in SciX have at least one link to another resource, other than the version of the paper. This percentage is 73% for astronomy papers. SciX continues to expand its available links through the metadata enhancement described in Section 2.1.5, through collaborations described in Section 2.2.2, and through integration of curated bibliographies.

[3] https://force11.org/
[4] https://force11.org/info/about-force11/
[5] https://essopenarchive.org/
[6] https://eartharxiv.org
[7] https://www.egusphere.net/
[8] https://sti.nasa.gov/research-access/

Through our API, **SciX provides (I)nteroperability with other organizations**, as well as individual researchers. All the publicly available data holdings ingested by SciX are accessible through our API (Lockhart, 2021), which also underlies the User Interface. Many groups are taking advantage of this capability, including astronomy publishers, the NASA Astrophysics Archives, the Department of Energy's science.gov, the European Space Agency's (ESA) Science Division, and NOIRLab's information services[10], among others.

	Metadata	Fulltext	References
Astronomy	99%	98%	94%
Planetary Science	94%	88%	85%
Heliophysics	98%	96%	88%
GeoRef Priority	87%	50%	68%

Table 1: Current average completeness levels per discipline, based on core journals. These levels were determined for the period 2005-2023. The entry GeoRef Priority refers to the collection of GeoRef Priority Journals[9] indexed in SciX.

Finally, **SciX is geared toward promoting the (R)e-usability of data**. ADS pioneered links to mission and observatory archives, which SciX is expanding. SciX also indexes high-level data products and other data sets of community interest to facilitate the re-use of data**.**

2. Progress Since Spring 2020

We have achieved two overarching and noteworthy efforts. First, we have kept ADS operating 24/7/365, uninterrupted for this entire period despite the pandemic, while constantly updating its content and services. Second, we developed and began implementing a detailed plan to provide ADS-level services to all SMD science domains as NASA desires, including complete indices of the relevant refereed literature.

Because of its technical and scientific reliability and functionality, ADS is the trusted partner through which essentially every astronomer on earth interacts with the scientific literature and, increasingly, software and data**.** Planetary scientists, heliophysicists, and space physicists benefited from this continuity of service alongside astronomers and astrophysicists, especially as local resources closed during the pandemic. We experienced the same difficulties switching abruptly from an on-site project to a fully remote one that our suddenly-working-from-home colleagues did. However, we overcame the impediments so gracefully that our users did not notice any change in service. SciX continues to build upon this robust infrastructure, which is essential to the development of advanced capabilities. Section 2.1 describes the growth in SciX capabilities and versatility during this period.

When completed, SciX will be more than ADS for all scientists working in SMD disciplines; it will establish a synergistic nexus for cross-disciplinary research. SciX will enable a transformation in open science and a holistic scientific understanding of our world. While the growth in SciX capabilities and versatility is bringing that vision to fruition, section 2.2 describes the structural changes we have taken to support the expansion and welcome new users. An on-line bibliography of recent publications relevant to ADS and SciX is available at https://www.scixplorer.org/public-libraries/0PQTAukwTY65q-goPxQvYA.

[9] https://www.americangeosciences.org/information/georef/priority-journals
[10] https://noirlab.edu/science/library/publications/metrics

2.1 Continuously Updating Content and Capabilities

Since its inception as ADS and continuing throughout its expansion, SciX is growing its collections by ingesting relevant literature and enhancing the metadata describing its content. To improve user search capabilities and expose more of the scientific literature in a meaningful way to each discipline, SciX takes an active role in adapting new technologies, such as AI and ML, to the best practices in information science.

2.1.1 New disciplines, greater interdisciplinarity

The SciX Content Model (Henneken 2023), illustrated in Figure 1 and informed by three decades of ADS use, determines what content needs to be represented in the SciX holdings. This model for inclusion is unique among databases for its innovative, tiered curation architecture, which features a core collection consisting of all content from its target research disciplines, complemented by articles that are connected to it via citations through a two-level system. This three-tiered approach provides a comprehensive, interconnected set of articles relevant to and focused on the core collection, while still allowing related content to be found within the system. In the core collection, users can expect SciX to be complete, coverage- and citation-wise.

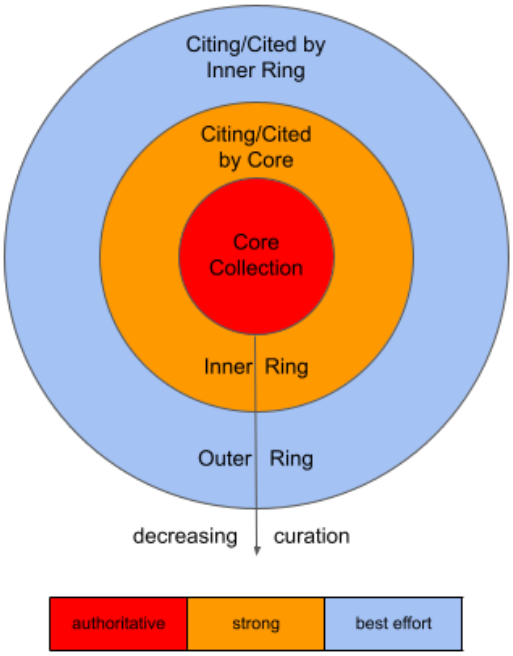

Figure 1. ADS's tiered curation model. The core collection represent disciplines where its curation is strongest and its coverage is authoritative. The surrounding tiers are connected to the core via the citation network.

Most of our curation efforts for this collection go into maintaining a high level of accuracy, quality, and completeness, ranging from the main refereed literature to conference proceedings, theses, and gray literature[11]. Based on the maturity of ADS, the AP core collection is well-established and fully authoritative. The coverage for HP and PS is 98% and 94% respectively. The ongoing expansion into ES requires that this discipline be fully represented in the core collection. Our census of all ES journals has identified about 1,200 journals that represent the most influential venues of scholarly publication in the field. Feedback from the ES project scientist will be instrumental in fine tuning this core collection.

	Current size (M)	Size in 2020 (M)	Growth since 2020
Astronomy	2.852	2.545	12%
non-Astronomy	17.488	11.739	49%
Full-Text	7.514	5.606	34%
Refereed	14.562	9.755	49%
Citations	190.18	129.3	47%

Table 2: Growth of SciX holdings since the last Senior Review (2020). While the increase in Astronomy records is due to the new content being published, growth in the non-astronomy portion of the database reflects the expansion of content due to the SciX effort. Further growth in full text holding for ES content will greatly increase the overall number in the next couple of years.

[11] Materials produced by organizations outside of the traditional commercial or academic publishing and distribution channels

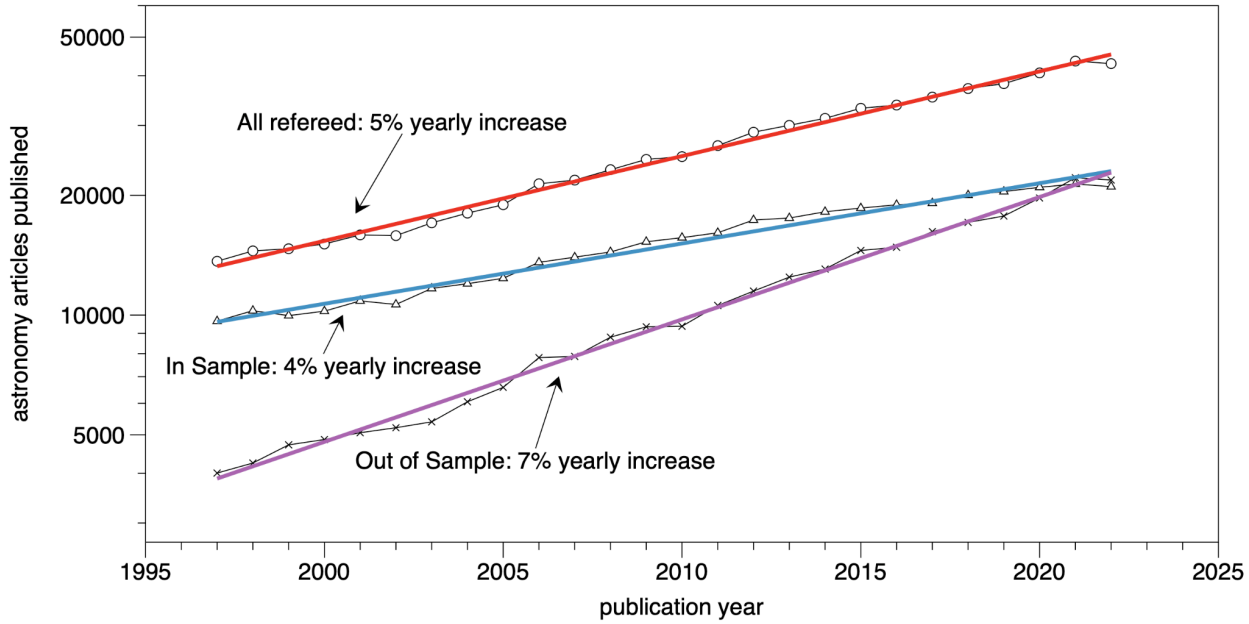

Figure 2: Growth of complexity as the increase of discipline-specific content in multidisciplinary journals in astronomy. The middle line shows a 4% yearly growth in the number of articles in the major Astrophysics journals. The bottom line represents the growth of articles in disciplines such as high-energy physics and geophysics that cite astronomy articles; it is increasing at 7% per year. The top line represents the weighted sum for both collections, which is increasing at 5% per year. SciX will be instrumental in enabling the discoverability of this growing content.

The long-term growth average for all science growth is 4% per year, or 12% for 3 years (de Solla Price 1961). The observed growth of the astrophysics collection in ADS reflects this trend. As shown in Table 2, larger growth of our database outside astronomy reflects the expansion of SciX scope. In addition to this, Kurtz & Henneken (2018) showed growth of interdisciplinarity over time. In Figure 2, we have extended this plot up to current time, to emphasize that these trends have continued. The middle line shows a 4% yearly growth in the number of articles in the major AP journals, which is consistent with Table 2 and the findings of de Solla Price (1961). The bottom line represents the growth of astronomy articles in disciplines such as high-energy physics and geophysics. **This direct measure of interdisciplinarity shows a yearly increase of 7%. Promoting such cross disciplinary research is an important reason for the creation of SciX** (NASA SMD, 2019).

2.1.2 Database Completeness

ADS has long been the gold standard for database completeness in the astronomical and astrophysical literature. Every modern refereed research article in astrophysics, whether in a main astronomy journal or an individual article in an obscure physics journal, is included. ADS also achieved a very high degree of completeness for the non-refereed literature and the historical literature. A large fraction of these articles, particularly of the modern refereed literature, is available for full text search and text mining, not simply for metadata (abstract, title, author) search. SciX is currently on schedule to achieve an equivalent completeness in PS, HP, and ES by 2029, as proposed in SciX22.

Achieving complete records for AP, PS, HP, and (especially) ES is both a very large task and an important SciX goal. Completeness can be assessed within specific document classes: recent or historic, refereed or non-refereed. Degree of coverage is also a consideration. Minimal coverage means having basic metadata: journal, volume, page, author, name, and title. Increasing degrees require having the abstract, then having the reference list, and further having the full text. We measure completeness in a discipline by the fraction of papers in the reference lists of papers in that discipline's core journals for which SciX has entries in our database. For recent astrophysics refereed papers, this measure is above 94%. Table 1 shows completeness measures by discipline.

Curating and maintaining the expanding SciX collections so that users can have confidence that they will find what they need is a difficult and ongoing task. While some large publishers routinely produce millions of documents, others produce an occasional one. Achieving a strong fraction of completeness for the most popular and larger article sources (e.g. Elsevier, AGU) is straightforward, but it is much more challenging for single conferences or small journals.

The success of ADS was built on interoperability with external collaborators that provide essential metadata to enrich the database. Extending these collaborations, SciX will achieve technical parity with astrophysics for the major NASA disciplines by the end of this funding cycle (completeness and curation of the BPS collection is discussed in section 2.2). As an example, we have negotiated an agreement with the American Geosciences Institute, purveyors of the GeoRef database. If approved, this arrangement will provide us with a large fraction of the essential metadata for the non-refereed ES literature. While GeoRef contains abstracts for most of its holdings, it does not have references nor the full text. We will continue our collection efforts to obtain this information, which enables users to have more powerful and productive interactions with the collection.

2.1.3 New connections, increased accessibility

An important feature of SciX is linking from journal articles to the archival data used in the research, which predates the federal government open science policy[12]. Implementing White House and NASA open science mandates means that an increasing number of data sets will be openly available to the public. Improving access to these data is a key SciX responsibility. In addition, SciX will also provide the means to develop metrics for measuring degrees of compliance.

This emerging, increasingly complex research environment is very much a reality. Current SciX holdings indeed show that new publishing modalities (such as software and data sets) add to the growing complexity. For cited data products, SciX always provides a link from our record of the citing publication to the data cited. In cases that meet criteria established by the SciX curation team, we also create a record for the cited data product and assign the citation. For example, we will assign citations for high-level, curated data products and reused data sets whenever a DOI has been assigned to them and sufficient, descriptive metadata is available. SciX currently holds close to 20k records for software products that were cited in the scholarly literature, and over 20k records for PDS data sets and VizieR catalogs[13], with over 10k total citations.

[12] https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-access-Memo.pdf
[13] https://vizier.unistra.fr/

SciX mines the full text of documents to discover actionable mentions of the data sources used in the paper; these are mentions that can result in data links or citations. Through collaborations with the astronomy archives and data centers, ADS has incorporated links between papers and data for years, for example 68% of all *Astrophysical Journal* papers published in 2023 have data links in SciX. Data mining, however, has been especially fruitful for American Geophysical Union (AGU) publications, where there is a formal mechanism for these data references[14]. The *Journal of Geophysical Research* recently began requiring data availability statements in their articles; consequently, 78% of their articles published in 2023 have data links in SciX. Overall, a test of mining data links from data availability statements in the full text of about 50 journals resulted in 80k data links. In addition, by analyzing DOIs[15] in reference data, we found 125k citations for data products. SciX now exposes all these links to its users.

We anticipate that the practice of assigning DOIs to data sets and referencing them in a method similar to traditional literature will be much more widely adopted and further improve this capability. For instance, in astrophysics, the International Virtual Observatory Alliance (IVOA) has recommendations for publishing curated data links. Because connecting the literature with the data sets and software is an important component of the open science ecosystem, sections 2.2.3 and 3.2.3 describes growing collaborations between SciX, archives, and data centers to increase the exposure of their products and the overall interoperability of such resources as well as expanded agreements with journals to allow more flexible use of their content.

2.1.4 Artificial Intelligence/Machine Learning (AI/ML)

From the initial ADS announcement (Kurtz et al 1993) to present-day SciX, we have implemented, and sometimes developed, cutting-edge digital library techniques for use directly by scientists and for use in our curation and enrichment processes. In its early stages, ADS featured natural language search using a relevance vector space similar to the vector embeddings currently popular for Retrieval-Augmented Generation (RAG) with large language models, such as GPT-4. Later implementations made use of more traditional information retrieval techniques based on open-source initiatives (Accomazzi 2024).

With the rapid changes and increasing investment in AI/ML technology, predicting the impact of these opportunities over the SciX user experience is difficult. However, we will remain engaged with these developments while seeking solutions to serious legal (e.g. copyright) and technical (e.g. hallucinations) challenges to their use. We will offer services combining language models and knowledge graphs as soon as they are feasible, legal, and trustworthy. The rest of this section describes the AI/ML activities we have pursued in the past four years, while section 3.5 discusses our plans for the future.

Recognizing the importance of the attention mechanism (Vaswani et al 2017), we proposed in the SR20 to develop a language model to support a variety of Natural Language Processing (NLP) tasks. We have hired a computer scientist for this work, and have created and publicly released an astrophysics-centric language model, astroBERT[16] (Grezes et al 2021), which has been used in

[14] Some other journals currently have this as well, including from AAS, Elsevier, Springer-Nature,and IoP
[15] A DOI has metadata bound to it that can be retrieved automatically from the repository where the DOI was registered
[16] https://huggingface.co/adsabs/astroBERT

several external projects (Timmer et al 2023). Use of **astroBERT and similar models improves the classification literature by discipline within SciX and enhances the available metadata** (see section 2.1.5). SciX will expand its use of these tools to provide researchers with more powerful and nuanced search capabilities.

During the past few years the use of language models has exploded. Large language models (LLMs) are becoming an important part of the world economy and have become central to the development plans for several of the world's largest industrial corporations. Although the development costs of such efforts are far beyond the reach of small projects, SciX is proactively responding to these changes as demonstrated by the four projects showcased below. The rapid growth in this technology makes predicting applications in five years imprudent. However, SciX will remain engaged with emerging developments and continue collaborating with the community in evaluating and implementing new capabilities based on these technologies (see section 3.5).

First, we have developed a system, the SciXBrain (Blanco-Cuaresma et al 2023), which runs on our local servers and can be used with external, publicly available LLMs, such as those from the Meta Corporation (Touvron et al 2023). This design allows us to test different LLMs to see which performs best for a specified task. We have used the SciXBrain system to test different language models and training techniques. Currently, we are using models behind the SciXBrain as part of our Named Entity Recognition (NER) metadata enrichment processes, which has identified named planetary features and will identify missions and instruments. As described in section 3.2.1, these models will enable researchers to execute more sophisticated searches using discipline-specific taxonomies.

Second, we are (re)negotiating our agreements with publishers/copyright holders to permit our use of the full-text data for AI/ML projects. SciX will maintain the crucial trust and beneficial relationships ADS established with scientific publishers over decades. What the copyright holders allow us to do with their content varies widely, especially regarding displaying the results of a language model, or publishing the model. Given the mix of licenses being used by hybrid journals, we are modifying our data models and ingestion procedures to honor our agreements on a per article basis. Identifying the appropriate use of an individual article allows SciX to give that article maximum exposure consistent with the relevant agreement, which in turn, maximizes the return on the author's, and funding agency's, investment. Managing permissions this way is also essential to offering effective AI-based discovery tools that respect intellectual property rights. To avoid copyright issues, the generative use of SciXBrain relies exclusively on liberally licensed content.

Third, recognizing the broad interest in LLM development, we are collaborating with several groups, both NASA affiliated and non-NASA, in efforts involving the development and use of these models. Because SciX already has the text relevant to NASA SMD research, we are becoming the central hub for these developments. We collaborated with NASA and IBM to build a science aware model,[17] which was released in December 2023. During the same period, SciX began working with the UniverseTBD collaboration[18], which recently released the first

[17] https://huggingface.co/nasa-impact/nasa-smd-ibm-v0.1
[18] https://universetbd.org

fine-tuned version of the popular LLaMA-2 model, named AstroLLaMA[19]. AstroLLaMA has outperformed closed-source models on selected astronomical tasks (Nguyen et al 2023). We are now supporting further fine-tuning AstroLLaMA in collaboration with UniverseTBD and colleagues at Oak Ridge National Laboratory (ORNL) who have access to significant super-computing resources, which are required for training models with greater than 1 billion parameters. This will provide us with more powerful discipline-focused LLMs to support a variety of AI/ML downstream tasks.

Fourth, we have organized workshops on information extraction from scientific papers (WIESP[20]) at the 2022[21] and 2023[22] International Joint Conference on Natural Language Processing-Asian Chapter of the Association for Computational Linguistics Meeting.[23] These workshops provided a venue to engage with the computer science community on a variety of NLP tasks beyond NER and related to information extraction and classification, metadata normalization and enrichment, language models and Knowledge Graphs (KGs). Participation in the workshops has yielded collaborative efforts with ORNL and the Observatory of Paris (Alkan et al 2022). Ultimately, these efforts will provide researchers with even more advanced tools for exploring the scientific literature.

2.1.5 Metadata Enrichment

SciX is harnessing recent technological developments, such as LLMs and graph neural networks, to present more nuanced information to researchers and to increase the efficiency of their discovery processes. SR20 discussed some benefits that would be achieved using these substantially more powerful tools for extracting information from text; SciX22 proposed the first projects. One pilot project is now a functional feature available in the new SciX User Interface (UI): the planetary feature names keywords (Shapurian et al 2023). A second project will implement classification of astronomy papers using the Unified Astronomy Thesaurus (UAT) keywords (Accomazzi et al 2022).

Using a supervised learning approach, SciX built a concordance between solar system features listed in the IAU/USGS Gazetteer of Planetary Place Names[24] and the articles that discuss them. We released an initial version of this concordance with an integrated user interface and links to the relevant sections of the USGS database, along with the beta SciX release at the December 2023 AGU meeting; Figure 3 illustrates this new discovery aid, which garnered appreciative comments there. This is a novel, enabling infrastructure for the planetary science community. It may be compared with the creation (by hand curation, half a century ago) of the Bibliographic Star Index (Cayrel et al 1974), which today remains at the core of the CDS/SIMBAD[25] database (Wenger et al 2000).

[19] https://huggingface.co/universeTBD/astrollama
[20] https://ui.adsabs.harvard.edu/WIESP
[21] https://aclanthology.org/volumes/2022.wiesp-1/
[22] https://ui.adsabs.harvard.edu/WIESP/2023/
[23] http://www.ijcnlp-aacl2023.org/
[24] https://planetarynames.wr.usgs.gov/
[25] https://simbad.u-strasbg.fr/simbad/

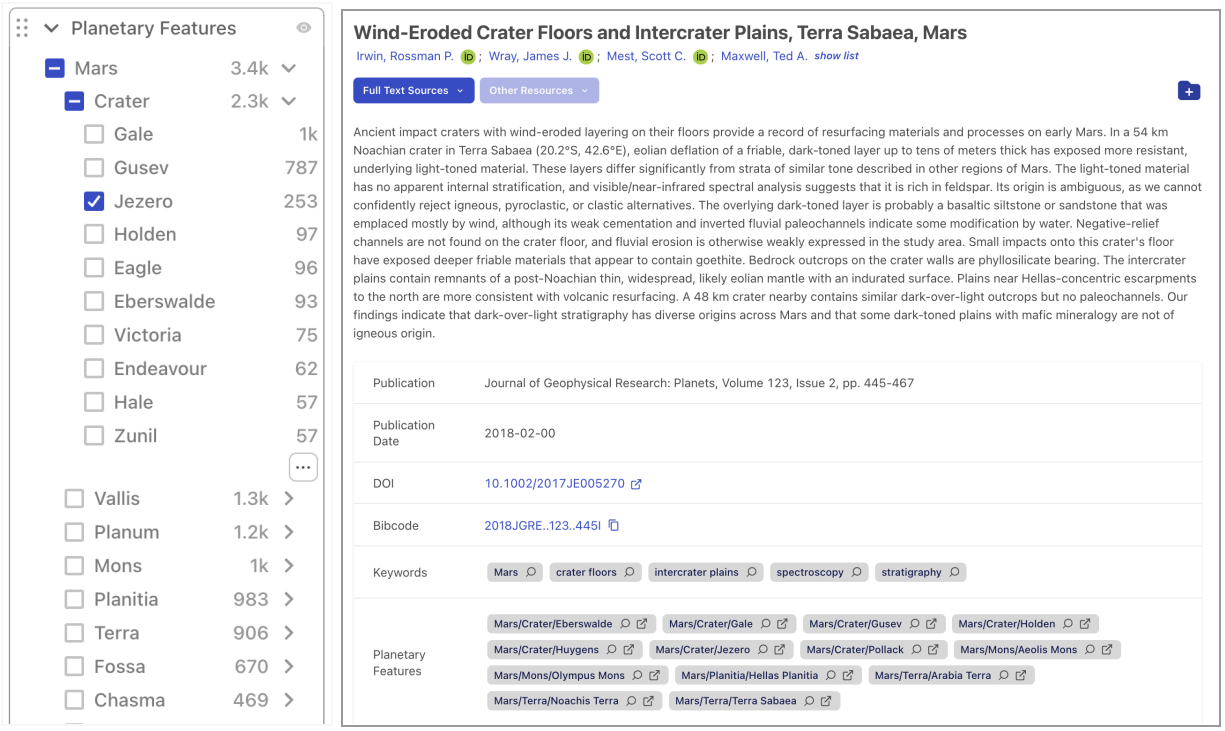

Figure 3: Implementation of the IAU/USGS Gazetteer of Planetary Place Names indexing in SciX. Left: the Planetary Names facet allows users to filter the literature based on individual feature names, types, or solar system bodies where they appear. Right: the abstract pages for papers found to contain named planetary features display the corresponding list, with links to the feature descriptions on the corresponding PDS node.

Because many planetary place names are highly ambiguous (e.g. Saturn, Cassini), creating the concordance was substantially more complex than simple text mining. We iteratively developed a learning set, used it in the ML training of a custom expert system making use of NER-optimized versions of astroBERT and SciXBrain. This first NER application benefits primarily PS, but the technologies we developed will now be used more generally and the development of similar processes for other disciplines will proceed more easily.

As noted in SR20, the technologies enabling NER (Grezes et al 2022) can be applied to a wide variety of people, places, concepts, and things (see section 3). By densely annotating documents (Allen 2023), we are creating training data to recognize organizations, missions, instruments, wavebands, grants, etc. automatically. With the permission of the copyright holders, we are sharing these annotated data sets with other organizations, primarily NASA data centers, which also want to do their own NER analysis. By making our data sets freely available, we are both **modeling the open science approaches NASA is promoting and enabling their re-use to increase the scientific return on our development work**.

For NER to function effectively, a system must be in place to recognize when different representations of the same entity are the same (e.g. GSFC, NASA Goddard, and Goddard Space Flight Center). We have an ongoing program to achieve this for organizational affiliations (Grant & Templeton 2020; Templeton & Grant 2021). Our normalized affiliations allow users to search on organizational name variants and Research Organization Registry (ROR)[26] identifications.

[26] https://ror.org/

They also include parent-child relationships so that departments within a university or centers within a facility can be distinguished. For example, searching the publisher-provided affiliation information for “NASA Goddard” or “GSFC” only provides a partial list of results[27], while searching the normalized institution field for “GSFC” finds an additional 8% of relevant records where the affiliation was listed in a different form[28]. It also finds records with an associated affiliation such as the Center for Research and Exploration in Space Science and Technology (CRESST), which also has its own entry in our normalized list[29].

Clear identification of institutional affiliations also provides the user with tools to disambiguate common author names, which will become increasingly important as the core collection grows. In addition, such identifications allow the grouping of papers by the institutions at which they were created, highlighting the research output of individual centers. Our system is sufficiently developed to examine the institutional changes in astrophysics over the last quarter of a century (Kurtz et al. 2024). For refereed publications in AP, PS, and HP, affiliations are currently matched to canonical affiliations at least at the 95% level, while in physics this accuracy is at 90%. Without having done any additional work, affiliations in ES are currently matched at the 70% level and, with a minimal amount of work, we expect to increase this to 80%.

In addition to our multiple NER programs, we, with a collaboration of researchers from the Harvard School of Engineering and Applied Science, developed techniques to assign keywords from a standard set to arbitrary documents automatically. We began with the UAT (Accomazzi et al 2022), which has become standard in astrophysics, and which is being expanded to cover heliophysics and planetary sciences via a collaboration among NASA HPD, SciX, and the UAT steering committee. The AI/ML assigned keywords will provide a more thorough and more consistent set of search terms than the handful of arbitrary author-assigned terms. Searching using these tools provides researchers more focused results, giving greater exposure to the most relevant papers. It also provides a way for them to leverage the UAT structure to easily find papers discussing similar, narrower, or broader topics. As discussed in section 3.2.1, SciX plans to add additional keyword systems in support of other disciplines building on the technologies developed for UAT classification.

2.2 Transitioning to SciX

SciX22 promised much more than an upsized ADS or an ADS with earth science literature. It envisioned discipline-aware portals providing access to a centralized discipline-agnostic digital library. **When complete, scientists from each SMD discipline will interact with a SciX interface that is responsive to the culture of their community and that applies the appropriate vocabularies, taxonomies, and conventions**. To accomplish this, SciX has redesigned its user experiences, expanded its external collaborations, and begun a thorough internal reorganization.

SciX will accommodate, support, and encourage interdisciplinary research, which will benefit all the SMD disciplines, consistent with historical trends (Kurtz & Henneken 2018). The trend towards interdependence of the disciplines will likely accelerate (van Noorde, 2015; Kurtz &

[27] https://www.scixplorer.org/search?q=aff%3A%22NASA+goddard%22+or+aff%3AGSFC
[28] https://www.scixplorer.org/search?q=inst%3AGSFC
[29] https://www.scixplorer.org/search?q=inst%3A%22GSFC%2FCRESST%22

Accomazzi 2019). Even considering the importance of our short and intermediate term goals of completing the literature database, linking to external data sources, and exploiting the new technologies, perhaps **the most important long-term goal of SciX is to enable cross and interdisciplinary communication, collaboration, and research.**

At this early stage in the SciX transition, different disciplines are supported differently. Astrophysicists are still the primary users, followed by heliophysicists and planetary scientists. The ES community is now becoming aware of SciX following its debut at the December 2023 AGU meeting, and we anticipate rapid growth in the number of users interested in ES-specific content as we grow our ES collection and continue to reach out to ES communities. SciX is not currently addressing the needs of BPS at the same level as the other four disciplines. However, SciX **already contains most of the physics literature behind BPS research and will ingest additional papers authored by NASA BPS employees or funded by the NASA BPS division.** Any more intensive effort would require direction from NASA.

2.2.1 Re-designed User Interface/User Experience (UI/UX)

The ADS user interface (UI), originally developed in 2015, was obsolete by the end of that decade, based on out-of-date technology, and importantly, not up to modern accessibility standards. We proposed replacing it in SR20 to improve its overall usability, accessibility and maintainability using state-of-the-art technologies. Furthermore, the multidisciplinary SciX portal must support queries from scientists unfamiliar with ADS and whose vocabulary differs from traditional astronomy terms.

We hired an additional UI/UX developer to achieve the critical mass necessary to complete this project. The new system allows for substantial customization, by both SciX staff and users, while meeting the Web Content Accessibility Guidelines[30] (WCAG 2.1 level AA). This flexible design permits the explicit instantiation of specialized user interfaces for each SMD discipline and can be extended as SciX adds new features and capabilities. We announced the beta release of the new UI at the December 2023 AGU, where it was well received by testers and potential users[31]. The new UI targets new SciX users, but, as its advanced features become more robust, we expect its adoption by astronomers, allowing us to unify all user communities on a single multi-disciplinary platform within five years.

As part of the redesign, we rebranded the look of our interface to reflect our commitment to serving the full range of SMD disciplines and cross-disciplinary discovery. The SciX name, web addresses, logo, help features, and promotional materials are welcoming to all disciplines represented in the expanded digital library. Rethinking these less technical aspects of the system addresses concerns that users without previous ADS experience would have a steep learning curve. The SciX design intends to get the user interacting with the literature and linked resources quickly, providing maximum relevant scientific content for their invested time through selection of discipline-specific collections that will be curated as part of this expansion effort. A satisfied user is both a return user and an ambassador to their community for the service.

[30] https://www.w3.org/WAI/standards-guidelines/wcag/
[31] https://scixplorer.org

With the expected increase in user base from the newly added disciplines, we expect a significant increase in usage of our search API and website infrastructure. With increased visibility, we also must ensure that our infrastructure and security are hardened appropriately.

2.2.2 Maintaining and Improving Computing Infrastructure

We continue to host our front-end and microservices in the cloud, using Amazon Web Services, which has proven to be stable and reliable for our users. Small updates have been made to our microservices as needed to add new collaborative functionality for our updated front-end interface, such as the ability to add and maintain private notes to entries within SciX libraries[32]. Overall, our cloud-based infrastructure and microservices are stable and working well.

Most back-office development efforts have been spent updating and evolving our data pipelines, hosted in our back-office, on-premises servers. Much of this effort has been focused on incrementally replacing legacy software components, an ongoing project for the team. The team replaced numerous metadata parsers, many of which were still using legacy PERL code. These parsers are in testing by the curation team and will eventually be a component of a new data ingest architecture, currently in development. This new ingest architecture will replace a combination of legacy PERL scripts that operate directly on files on disk and the current more modern back-office architecture that fundamentally relies on the legacy code. The new architecture incorporates Apache Kafka[33], an event streaming system that allows us to perform event-driven incremental updates on our data instead of relying on larger scheduled jobs.

Other legacy software replacement components are either in testing or in production. Reference parsing for arXiv papers relies on a combination of parsing TeX and PDF documents. Our TeX distribution has been updated to match arXiv's and a new extraction method, using GROBID (2008), for PDF documents is in testing. Our new reference resolving architecture, consisting of a microservice and pipeline, is nearly complete and is undergoing testing. Our new document matching system (Koch et al 2022), which cross-correlates pre-prints with their corresponding published versions, has been in production for a year. Finally, our new scan explorer interface and service, which allows users to page through ADS-scanned papers and is the last remaining user-facing legacy component, is in testing.

In addition to on-premises software updates, our on-premises servers themselves are periodically updated. New redundant NetApp network storage units were purchased in early 2023 to meet the demands of a growing volume of data being added to SciX. The new data ingest architecture will be migrated to a recently installed 8-node Dell computing cluster in 2024. We will have the WEKA file system[34] (a parallel high performance file system) and Kubernetes (a popular container management system) installed on this cluster, to enable high I/O operations and a redundant computing environment. In addition, we also currently host an on-premises GPU server (2 x NVIDIA V100) that allows us to do small model training and run inference on small LLMs (less than 13 billion parameters). We currently have one pipeline that relies on these hosted LLMs, our planetary names feature described in section 2.1.5, in production, and have the

[32] SciX libraries are an extension of ADS libraries, user-managed collections of records that can be kept private, shared with collaborators, or made publicly accessible.
[33] https://kafka.apache.org/
[34] https://www.weka.io/

SciXBrain interface described in section 2.1.4 that allows us to experiment with new open source LLMs as they become available.

2.2.3 Expanding Collaborations

Scholarly documents, especially refereed journal articles in English, are the *lingua franca* of science. As the digital gateway to scientific scholarship, SciX is the center of a complex web of authors, readers, publishers, societies, data centers, libraries, agencies, and others. To serve the global research enterprise, SciX uses not just documents, but substantial additional information from multiple sources. Our effectiveness lies in active collaboration with many outside people and organizations.

As ADS expands into SciX, the number and nature of our collaborations is also expanding. We have hundreds of agreements at widely different levels of formality, from contracts to handshakes. Considering our new mission, every agreement must be revisited, and many more need to be established. Most of our agreements, however, fall into two categories: publishers and data archives.

Publishers provide the literature, which forms the core of SciX services. To help us navigate our evolving relationships with publishers, we hired a part-time, former publishing executive as a consultant. In addition to establishing relationships with new publishers covering new disciplines, especially ES, we are renegotiating existing agreements to access a larger set of journals from some publishers and to enable the use of copyrighted material with emerging technologies. Section 3.5 describes our use of AI/ML to provide advanced capabilities while respecting our contractual obligations.

SciX enhances the literature by indexing and linking software and data products. Our newly hired discipline project scientists will increase the systematic interoperability pioneered by ADS in astronomy that **makes data sets more discoverable, encourages their reuse, and increases the science resulting from their creation**. The project scientists are establishing science-based priorities and implementation plans in collaboration with the archives and data centers. We are also collaborating with the NASA Science Discovery Engine[35] to achieve these goals without duplicating efforts.

One ongoing collaboration with NASA archives and data centers is the **incorporation of NASA-curated bibliographies into SciX,** cementing it as an authoritative resource for NASA publications, including otherwise less-accessible, non-refereed materials. In addition, the curation of these specialized bibliographies within SciX exposes their contents to a broader audience. Figure 4 describes the ongoing communications between the SciX team and various NASA stakeholders.

[35] https://sciencediscoveryengine.nasa.gov/app/nasa-sba-smd/#/home

Bibliography Source	Status
NASA Ames Space Science & Astrobiology (ARC/SS)	Complete/Maintenance
NASA PubSpace (STI/NTRS)	Ingest
NASA Socioeconomic Data and Applications Center (SEDAC)	Curation
NASA Goddard Earth Sciences Data & Info. Services Center (GES DISC)	Curation
NASA Astromaterials Data System (Astromat)	Planning/Prep
National Snow & Ice Data Center (NSIDC)	Planning/Prep
ORNL Distributed Active Archive Center (ORNL DAAC)	Planning/Prep
NASA Goddard Sciences and Exploration Directorate (SED)	Communications

Figure 4. Representation of the steps necessary for SciX to incorporate specialized bibliographies along with the status of several NASA bibliographies in progress.

The SciX ingest of the larger NASA PubSpace collection[36] is in progress. This collection is of particular importance because it is the designated public-access repository for refereed journal articles funded by NASA. It is now administered by NASA Scientific and Technical Information Services (STI) but was formerly held by the National Institutes of Health's (NIH) PubMed Central (PMC). **PubSpace is now integrated as a bibliographic group in SciX and can be easily accessed using a simple query[37];** currently, it holds 24,714 records (92% of the collection in NTRS).

With NASA PubSpace accessible through SciX, users can use sophisticated search techniques to further explore this collection; all publications within this collection with at least one link to data products can be found by adding the filter **property:data**[38] and, similarly, all publications with at least one link to an Open Access version can be found via the filter **property:openaccess**[39]. These examples illustrate how simple SciX queries can find measures for the success of NASA's Open Source Science Initiative (OSSI) program[40], as expressed in the scholarly output of NASA SMD. In addition, the SciX visualization capabilities can explore the contexts in which NASA data products have been made available via the publications to which they have links.

Similar powerful assessments will be possible through collaborations between other NASA stakeholders and SciX. Projects are underway to integrate bibliographies from the Socioeconomic Data and Applications Center (SEDAC), Goddard Earth Sciences Data and Information Services Center (GES DISC), Goddard Sciences and Exploration Directorate (SED), Astromaterials Data System (Astromat), the National Snow & Ice Data Center (NSIDC), and Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC). The SciX

[36] https://sti.nasa.gov/research-access/
[37] https://scixplorer.org/search?q=bibgroup%3A%22NASA+Pubspace%22
[38] https://scixplorer.org/search?q=bibgroup%3A%22NASA+Pubspace%22+property%3Adata
[39] https://scixplorer.org/search?q=bibgroup%3A%22NASA+Pubspace%22+property%3Aopenaccess
[40] https://science.nasa.gov/researchers/open-science/

project scientists will continue building relationships with these and other NASA communities to expose their research products to as wide a range of interested scientists as possible.

The SMD Science Discovery Engine is an initiative to improve data discovery across all SMD disciplines by creating a central catalog of all NASA research data sets. While NASA SciX will focus its efforts on the literature, there are common threads between the two projects that we intend to exploit. On a technological level, both projects plan to use NLP tools to improve the discovery process of their collections. On the disciplinary level, both projects will likely use similar knowledge bases to support semantic search and ranking of results. The development of annotated data sets and language models described section 3.5 will also be an extremely useful development for the Science Discovery Engine team.

2.2.4 Evolving Organizational Structure

In order to accomplish the SciX vision, our current operations also need to undergo a significant expansion and enhancement. With a greater amount of the scholarly literature to cover, all the activities associated with data ingestion, transformation, normalization, enrichment, access, and usage will need to scale up accordingly. The increase in interdisciplinary content also requires the expansion of staff experts in the new research areas being covered.

SciX will be more than twice the size of ADS in terms of use and impact, which requires a considerably larger and more intellectually diverse team than ADS needed. Accomplishing this re-alignment demands substantial changes to hiring and management practices. The SciX staff is expected to be twice the size in terms of personnel and not all the senior staff will be astronomers.

To achieve its vision, SciX22 laid out a new project structure, described as the *pentapus*[41] shown in Figure 5, which is taken from that proposal. Like the familiar octopus, a pentapus has six brains, which think independently. Similarly, SciX will have a central body, which manages the system and its data, and a set of scientific tentacles, which keep the system grounded in the needs of scientists.

We are on schedule to complete the necessary reorganization within Year 1 of this funding cycle. At that time, we anticipate the new organization and management structure described in SciX22 and illustrated in Figure 5 will be fully realized. As we hire additional people to build and staff SciX, we are adjusting roles and responsibilities appropriately. We implement each change deliberately, conscientious that SciX must operate without interruption of service or continuous updates 24/7/365.

To address the expanded SciX disciplinary focus, each tentacle contains an independent scientist. Active researchers on the team ensure SciX maintains its focus on serving scientists in the disciplines it covers. Furthermore, each tentacle responds to its own disciplinary culture while also being part of a multidisciplinary effort. We hired two associate project scientists to fill these functions, one each in astrophysics and planetary science, and are currently advertising an open position in earth science. We plan to complete the team by adding a heliophysicist during Year 1.

[41] A pentapus has five tentacles rather than the eight of an octopus

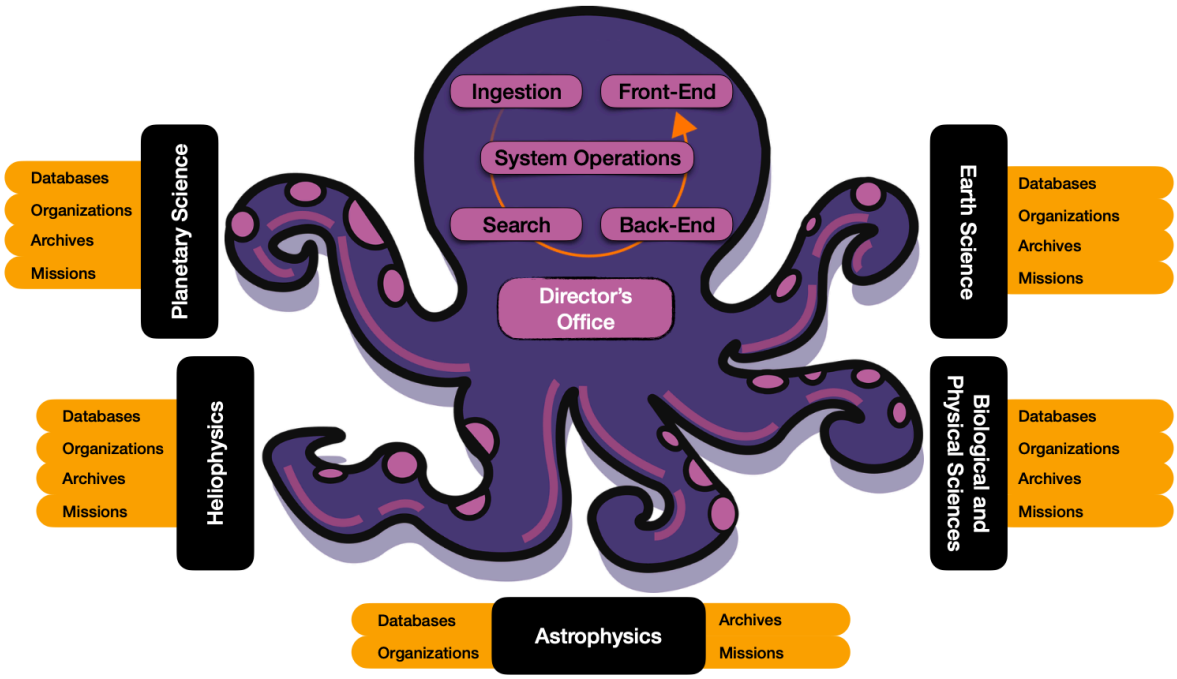

Figure 5. Structure of the enlarged project: a discipline-agnostic core group consisting of the project leadership and five operational teams will interact with disciplinary teams staffed by subject matter experts who will interface with the Databases, Organizations, Archives, and Missions specific in each research field.

SciX now uses consultants for tasks that are either short term or require substantial, part-time expertise. For example, SciX contracted with consultants to build a new interface for our historical, scanned material and is using a consultant to negotiate new agreements with publishers.

2.2.5 Expanding to New Communities

The community engagement plans for SciX will broaden significantly and we have hired a community engagement coordinator to facilitate expansion into new communities. While we continue our engagement with the astronomy community through conferences, social media, and events like hackathons, we will also **develop tailored strategies to connect to the remaining NASA SMD communities**. We envision developing training materials (presentations, infographics, demo videos on YouTube, tutorials for user objectives, Jupyter notebooks) for researchers, students, and librarians from our expansion communities; increased social media presence; user studies conducted at a regular cadence including UI/UX testing; more frequent and focused outreach including conference presentations and workshops; collaborations with academic and research institutions; collaborations with NASA initiatives such as TOPS and SCoPE; and usage log analysis informing continuous enhancement of the platform.

Central to our strategy for engaging these new disciplines is the implementation of the **SciX Ambassadors program**. To support this program, we have partnered with leadership in the

NASA SCoPE[42] and Infiniscope[43] projects to engage with large groups of NASA Subject Matter Experts, Science Activation Teams, and science educators to advertise the platform and recruit ambassadors. Our ambassadors will be selected to represent the main SciX user communities including scientists, students, librarians, and educators across each discipline. We will support a cohort of 20 ambassadors (rotating every two years) to receive training through a 2.5-day workshop at the Harvard & Smithsonian Center for Astrophysics (CfA) to utilize and promote the SciX platform effectively within their communities. A dedicated page on the SciX website will showcase the ambassadors, each of whom will commit to conducting a minimum of four SciX training sessions or presentations annually for their communities, participating as beta testers for new features, and contributing to our user studies. This initiative strategically accelerates SciX's expansion into diverse scientific disciplines with a focus on interdisciplinarity, aligning closely with NASA's TOPS mission to cultivate an inclusive culture of open science rapidly.

In addition to direct feedback from individuals, we continue to analyze our logs to measure the usage and adoption of the system by different segments of the research community and content-defined subgroups. Usage analysis will also guide the design and re-design of user facing features.

To ensure that feedback from our expanded user community is effectively captured and addressed, we are establishing an advisory group focused on guiding NASA SciX's expansion into communities beyond astronomy. This group will mirror the structure of the ADS Users Group[44] (ADSUG), providing vital input and recommendations for the SciX operations overall. The group will play a crucial role in representing our expansion communities through contributing advice on content management, community engagement, and prioritizing tasks.

3. Technical Development Plan

This section describes how we plan to develop the SciX system such that it achieves the vision outlined in the previous section. The goals guiding the development are: **achieving technical parity for the four main disciplines at the end of the five-year plan; making continuous improvements to the system; and keeping the system current in our extended holdings and available to our expanding user base 24/7/365.**

The previous section highlights the solid progress SciX has made since the last Senior Review in expanding our services to all SMD disciplines. We will be building upon the developing infrastructure and curation initiatives so that we can achieve the desired technical parity and serve each discipline fully.

Technical parity means that all SciX disciplines will have literature databases equal, or superior, to the best in the world, as astrophysics does now. It means that SciX will connect papers and other documents in all four disciplines directly to the data and software used in writing them at several NASA data centers and other FAIR repositories. It means there will be substantial value-added enrichment and curation, some of which is discipline specific.

[42] https://science.nasa.gov/sciact-team/smd-community-of-practice-for-education/
[43] https://infiniscope.org/
[44] https://ui.adsabs.harvard.edu/about/adsug/adsug

Thanks to a solid technical foundation built on top of cloud services and scalable infrastructure, **SciX will be continuously available and up-to-date, even as the amount of ingested data, the number of users, and the activities associated with ingest and curation of each paper more than doubles**.

In addition to expanding the collections, continuous improvements of SciX will come from exploiting new AI technologies to develop new services; from improving and expanding our graph-based visualizations; from developing and improving enrichment techniques, such as NER; and from developing user interfaces to allow our users to understand and make full use of the new features and capabilities.

In the remainder of this section, we will outline the proposed technical development over the next five years. This is organized around several axes, as defined in the goals listed in section 1.2, and collected here in the broad categories of data ingestion, enrichment, discovery, and management. We will conclude by describing the R&D efforts required to support much of these initiatives.

3.1 Data Ingestion

As the SciX corpus grows, including new sources of data, users will no longer exclusively rely on simple metadata searching to find papers of interest. New methodologies will be required to track, ingest, and curate these additional data properly. Much of the team's effort here will be on building workflows and products to automate routine activities that enable relevant discovery.

3.1.1 Data Ingestion Pipelines

The SciX team has begun development of a new data ingestion pipeline to replace several generations of previous pipelines. The new pipeline is designed to be more flexible and efficient than the current pipelines, allowing our processing to keep up with the increase in content. Work on this will be ongoing for the next two years and will require a significant amount of effort. The new ingestion pipeline architecture uses Apache Kafka, an event streaming system, to enable real-time processing to integrate content from new publishers, NASA repositories, and disciplinary archives with internal enrichment activities. The new architecture removes assumptions built into the ADS legacy pipelines, such as the format of its internal identifiers, the underlying bibliographic data model, and arXiv as the single source of preprints ingested in the system. It properly identifies and cross-links records of different kinds (articles, software, data sets) while retaining information about their provenance. Importantly, it retains licensing information to assess Open Access compliance for different collections.

3.1.2 Citation Processing

For recent astronomy journal articles, we match around 98% of citations in the bibliographies to records in our holdings. For similar earth science content, that drops to about 50%. These figures are lower overall for records from the early 2000s or before. To improve the resolution of citations in earth science literature, we are adding parsers and heuristics for the appropriate citation styles and publication names found in newly ingested papers and data collections.

The effort of extending the citation graph with earth science citations is essential to support interdisciplinary research and has positive consequences for historically core disciplines as well. Because ADS only shows citations that correspond to existing ADS records, citations to earth science papers from astronomy papers were generally missed in the past. This will no longer be the case with SciX.

3.1.3 Text Mining

Our current text mining processes require manual intervention and curation. Activities in this area will be aimed at creating and implementing robust pipelines to identify, validate, index and/or link data products mentioned in the available full-text papers automatically. Starting with data cited via DOI, this capability will expand to data links in FAIR data availability statements when available. We will also extract and index funding information from manuscripts when available, integrating this information with existing services provided by CHORUS and Crossref.

Data links and funding information are sometimes available in acknowledgements sections or footnotes; mining these sources will require more robust text mining approaches, which will be implemented incrementally over the five-year plan.

3.1.4 Metadata Normalization

The SciX curation staff performs a number of tasks to normalize data across publishers and institutions, including identifying and normalizing journal names, their abbreviations, and institutional affiliations; disambiguating author names; and mapping keywords. These efforts currently rely on a combination of scripts and manual curation. Work in this area will focus on automating this process as much as possible in a series of pipelines that interface appropriately with the new data ingestion pipeline architecture. The disciplinary project scientists in our expanded subject areas will ensure the accuracy of this processing.

3.2 Data Enrichment

We will enrich records in our database with text mining, labeling, and disambiguation methods to enable more sophisticated document retrieval and discovery.

We are currently developing a robust machine learning-based methodology to classify records upon ingestion so that we can automatically assign new papers to a particular collection. This classification will make it easier for researchers to limit their searches to the relevant literature, and for SciX to develop a discipline-specific view over its data holdings. We are also exploring various NER methods, including traditional text mining and machine learning approaches as well as the use of LLMs. Robust NER approaches, integrated into our data ingestion pipelines, will allow us to identify and extract from the text known entities corresponding to concepts, objects, and artifacts to uniquely identify relevant disciplinary knowledge.

3.2.1 Metadata Enrichment

Vocabularies and taxonomies improve the efficiency and effectiveness of information systems by facilitating interoperability and enhancing information retrieval (Cox et al., 2021). The SciX interface inherited from ADS several astronomical taxonomies in its filter section (keywords,

astronomical objects, institutions, etc.) that provide users effective ways to narrow searches and discover links to relevant information within and across disciplinary boundaries.

The SciX team has adopted the Gazetteer of Nomenclature and the Unified Astronomy Thesaurus (UAT; Frey & Accomazzi 2018) as the first two taxonomies with which to label all records in its core collection via NLP techniques. SciX released a UI capable of three-level facet searches on Gazetteer terms in December 2023 (see Figure 3) and will release a similar search of UAT keywords in 2024. While these two capabilities improve the search and discovery of concepts related to planetary science and astrophysics, they also demonstrate how we can leverage a variety of NLP techniques for exploring multiple levels of hierarchical concepts within the literature, which SciX plans to expand to taxonomies in other disciplines. Additionally, the Gazetteer contains geolocation data for the planetary features; we will implement a search capability over these data, then expand this for earth science data.

With expansion into earth sciences, we will develop similar capabilities for discipline-specific discovery using NASA's Global Change Master Directory keywords[45] (GCMD; Parsons et al 2023). We will develop NLP algorithms to classify records with keywords from the GCMD Earth Science and Earth Science Service categories as well as ways to expose the other categories of keywords to the extent that those keywords are already associated with ingested records. In collaboration with the Science Discovery Engine, we also anticipate using the Data Centers/Service Providers, Projects, Instruments/Sensors, and Platforms/Sources keyword categories to develop acronym and synonym lists relevant to earth science communities; these remain likely additional concepts for further expansion of NLP classification. Identification of additional tools to support earth science discovery will be based on an assessment of ways in which early adopters of SciX from these communities are using the service and explicit feedback from those communities regarding their prioritized requirements.

Building on our experience of indexing astronomy content using the Gazetteer and UAT and growing experience with indexing earth science content using the GCMD, we will assess additional relevant taxonomies used in each discipline and their potential value in supplementing the keyword systems already available. We will work with data providers in each discipline in order to provide greater interoperability of the indexed literature with the relevant research communities. For the most promising systems, we will attempt to develop tools to classify the literature with appropriate concepts of interest from each collection automatically. Systems under consideration include: US Geological Survey (USGS) Thesaurus[46], the Astrobiology Resource Metadata Standard Keywords (AHED)[47], Mars Target Encyclopedia[48], Space Physics Archive Search and Extract (SPASE) Dictionary[49], Physics Subject Headings[50], Semantic Web for Earth and Environmental Terminology (SWEET)[51].

SciX has also inherited from ADS the ability to resolve the names of stellar, galactic, and extragalactic objects and search for these objects through collaborations and integrations with Set

[45] https://www.earthdata.nasa.gov/learn/find-data/idn/gcmd-keywords
[46] https://apps.usgs.gov/thesaurus/
[47] https://ahed.nasa.gov/help/help-keywords
[48] https://pds-geosciences.wustl.edu/missions/mte/mte.htm
[49] https://spase-group.org/data/model/spase-2.6.0.pdf
[50] https://physh.org/
[51] https://github.com/ESIPFed/sweet

of Identifications, Measurements, and Bibliography for Astronomical Data (SIMBAD; Wenger et al 2000) and NASA/IPAC Extragalactic Database (NED; Mazzarella et al 2017). **The ability to resolve solar system objects and spacecraft similarly, and to search for relevant literature will benefit the planetary science, astrophysics, heliophysics, and earth science communities.** To do so, we will develop integrations with the Virtual Observatory's Solar System Open Database Network for access to aggregated solar system taxonomies maintained by multiple IAU working groups and also corresponding NLP algorithms to tag these entities in the literature. To resolve terrestrial locations, we will investigate ways to incorporate information from gazetteers such as the Central Intelligence Agency (CIA) World Factbook (2023) and Getty Thesaurus of Geographic Names (TGN; Getty Trust 2017), possibly through SWEET, which provides Web Ontology Language (OWL) wrappers for such external resources.

As an early digital library and as the leading astrophysical literature service, ADS had great leeway in establishing its own metadata and API standards. SciX serves a broader community; as such, it must link to and be accessible from a wider range of tools serving all of the SMD disciplines. To facilitate that interoperability, **SciX will review the recommendations of Science on Schema.org** (SOSO; Shepherd et al 2022) and similar efforts, such as Basic Formal Ontology[52] (BFO), to identify areas in which its metadata and fields should be adjusted and to develop a prioritized timeline for such improvements. Furthermore, SciX will participate in future development of these community standards.

3.2.2 Author Profiles

The increased complexity of author and affiliation management gives us the opportunity to improve our system to support functionality that will solve many of the most vexing problems associated with name ambiguity: the creation of internal author profiles.

Researchers' professional careers typically follow certain common patterns, where, for instance, junior scientists tend to change institutions with a certain cadence until they finally stabilize in more permanent mid-level/senior positions. Additionally, most researchers usually have very specific areas of interest where they will publish regularly, with topics that may rise and fade or evolve. Academic genealogies, such as those developed for astronomers,[53] physicists, and mathematicians,[54] also reveal researcher networks (Tenn 2016, Mulcahy 2017). This information is already implicitly contained in the data that the current and expanded system has or will have, such as affiliations, research topics, frequent collaborators and mentors, academic genealogy, identifiers, email addresses, read/access information. However, it is not straightforward to make it explicit by extracting, combining, and clustering all the key features in an automated and accurate way.

Despite the fact that ADS users can already search and claim papers manually with their ORCiD profile, many are not using ORCiD for reasons such as lack of time, knowledge, or perceived usefulness. While ADS has greatly facilitated the claiming of authorship for its users (over 1,030,000 claims since ORCiD integration in 2016), the activity of building and sharing one's bibliography has always required an ongoing effort which few keep up with, especially given the

[52] https://basic-formal-ontology.org/
[53] https://astrogen.aas.org/front/index.php;
[54] https://www.genealogy.math.ndsu.nodak.edu/index.php

lack of immediate reward. At the same time, discovering an author's complete and unambiguous list of papers has become more and more difficult as the scientific literature has seen a steady increase in the number of active researchers.

We believe that removing friction from the current paper claiming workflow will promote an increased adoption of ORCiD for active authors and lead to a corresponding decrease in ambiguity for users searching the system, resulting in a significant increase in research productivity. To make this possible, a substantial effort is required to develop a system that leverages the large amount of data in our possession in order to identify and expose user profiles automatically and link them to papers, enhancing the current ORCiD integration. Such a system will accomplish multiple goals:

- It will help active researchers to claim new papers (e.g., enabling the user interface to propose potential papers that the user might be a co-author of and which they might want to claim and add to their ORCiD profile).
- It will assist researchers maintain their ORCiD profiles, including identification of authors who have mistakenly created multiple ORCiD IDs.
- It will improve author disambiguation, which is expected to become an even greater challenge with the increase of records and authors in SciX.
- It will lay the foundation for creating a "people database" that will facilitate finding authors, potential collaborators, and disciplinary experts such as referees.

Rather than starting from scratch, **we will build upon our existing ORCiD infrastructure and make use of prior art and technology that has been successfully deployed by our collaborators in data science studies** (Mihaljević & Santamaría 2021) and digital library systems (King & Feldman 2021; INSPIRE 2022; OpenAlex 2023). Our author profiles are not meant to replace ORCiD, but rather augment it. Integration of this system with ORCiD will ensure maximum interoperability with the larger scholarly publishing ecosystem and provide efficiency and transparency in reporting research outputs.

Our author disambiguation will extend beyond ORCiD, to encompass authors without such an identification (whether due to ORCiD unfamiliarity or pre-deceasing its implementation). The assignment of internal author identifiers will naturally overlap with ORCiD and will allow us to assist with ORCiD curation and author searches within and across disciplines.

3.2.3 Curation and Linking of Bibliographies and Data Collections

Section 2.2.3 illustrated the collaborative work we have begun with some data providers within the disciplines covered by SciX. As we expand our efforts, we will work with all SMD repositories and make a concerted effort in supporting and incorporating the curation of their bibliographies in SciX. In addition to making this information more accessible, maintaining these bibliographies within NASA SciX will increase the transparency of the reporting process and make it possible to better evaluate impact and OA compliance of the relevant missions and projects.

To facilitate these efforts, we will use text mining and NER techniques to identify papers that mention elements relevant to NASA (e.g., the planetary data ecosystem elements[55]), with a focus

[55] https://science.nasa.gov/planetary-science/data/pde-elements/

on missions and the most easily identifiable elements, increasing granularity as we improve upon the technical capabilities of our models. These techniques will allow us to classify NASA data ecosystem element names (e.g., missions, archives, projects) and disambiguate them from other terms commonly found in the scientific literature indexed in SciX.

For example, the names for two of NASA's missions, "Cassini" and "Wind" have multiple meanings within each discipline, and human curation is currently required to identify the proper classification for each of them. Reliable mission, instrument, and archive identification will simplify the work of librarians, principal investigators, and curators who currently spend significant time creating program bibliographies. It will also help, though not eliminate, the amount of curation involved in maintaining links between papers and data sets hosted by the archives.

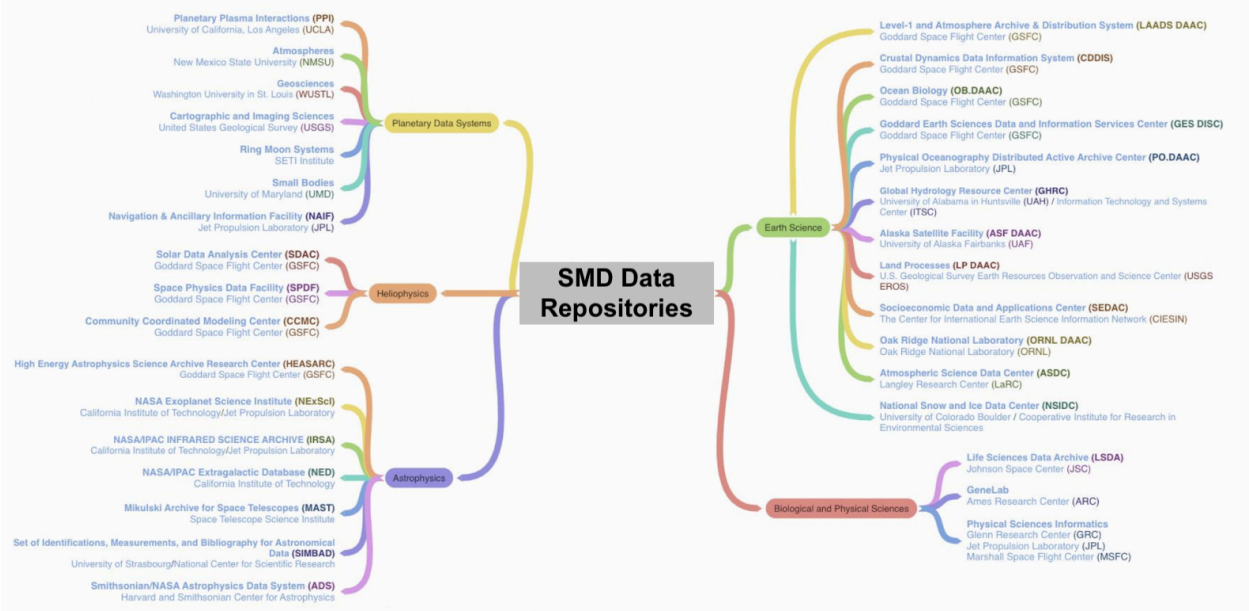

Figure 6. The SMD Data Repositories. ADS developed collaborations with the AP archives over the past 30 years and, more recently, with the PS and HP repositories. SciX has already started an effort to engage some of the ES DAACs which are maintaining bibliographic databases.

To further support these efforts, we will engage with all 32 SMD repositories (illustrated in Figure 6), as well as missions, organizations, and databases (hereafter collectively referred to as Data Centers or DCs). **We will work with DCs to ensure that their bibliographies and data collections are properly represented in SciX and that the related literature is properly ingested.** This will be a multi-year collaborative effort involving a sequence of incremental steps:

1. Develop formal collaborations with each DC; whenever necessary and feasible, develop collaborations at the sub-group or project level (e.g. AHED[56]).
2. Establish and maintain (both through text mining and curation at each DC) complete bibliographies of the work performed at each center.

[56] https://ahed.nasa.gov/

    a. Include in SciX all items in the bibliographies of the DC, and, where appropriate and feasible, include complete journals where these papers appear.
    b. Where appropriate and feasible, include in SciX all publications in the reference lists of items in the bibliographies of the DC, and the journals they are published in.
3. For each high-level data set at each DC, create a record for it in SciX so that it can be more easily discovered and cited.
    a. Index in SciX smaller data sets if they are re-used (via formal citation) outside of their original publication.
4. Establish and maintain through text mining at SciX and curation at each DC linkages between research publications and the archival data used to perform that research.
    a. Leverage mentions and formal citations of data products via DOIs when feasible
5. Use new technologies as technically and operationally feasible to develop new and enhance existing services in support of text mining efforts, as described in section 3.5.

3.3 Data Discovery and Search Infrastructure

A decade ago, NASA ADS adopted the popular open source project Apache Solr[57] for its search engine. Since then, it has been heavily customized and extended[58] to fulfill requirements based on the needs of our community: advanced query parsing syntax (e.g. specify first author searches), implementation of second order operations (e.g. search for the citations of a given set of papers), expansion of author synonyms and name variations (e.g. due to transliterations), and custom relevance scoring (taking document citations and usage into account).

The current implementation of the search engine and the pipeline that updates it is designed to provide timely indexing and efficient retrieval of documents that are relevant to a scholar, for instance: the latest papers written by a person, the latest citations made to a researcher, or current trending papers written on a topic. These capabilities are possible thanks to citation and usage networks that are built into the search engine and updated every time new data is ingested. With the proposed expansion, the number of indexed documents will significantly increase, while the citation and co-usage data will grow factorially, due to their network effect. At the same time, the number of users of the system will multiply as more scientists become frequent users of SciX.

All these factors require that we plan for **alternative strategies that can guarantee scalability and constant response times of our searches without sacrificing the advanced features built into the current system**. Architectural changes that may be required involve the pre-computation and replication of usage and citation networks, combined with an optimization strategy to distribute the search load among multiple machines when possible. Additional exploratory efforts will allow us to determine if recent developments in text embeddings and graph databases are suitable for use in our search ecosystem.

Additional finding aids will be necessary to **help users discover their desired content efficiently and view it in the corresponding context**. The experience with past ADS expansions has shown that as the size of the collection grows, providing users with additional tools to help them find, filter and analyze search results is important. The first improvement will

[57] https://solr.apache.org/
[58] https://github.com/adsabs/montysolr

consist of expanding the search filter selections by including discipline-specific taxonomies, and enriching records with corresponding concepts, which will allow users to easily filter search results. A second improvement will consist of modifying the relevancy ranking algorithm, now the default sort order in SciX, to take the user-selected discipline(s) into account when presenting search results to a user, thus elevating papers drawn from the relevant disciplinary literature, initially defined by curated collections within SciX. A third improvement will be the extension of the use of search highlights, including an exploration of text embedding techniques to enhance this feature, which allows users to see their search terms in context. This context will be particularly useful in those cases where semantic ambiguity exists, particularly across disciplines. As an example, the acronym HDF in an astronomy paper typically refers to the Hubble Deep Field whereas in an earth science context it refers to the Hierarchical Data Format.

With the expansion of the system into a multidisciplinary database, expending additional effort to **understand the user intent in order to help them find what they desire** will be necessary**.** This can be achieved at two different levels. The first level is related to the individual query, which can be analyzed to identify key concepts (drawn from a taxonomy based on the user-selected discipline) and spelling mistakes (typical for author names). The second is contextual, and it requires us to have a recent history of past user queries that might help disambiguate its intent. This is especially important when a query returns zero results or when the user issues consecutive queries without interacting with the system (i.e., not clicking on any result). These are situations where the system could assist the user by providing query suggestions (e.g., *"did you mean?"* feature, or auto-completion) that might help them achieve their goals and reduce the frustration of trying to find useful results in a constantly growing multidisciplinary database.

3.4 Data Management

The purpose of SciX is to build an organization which will collect, collate, correct and curate information centered on research publications from thousands of sources in hundreds of formats. These data will be organized into a highly interconnected complex system with many sophisticated avenues for user exploration and discovery. The use of SciX is completely open and available equally to any person on earth. **SciX is, itself, a long term repository for the data it contains and has taken steps to guarantee the preservation of the unique content it maintains**. This includes the implementation of a hybrid Disaster Recovery (DR) environment (on-premises and cloud), and the development of agreements with publishers and digital preservation services from Portico.

Much of the data will come from non-public sources with the distribution regulated by contracts and licensing agreements. To the extent that the agreements allow, the data will be open. When both open and non-open versions exist, especially of research articles, we will provide access to both. The versions of full-text documents stored in the SciX database (typically source XML documents) are not suitable for public display, but rather are meant for indexing purposes. The online versions of record of research articles are instead maintained and served by archives, publishers, and preprint servers.

The copyrighted nature of much of the full text and source material requires security and abuse prevention measures to ensure compliance with our terms of service. Users are subject to rate

limits, designed to prevent wholesale downloading of our database. These limits have been revised over time and are set to be high enough that individual researchers never reach them, so there is no effective restriction on the openness of the SciX data to them. However, the collection of such a quantity of literature in one database is invaluable for bibliometric and other studies of the literature. We will continue to work with large projects and researchers needing greater access to the database for particular purposes and can provide access to the full text content if needed.

The main data products of SciX are not static data tables, rather they are complex sets of documents, statistics, and links resulting from a query. The primary output consists of the relevant search result pages, but other options exist. Lists of articles can be exported in a variety of formats, providing users with a simple way to create bibliographies for inclusion in scientific papers. All the histograms and visualizations displayed by the UI can be conveniently downloaded as CSV files. For greater access to the underlying data, all functions of the system are accessible via the public API, which provides the entirety of SciX's public data via a modern JSON-based REST interface.

Additionally, SciX will produce labeled data products used for its curation efforts. The format and release frequency of these data sets will vary greatly, and will augment the data sets already made publicly available by the ADS project. Examples include the annotated datasets created for the 2022[59] and 2023[60] Workshops on Information Extraction from Scientific Publications (WIESP, held in conjunction with IJCNLP-AACL); the list of canonical institutions used by the ADS project[61] (shared with RoR and ADS users); the lists of synonyms/acronyms used by the search engine (shared with the Science Discovery Engine group); and any metadata crosswalks between taxonomies developed by the project.

All papers produced by the project's staff will continue to be published with an OA version, following current NASA guidelines. Software for the project is being developed in the open on GitHub[62]. Much of it is based on open-source projects such as Apache Solr and PostgreSQL. All code developed by the project is made available via permissive open source licenses for the benefit of the community.

To improve search quality, detect abuse, help design the system, and perform bibliometric research (Kurtz & Bolen 2010), SciX maintains a complete set of usage logs and access information. These data contain personally identifiable information and will never be made public due to the privacy regulations of the host institution[63]. They are strongly protected using AWS security infrastructure (AWS Identity and Access Management or IAM) and complying with the security policies from the Smithsonian Institution for the on-premise servers (e.g., relying on a demilitarized zone or DMZ, separate accounts to log into the gateway and servers, strong password rules).

SciX data holdings will be relatively small by modern standards, amounting to less than 100 terabytes. They will be continuously backed up off-site using commercial cloud storage

[59] https://huggingface.co/datasets/adsabs/WIESP2022-NER
[60] https://huggingface.co/datasets/adsabs/FOCAL
[61] https://github.com/adsabs/CanonicalAffiliations
[62] https://github.com/adsabs
[63] https://ui.adsabs.harvard.edu/help/privacy/

solutions. We expect the implementation of disaster recovery plans to change in the future as NASA refines its cloud storage and computing strategy.

3.5 Incorporating AI for Curation and Information Discovery

Recent advances in deep neural network models have led to a variety of improvements in the field of natural language understanding and information retrieval (Devlin et al 2018; Zhang et al 2022). In particular, embedding models (Pennington et al 2014; Dar et al 2022) generated from large text corpora excel in capturing the semantic nuances of language, which is important for understanding evolving terminology of scientific text within and across disciplines. These models have been successfully used to implement query interpretation and expansion, improved relevance ranking, and recommendations. Given the planned growth of content in SciX, its search capabilities must grow as well, especially in the context of interdisciplinary discovery. **A continued investment is required in the technologies underpinning SciX's knowledge representation and information discovery capabilities.**

Along with the promise of an exciting future, the latest AI technologies bring with them a lot of questions related to trust. While today's LLMs can be valuable tools in scientific investigation, the lack of transparency about the data on which they have been trained and their complexity means that their output requires interpretation, validation, and contextualization by human experts. As scientists, we have been trained to reject claims that cannot be substantiated, and therefore require knowledge systems based on evidence accrued through the scientific peer review process. Given the completeness of its knowledge base, **SciX is the obvious, authoritative organization that should curate and share data to train LLMs used in the earth and space sciences.**

Similarly, Knowledge Graphs (Hogan et al 2020) are a way of structuring and integrating knowledge based on relationships and entities, and are commonly used in various scientific domains. KGs are not only useful in formalizing and reconciling knowledge representations in multiple disciplines, but have applicability in search, recommendations, information extraction, and metadata normalization. Researchers are currently working on a unified approach to KGs and LLMs where the structured knowledge of the former guides the generative capabilities of the latter (Pan et al 2023). **SciX has used both LLMs and KGs in its first AI-enhanced pipeline (Shapurian et al 2023), and we expect that this will become common practice for future services and curation activities.**

As described in section 2.1.2, the SciX team has been investigating and using AI technologies for information retrieval, reasoning, and metadata enrichment. It has also collaborated with a variety of projects and NASA partners in advancing the use of AI agency-wide, including contributing to cross-divisional SMD initiatives[64]. Given the speed at which AI is evolving, and the central role that SciX will increasingly play in providing information services to the scientific communities it serves, it is essential that the project maintain in-house expertise on these technologies. The goal of this effort is not to have an AI-focused R&D group, but rather a dedicated person who can effectively follow these developments, collaborate with groups

[64] https://science.nasa.gov/researchers/open-science/smd-ai-initiative/

actively developing, fine-tuning, testing and evaluating these models, and integrate this technology into SciX.

SciX is engaged with multiple collaborations, some of which are described in section 2.1.2. In all cases, community members and collaborators turn to the SciX team given its unique ability to provide data and knowledge not otherwise available elsewhere, e.g.:

- Obtaining authoritative data sets to be used for model training (e.g. using SciX's corpus collection for model training);
- Creating labeled data to be used for model validation and evaluation (e.g. using the NER labeled data created for the WIESP workshops);
- Developing/adapting/evaluating models to be run on new content ingested by SciX in support of curation efforts (e.g. maintaining telescope bibliographies);
- Providing embedding APIs that can be used by the community to generate meaningful representation of scientific text (e.g. for exploration and classification tasks)
- Providing KG APIs returning entities extracted from the supplied text for NER support (e.g. metadata enrichment, text classification and disambiguation)
- Supporting applications that leverage SciX's infrastructure to provide value-added capabilities (e.g. AI agents to support research activities).

Given the current proliferation of open-source AI initiatives, we expect that such requests will accelerate. In order to meet the needs of the project and the larger community, SciX will need to adapt, evaluate, and extend its discovery services based on emerging AI technologies. This will involve an active and expanding set of collaborations, some of which have been described in section 2.1.4. **SciX's ability to innovate its capabilities and meet the needs of the community will require continued integration of AI technologies.**

A side-effect of having in-house expertise for AI/ML initiatives will provide us with the ability to develop new and improved services in SciX, such as improved search support, summarization of search results, better recommendations, and more accurate notifications.

4. Management and Budget

Enhancing SciX with expanded disciplinary content is an interconnected effort that involves collaborating with NASA funded archives, missions, and community outreach programs. Our key initiatives are based on the development plan outlined in section 3 and include building specialized collections for each discipline, enhancing search relevance and accuracy for discipline-specific terms through advanced AI/ML techniques, implementing synonym-aware search systems, disambiguating terms and concepts relevant to each discipline, establishing dedicated bibliographic groups and data collections, and partnering with community engagement projects.

The anticipated outcomes following the completion of this roadmap include technical parity across the main NASA SMD disciplines, a greater than 20% estimated increase in citations for data-linked papers (Henneken & Accomazzi 2012; Dorch, Drachen & Ellegaard 2016; Colavizza et al 2020), amplified scientific impact from data sets, and an increased impact of archival data (White et al 2009; Rebull et al 2017; Peek et al 2019). Additionally, by partnering with NASA

community outreach projects, we will establish a pathway for communication with end-users improving discoverability and accessibility of content produced within the NASA data ecosystem as a pillar of NASA infrastructure enabling open science.

4.1 The SciX Team

The SciX staff has increased and is still growing in response to an increase in the project's scope and development efforts. SR20 approved an ADS funding of 15.6 FTEs, and SciX22 increased funding for a total workforce of about 30 FTEs by FY25. This has led to a reorganization of the project's structure and an increase in shared responsibilities across the SciX team, as shown in the organization chart in Figure 7. The current proposal aims to achieve the staff levels awarded in the previous proposals and maintain them for the next 5 years.

With only minor changes, all the positions in the orgchart have been approved as described in the SciX22 proposal and funded through it. They are required to successfully execute the expansion described in this proposal. Vacant positions are being advertised according to the development schedule provided in SciX22 and are expected to be filled in Year 1.

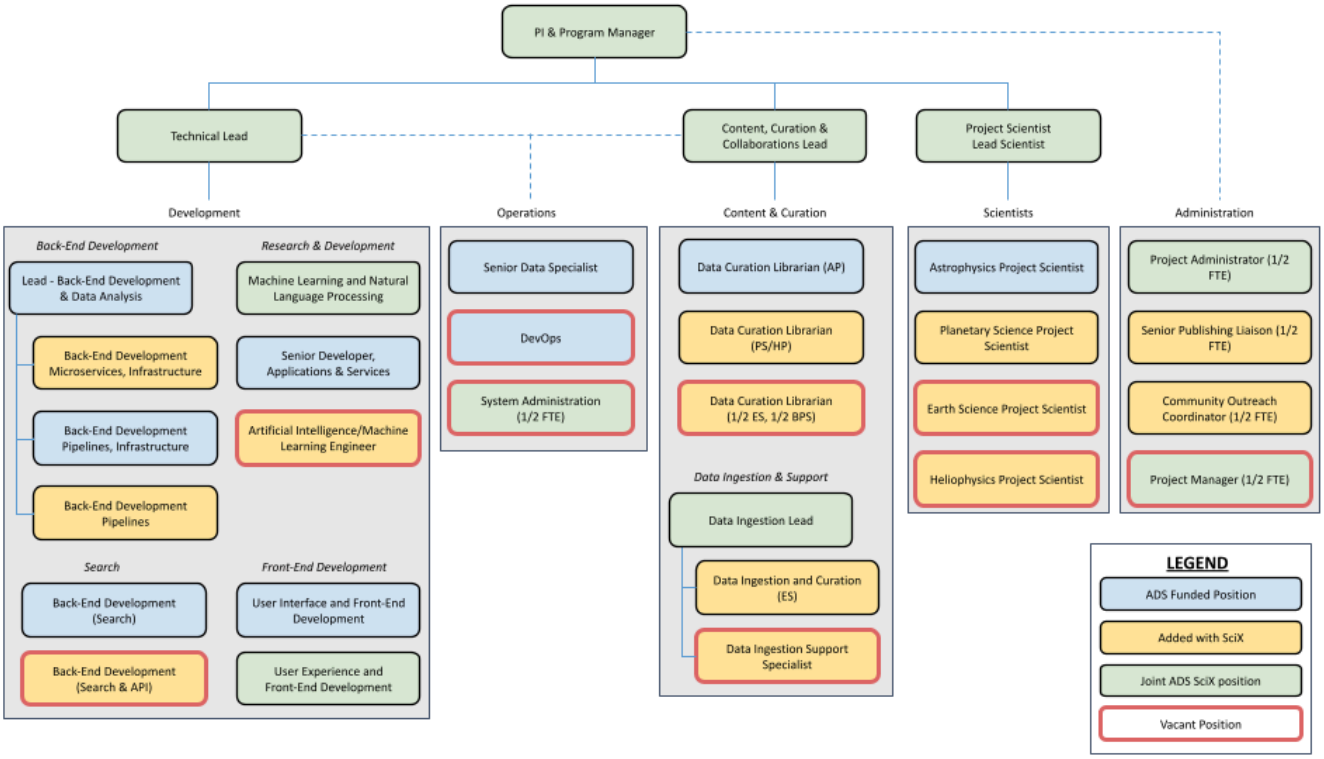

Figure 7: Organizational chart of the SciX team, which will be fully staffed in Year 1. The current SciX team members are listed on its website at https://www.scixplorer.org/scixabout/team/

4.2 System Development & Operations

Most of the effort involved in operating SciX will focus on a set of core activities, such as ingestion of content, user interface updates, user support, interoperability with research systems, collaborations with publishers and data providers, and maintenance of system infrastructure. These essential tasks will allow scientists using SciX to discover new and newly relevant content easily by narrowing or expanding the scope of their search by collection and exploring connections between documents through the citation and usage networks.

Most of the effort involved in the development of new SciX capabilities will focus on services and features that benefit all disciplines and communities, irrespective of their research focus. These fundamental services include author disambiguation, search infrastructure, text mining, citation processing, and metadata enrichment pipelines. However, the addition of new content and the requirements for its increased discoverability in new domains will require additional effort in various disciplinary areas such as parsing and normalization of new content, incorporation of new knowledge management systems, development of new user interface components, and fostering new collaborations.

Turning the SciX vision laid out in this proposal into reality critically depends on our ability to leverage both deep expertise in the disciplines we cover and the technological advances offered by the latest developments in AI/ML. Given the scope of the expansion, there is no scenario under which traditional ingest and curation alone can scale up to enable an ingestion of SciX content twice the size of the current ADS system. **Investment in technology provides the only viable path to the increase in efficiencies required to accomplish our goals, and informs the choices we have made in creating the SciX team**. It also informs the programmatic plan outlined in section 4.3.

Integration of AI/ML technologies requires both a dedicated engineering position within the team and computing infrastructure that can support KG and LLM integration in production systems. In addition to personnel, this requires us to develop a mixed computing environment which permits:

- Training: we will leverage the Smithsonian Institution's HPC cluster and opportunities provided by collaborative efforts with larger HPC centers such as ORNL.
- Inference (pipelines): we will use dedicated nodes in our on-premises computing cluster with restricted access to our proprietary data to run models over our full text and metadata enrichment pipelines.
- Inference (cloud): in due time, we will explore the use of user-facing AI tools and interfaces within our cloud-based API and UI, which will require a deployment of models in a scalable cloud environment.

The budget section describes the associated costs.

4.2.1 Disciplinary Developments

In addition to providing an up-to-date discovery platform for astronomy and astrophysics content, the ADS has also fulfilled the role of a literature archive providing free access to the historical literature published in all the major astronomy journals, conference proceedings and

observatory publications. This corpus of full-text, initially digitized by ADS in the mid- to late-90s, has been complemented by the availability of new content published electronically by the learned societies and journals in the field, leading to the current situation where virtually all of the research literature in astronomy is publicly available and accessible to anyone.

Given its role and responsibility in maintaining this full-text astronomy archive, ADS will always remain a unique resource for astronomy, providing a level of content and curation for AP which is beyond the scope of SciX. The level of funding from NASA AP reflects the dominant role that ADS has in the AP community, where its penetration reaches 100% of active researchers. This proposal aims to develop SciX into the feature-rich discovery platform for NASA SMD content that ADS has been for AP, but without the corresponding full-text digitization and archival component. This means, for instance, that 100% of the research papers relevant to ES will be indexed and discoverable (but unlike the AP content, not hosted) by SciX.

Given the different state of completeness and feature-richness of each SciX discipline, the development plan follows a natural progression where new capabilities are first developed and introduced in the more mature disciplines (AP, PS and HP), and later introduced in the remaining ones (ES and to some extent, BPS). For the later years, the timeline is tentative and subject to further refinement and prioritization based on community input.

Bringing ES content and services up to the level of AP is a substantial endeavor, as pure ES content is larger than pure AP content and NASA ES has twice the number of major archive centers as AP. Additionally, all SciX services are continually being enhanced, so reaching parity between ES and AP will be a moving target. As discussed in SciX22, we have developed a plan to achieve this goal by the end of FY2029, assuming steady funding that allows us to hire and retain a core set of individuals for the duration of this performance period. The schedule in Table 3 outlines the timeline for the development of both discipline-specific and discipline-agnostic (core system) activities.

4.3 Schedule

In this section we provide a schedule for the development and implementation of the major features in SciX (Table 3). Disciplinary-specific activities can be found on page 36 and disciplinary-agnostic activities on pages 37 and 38. Since the baseline funding scenario does not allow us to retain the optimal staff level (see 4.4.2), items marked in **orange** on the schedule correspond to tasks that will be delayed under baseline funding, while items marked in **red** are items that cannot be accomplished under baseline funding. Under the augmented budget outlined in section 4.4.1, both categories of items will be accomplished following the provided timeline.

Developing a plan under the flat baseline budget forces us to make difficult choices regarding which positions are reduced or eliminated. Such decisions have been guided by trying to minimize the impact of the cuts on the scientific functionality and capabilities of the overall system, while at the same time respecting the different funding streams (AP vs. other disciplines). The cuts attempt to strike an optimal balance between two axes: disciplinary focus vs. system infrastructure on the one hand; hand curation vs. ML-driven enrichment on the other.

DISCIPLINE SPECIFIC	Year 1	Year 2	Year 3	Year 4	Year 5	Year 5 + 1
Astrophysics *1 FTE scientist - 10% Data Curation & Pres., 30% Management, 60% User Support* *1 FTE librarian - 80% Data Curation & Pres., 20% User Support*	- Integrate data links from AP journals via text mining (2.1.3) - Normalize and curate data links across AP archives (2.2.3) - Continuing ingest of historical lit., dissertations, and software (2.1.2) - Continuing support for AP bibliographers for public library development (2.2.3) - Train tagging of Mission/instrument via NER, using existing labeled data (3.2.1) - Prototype alternative relevance ranking to improve retrieval (3.3)	**- Promote IVOA standard enabling archives to publish curated data links (2.1.3)** **- Implement improved relevance ranking algorithms for improved retrieval (3.3)** - Continue ingest of historical lit., dissertations, and software (2.1.2) - Continuing support for AP bibliographers for public library development (2.2.3) - Implement tagging of mission, instrument in literature via NER (3.2.1)	**- Explore using AI/ML to identify waveband information for observational papers (2.1.5)** **- Continuing support for AP bibliographers for public library development (2.2.3)** **- Continuing ingest of historical literature, dissertations, and software (2.1.2)**	**- Include mentorship information from AstroGen into author profiling (3.2.2)** **-Intensify community outreach to transition ADS users to SciX UI (2.2.1)** **- Continuing support for AP bibliographers for public library development (2.2.3)**	**- Develop techniques to text mine and integrate AP concepts in papers (3.2.1)** **- Transition remaining ADS users to SciX UI (2.2.1)**	**- Discontinue final support for independent ADS UI (2.2.1)**
Planetary Science *1 FTE scientist - 10% Data Curation & Pres., 30% Management, 60% User Support* *1/2 FTE librarian - 84% Data Curation & Pres., 16% User Support*	- Normalize and curate PS data links across PDS nodes (3.1.1) - Create and curate PS collection (3.3) - Prototype PS-specific relevance ranking algorithms to improve retrieval (3.3)	- Implement PS-specific relevance ranking algorithm for improved retrieval (3.3) - Curate and incorporate bibliographies of NASA PS missions (2.2.3) - Implement tagging of Mission/instrument via NER (3.2.1)	**- Integrate APIs for object aliases (SSODNET, JPL Horizons) (3.2.3)** -Curate and maintain lists for solar system objects (3.2.1) -NER for disambiguating planetary objects and concepts (3.2.1) -Improve synonym-aware search utilizing existing taxonomies (3.2.1)	**- Develop ML-based systems for NER of PDE elements (3.2.3)** - Index funding source to corresponding analysis program (i.e., DAPS) (3.1.3) - Curate collections for data analysis programs (3.2.3)	**- Continue NER of PDE elements (3.2.3)** - Index literature-software links that support PS data analysis (1.4) - Curate PS software collection (1.4)	**- Develop techniques to text mine and integrate PS concepts in papers (3.2.1)** - Normalize and curate data links across PSA (3.1.1) - Normalize and curate data links across PDS-like archives (e.g., CNEOS) (3.2.3)
Heliophysics *1 FTE scientist - 10% Data Curation & Pres., 30% Management, 60% User Support* *1/2 FTE librarian - 84% Data Curation & Pres., 16% User Support*	- Create and curate HP collection (3.3) - Incorporate bibliographies of NASA HPD and HP missions (2.2.3) - Begin UAT collaboration for expanded HP coverage (3.2.1)	**- Incorporate bibliographies of ESA HP missions (2.2.3)** **- Normalize and curate data links across HP repositories (2.2.3)** - Implement tagging of Mission/instrument in literature via NER (3.2.1) - Normalize and curate HP data links across HPD nodes (2.2.3)	**- Make substantial progress on HP gray literature (2.1.2)** **- Map UAT concepts with existing HP knowledge bases such as e.g. SPASE (3.2.1)**	**- Provide deeper integration with NASA HPD, e.g. the Heliophysics Digital Resource Library (HDRL) (2.2.3)**	**- Liaise with HSO Connect to support recommendations to community curated resources from user queries (2.2.3)**	**- Develop techniques to text mine and integrate HP concepts in papers (3.2.1)**
Earth Science *1 FTE scientist - 10% Data Curation & Pres., 30% Management, 60% User Support* *1.5 FTE librarian - 42% Data Curation & Pres., 8% User Support*	- Harvest and ingest metadata for all records from ES priority journals (3.1.1) - Incorporate bibliographies of all DAACs that maintain them (3.2.3, 2.2.3) - Investigate tagging of Mission/instrument via NER (3.2.1)	**- Process references for all records from ES priority journals (3.1)** **- Make substantial progress on ES gray literature (2.1.1)** **- Implement initial enrichment with GCMD keywords (3.2.1)** - Index full text for all records from ES priority journals (1.4, 2.1.1) - Collaborate with all DAACs to create bibliographies (2.2.3,3.2.3) - Implement tagging of Mission/instrument via NER (2.1.5,3.2.1)	- Complete acquisition of ES refereed literature (2.1.2) - Improve ES search semantics and ranking (3.2) **- Cited literature for refereed ES collection ingested at 90+% level (2.2.1)** **- Investigate geolocation tagging of literature via NER (2.1.1)**	- Normalize and curate data links across ES repositories (3.2.3) **- Make substantial progress on ES gray literature (50% completeness) (2.1.1)** **- Develop geolocation tagging of literature (2.1.1)**	**- Improve coverage ES gray literature (80% completeness) (2.1.1)** **- Cited literature for gray ES literature ingested at 50+% level (2.1.1)**	**- Cited literature for gray ES literature ingested at 80+% level (2.1.1)**
Biological & Physical Sciences *1/2 FTE librarian - 84% DC&P., 16% User Supp.*	- Begin census of NASA-funded BPS refereed literature (2.2)	**-Start ingestion of NASA-funded BPS literature (2.2)**	**- Complete acquisition of NASA-funded BPS refereed literature (2.2)**	**- Make substantial progress on NASA-funded BPS gray lit. (2.2)** **- Merge relevant preprint content for NASA-funded BPS lit. (3.1.1)**	**- Index/link data and software cited in NASA-funded BPS literature via DOIs (2.1.3)**	**- Promote SciX at BPS conferences and audiences (2.2.5)**

DISCIPLINE AGNOSTIC	Year 1	Year 2	Year 3	Year 4	Year 5	Year 5 + 1
Software Development & Maintenance *(API support, user interface, tools & applications, search & database technologies, data processing pipelines, etc)* *9 FTE/yr IT Specialists*	- Improve software and data citation capture pipeline across disciplines (2.2.3) - Implement NER algorithm for Mission/instrument tagging (3.2.3) - Implement search relevance algorithm as a function of discipline selection (3.3) - Improve pipeline for text mining of software and data links (3.1.3) - Implement article classification in collections and filtering via UI (3.2)	- Improve NER of Mission/instrument using SOTA models (3.2.3) - Implement ORCID/author name lookup for registered users (3.2.2) - Fine-tune and adapt existing LLMs for data curation tasks (2.1.4, 2.1.5)	**- Incorporate KG techniques in author disambiguation infrastructure (3.2.2)** **- Provide text embeddings API endpoint (3.5)** **- Implement author profiles infrastructure (3.2.2)** **- Provide taxonomic extraction of terms from an API endpoint (3.2.1)**	**- Enhance upkeep of author profiles via "Is this your paper?" functionality (3.2.2)** **- Provide pipeline execution environment to partners and collaborators (3.5)** **- Implement public author profiles (3.2.2)** **- Expose knowledge graphs via API (3.5)**	**- Incorporate LLM and KG technologies in search engine (3.3, 3.5)** **- Enable taxonomy-based disambiguation of user queries (3.3, 3.5)**	**- Create links based on crosswalks between disciplinary taxonomies (3.4)** **- Integrate author profiles into researcher network database promoting interdisciplinary collaboration (3.2.2)**
User Support *(documentation and tutorials; helpdesks; outreach such as workshops, webinars and conference booths; user groups)* *3.4 FTE/yr aggregated from partial FTE contributions across organization*	- Liaise with ES NASA partners and community groups (2.2.5) - Liaise with Community groups e.g. NASA SCoPE, TOPS (2.2.5) - Collaborate with UAT for expanded PS coverage (2.2.5) - Create formal partnerships with NASA HPD archives (2.2.5) - Establish and organize ambassador program and workshop (2.2.5) - Create training materials (2.2.5) - Attend conferences: presentations, booths, workshops (2.2.5)	- Liaise with academic institutions to develop resources for K-12 (NASA K-12 education portals), community college (NASA community college network), and 4-year undergraduate programs (4.3.2) - Liaise with missions for ingestion of mission-related content (4.3.2) - Continue conference activities **(skip 30% of these planned activities) (2.2.5)**	- Continue liaise and collaboration activities (2.2.5) - Ongoing user outreach (2.2.5) - Maintain user documentation (3.3) - Continue conference activities **(skip 10% of these planned activities) (2.2.5)**	- Continue liaise and collaboration activities (2.2.5) - Ongoing user outreach (2.2.5) - Maintain user documentation (3.3) - Continue conference activities **(skip 1 conference) (2.2.5)**	- Continue liaise and collaboration activities (2.2.5) - Ongoing user outreach (2.2.5) - Maintain user documentation (3.3) - Continue conference activities **(skip 30% of these planned activities) (2.2.5)**	- Continue liaise and collaboration activities (2.2.5) - Ongoing user outreach (2.2.5) - Maintain user documentation (3.3) - Continue conference activities **(skip 40% of these planned activities) (2.2.5)**
Operations & Infrastructure *(computing, storage, licenses, security, etc.)* *2.7 FTE/yr - DevOps and IT Specialists*	- Improve interoperability with the Science Discovery Engine (2.2.3) - Install multi-GPU node (4.4.1) - Maintain cloud infrastructure (2.2.2) - Maintain secure on-premises servers (2.2.2)	- Install compute nodes (4.4.1) - Maintain cloud infrastructure (2.2.2) - Maintain secure on-premises servers (2.2.2)	- Install network storage unit (4.4.1) - Maintain cloud infrastructure (2.2.2) - Maintain secure on-premises servers (2.2.2)	- Install compute nodes (4.4.1) - Maintain cloud infrastructure (2.2.2) - Maintain secure on-premises servers (2.2.2)	- Install multi-GPU node (4.4.1) - Maintain cloud infrastructure (2.2.2) - Maintain secure on-premises servers (2.2.2)	- **Install compute nodes**, network switches (4.4.1) - Maintain cloud infrastructure (2.2.2) - Maintain secure on-premises servers (2.2.2)
Data Curation & Preservation *(content selection & ingest; data harvesting & collection; data extraction, transform, loading; metadata enrichment)* *6.55 FTE/yr - Librarians and IT specialists*	- Implement ingestion of data and software according to new indexing policies (2.1.3) - Develop and deploy data harvesting procedures (3.1.1) - Start design of SciX record identifier architecture (3.1.1) - Replace legacy public scan explorer with new solution (2.2.2) - Design and implement workflow for massive harvest and indexing of DOI-based data (2.1.2)	- Extend preprint/journal article matching (2.2.2) - Improve coverage of Green Open Access versions of articles (1.4) - Start exploring OpenAlex as metadata enrichment source (3.2.2) - Switch over to new reference data processing pipeline (2.2.2) - Complete implementation plan for SciX record identifier architecture (3.1.1)	**- Implement Science-on-Schema metadata encoding (3.2.1)** - Publish Open Corpus collection of OA content (4.3.2) - Utilize OpenAlex to augment existing SciX metadata (3.2.2) - Start with design of data store for reference data (2.2.2)	- Implement preservation solution for digitized content (e.g. Portico) (3.4) - Implement new data store solution for reference data (2.2.2)	- Generate searchable PDF/A documents for content digitized by ADS (3.4) - Implement ingest and indexing of papers with code and Jupyter Notebooks (2.1.3) - Design dashboard for curation team (3.5)	**- Curate bibliographies from interdisciplinary projects (e.g. Research Coordination Networks) (4.3.6)** **- Implement dashboard for curation team (3.5)**

Management *(partner relations, hiring, budget, schedules, reviewing, reporting)* *5.35 FTE/yr - PI, Fund Manager, Division Administrator, Project Manager, Publisher Liaison*	- Hire HP scientist, AI/ML engineer (4.1, 3.5) - Liaise with ESA partners (3.2.3) - Negotiate agreements with publishers and data archives (1.4, 2.1.3) - Organize team meetings and staff mentoring (4.1) - Organize SciX/ADSUG advisory board meetings (2.2.5)	- Negotiate agreements with publishers and data archives (1.4, 2.1.3) - Develop collaborations and partnerships (2.2.3, 3.2.3, 3.5) - Organize team meetings and staff mentoring (4.1) - Organize SciX/ADSUG advisory board meetings (2.2.5)	- Negotiate agreements with publishers and data archives (1.4, 2.1.3) - Develop collaborations and partnerships (2.2.3, 3.2.3, 3.5) - Organize team meetings and staff mentoring (4.1) - Organize SciX/ADSUG advisory board meetings (2.2.5)	- Negotiate agreements with publishers and data archives (1.4, 2.1.3) - Develop collaborations and partnerships (2.2.3, 3.2.3, 3.5) - Organize team meetings and staff mentoring (4.1) - Organize SciX/ADSUG advisory board meetings (2.2.5)	- Negotiate agreements with publishers and data archives (1.4, 2.1.3) - Develop collaborations and partnerships (2.2.3, 3.2.3, 3.5) - Organize team meetings and staff mentoring (4.1) - Organize SciX/ADSUG advisory board meetings (2.2.5)	- Negotiate agreements with publishers and data archives (1.4, 2.1.3) - Develop collaborations and partnerships (2.2.3, 3.2.3, 3.5) - Organize team meetings and staff mentoring (4.1) - Organize SciX/ADSUG advisory board meetings (2.2.5)

LEGEND
Tasks that will be delayed under baseline budget
Tasks that will be canceled under baseline budget

Table 3: Schedule of activities. Top: discipline-specific tasks, broken down by proposal year. Bottom: discipline-agnostic tasks, broken down in the categories requested by the NASA budget template. Each task refers back to the proposal section (in parenthesis) in which the corresponding activity is described. Tasks that are imperiled or delayed by baseline budget funding are indicated in red and orange. For a full description of the budget scenarios and the cuts associated with them please refer to section 4.4.

4.4 Budgets

Keeping the SciX team fully staffed at all times has been a challenge. While NASA has granted the project funding to expand its staff, resignations and job market pressures have prevented the team from being fully staffed for the past three years. However, we have been successful in recruiting and promoting staff members to fill the critical positions of the newly expanded organizational structure. We seek to continue on this path in order to complete the effort outlined in SciX22 and further refined in this proposal. In the next sections, we will first describe the budget that brings the project capabilities to the capacity required to implement the SciX vision throughout the period of performance (augmented budget), and then the cuts required to the program to meet the agency's guidelines (baseline budget).

4.4.1 Augmented Budget

The budget guidelines provided to us by NASA are the result of the merged resources of the ADS project and the SciX expansion, which is a total staffing level of approximately 30 FTEs (see table 4). The Augmented Budget is tailored to provide stability and longevity to the effort so that all of the outlined goals can be met within the five-year nominal timeline. Importantly, it will provide the project with the capabilities needed to meet NASA's Open Science goals going forward, rather than generate an initial burst of initiatives without the sufficient long-term support required for SciX to succeed.

Annual hardware purchases range between $100K - $200K in Years 1-5. Every year, we perform system refreshes by investing in a multi-GPU node (Y1 and Y5), compute nodes (Y2, Y4, and Y5+1), network storage unit (Y3), and network switches (Y5+1). Cloud computing costs (AWS) average $300K/year and allow us to keep our services responsive and available 24/7 using our tested computing environment. Other services include payment of annual support fees for the SIMBAD database ($140K), and the cost of memberships in scholarly publishing organizations (ORCiD, CrossRef, publisher fees, amounting to $15K). Advisory board meeting and Ambassador program costs are estimated at $40K/year.

Travel costs average $380K/year to cover all expenses associated with attending community events where SciX needs to be promoted, and conferences and workshops focused on open science, earth and space science informatics, and data science. The former include meetings of the American Astronomical Society (AAS), the AAS Division of Planetary Science (DPS), Lunar and Planetary Science Conference (LPSC), the American Geophysical Union (AGU), American Meteorological Society (AMS), Triennial Earth Sun Summit (TESS), the Geological Society of America (GSA), and select European Geophysical Union (EGU) General Assembly Meetings. The latter includes meetings of the Earth Science Information Partners (ESIP), Astronomical Data Analysis, Systems and Software (ADASS), the NASA SMD Repository Workshops, selected Research Data Alliance (RDA) Plenaries, the International Conference on Computational Linguistics (COLING), and the International Joint Conference on Natural Language Processing (IJCNLP). Additionally, a part-time community engagement consultant with a background in ES is working with the SciX team to promote the project.

The largest portion of the budget is allocated to staff salaries. While our program has been carefully planned to minimize costs through the use of part-time consultants where appropriate, it still requires funding for an annual increase in costs due to mandatory Cost of Living Adjustments (5% per SAO's inflation index), and the natural progression of performance-based salary increases necessary to retain high-performing staff.

4.4.2 Baseline Budget

NASA provided budget guidelines with a breakdown between AP-funded activities and SciX expansion activities. The guideline for the AP portion of the budget is $4.4M in Year 1 and $4.5M in Years 2-5+1. The guideline for the non-AP portion of the budget is $4M for Years 1-5+1. These guidelines allow us to reach the level of optimal staffing in Year 1, but fail to maintain cost of living adjustments for Years 2-5+1, causing a severe shortfall in funding that would force us to decrease staff levels and miss the goals laid out in this proposal. This uncertainty is particularly damaging because the project is currently in an expansion phase and attempting to hire talent to staff its team. Given the challenges SAO faces in matching the high salaries typical for the tech industry, offering stable, multi-year positions is critical for attracting and retaining skilled IT professionals to contribute to our efforts. Most importantly, without prospects of working in a stable and engaging environment, the best performing employees will be the first ones to leave.

In order to meet the baseline budget, a number of cumulative cuts will need to be made on a year-to-year basis. We briefly discuss them below, qualifying the source of the cuts by specifying "ADS" for the AP-supported effort, and "SciX" otherwise.

Year 2 cuts (shortfall of $421K): reduce HP scientist position to 0.5 FTE (SciX); reduce ES/BPS librarian to 0.5 FTE (SciX); reduce SciX travel by $53K; reduce AP curator position to 0.55FTE (ADS) and reduce ADS travel by $11K.

Year 3 cuts (shortfall of $954K): in addition to the above, remove back-end developer (SciX); remove ingest support specialist position (SciX) and reallocate 0.35FTE from ADS ingest to SciX ingest; reduce SciX travel by $15K; remove 0.55 FTE AP curator position (ADS) and reduce ADS travel by $11K.

Year 4 cuts (shortfall of $1,334K): in addition to the above, reduce AI/ML engineer position to 0.6 FTE (SciX); reduce SciX travel by $4K; remove 0.5 FTE from UI/UX developer position (ADS).

Year 5 cuts (shortfall of $1,886K): in addition to the above, remove 0.5 FTE UI/UX developer position (SciX); remove 0.5 FTE ES/BPS librarian position (SciX); remove 0.6 FTE AI/ML engineer position (SciX); reduce SciX travel by 42K; reallocate 0.3 FTE front-end developer from ADS to SciX; reduce ADS travel by $38K

Year 5+1 cuts (shortfall of $2,089K): in addition to the above, remove purchase of all compute nodes; reduce SciX travel by $63K; reduce ADS travel by $48K.

The corresponding budgets and staff reductions are summarized in the "Total" row of the augmented budget, at the bottom of Table 4. By year 5+1, the project will have faced a loss of

6.51 FTEs over Year 1, or 22% of its staff. The consequences of this shortfall will have multiple negative impacts:

1. The reduction of a HP Project Scientist in years 3-5+1 poses a significant threat to the interdisciplinarity of SciX. The absence of a dedicated full-time HP researcher compromises the project's ability to properly serve the heliophysics community and undermines our commitment to our vision of an interdisciplinary information nexus for **all** NASA SMD disciplines.
2. The reduction in Year 4 and elimination in Year 5 of the AI/ML Engineer significantly hinders SciX's technological advancement. Without a dedicated AI/ML expert, we face a severe setback in our ability to incorporate and use the rapidly evolving AI technology stack, which in turn undermines our commitment to maintaining a forefront position in AI/ML advancements for NASA SMD disciplines.
3. The reduction in Year 4 and elimination in Year 5 of the UI/UX Developer hinders system usability, team redundancy, and user experience design leading to user dissatisfaction and ultimately loss of users, especially in newer disciplines
4. Removal of Back-End Developer and AI/ML Engineer leads to potential system downtimes, inefficiencies, and reduced ability to keep pace with technological advancements.
5. Elimination and reduction of librarian roles (AP and ES/BPS Librarians) significantly affects the selection, ingestion, retrieval, and management of scientific data for those disciplines, reducing the interdisciplinarity of SciX.
6. Consistent reduction in travel budget across years limits our ability to connect with newer communities (ES + HP most affected) reducing their adoption of SciX, which in turn delays the attunement of SciX to their needs along with numerous unique SciX benefits to researcher productivity.
7. The effect of hiring short-term, and/or part-time employees for key positions like HP Scientist and ES/BPS librarian leads to lower employee engagement and commitment, posing challenges in attracting and retaining top talent.
8. The reduction in FTE for several roles could result in decreased morale and productivity among the team, affecting overall project momentum and success, which could result in additional turnover in currently held key positions.

Taken cumulatively, these cuts significantly diminish the nature and impact of the SciX platform, rendering it far less useful as a unifying research tool for NASA. The risks of underfunding the project during its critical expansion phase carry implications not just for the project but for NASA SMD’s open science ecosystem and for the global scientific community that depends on it. It is imperative that the necessary resources are allocated to SciX, ensuring its position as a cornerstone of open science and a model for future scientific collaboration and discovery

I. FY25 - FY30 NASA Full-cost Guideline ($K)

	FY25		FY26		FY27		FY28		FY29		FY30	
Budget Guidelines	$4,400 / $4,000		$4,500 / $4,000		$4,500 / $4,000		$4,500 / $4,000		$4,500 / $4,000		$4,500 / $4,000	

II. FY25 - FY30 '5-way' Functional Breakdown ($K+FTE/WYE) **for Astrophysics-funded work**

	FY25		FY26		FY27		FY28		FY29		FY30	
	Total Cost	FTE/WYE	Total Cost	FTE/WYE	Total Cost	FTE/WYE	Total Cost	FTE/WYE	Total Cost	FTE/WYE	Total Cost	FTE/WYE
1. Operations and Infrastructure	$832	1.59	$846	1.60	$938	1.60	$924	1.60	$1,017	1.60	$944	1.60
2. Software Dev. & Maintenance	$1,207	4.99	$1,302	4.99	$1,415	4.99	$1,367	4.52	$1,327	4.20	$1,378	4.20
3. Data Curation & Preservation	$870	3.03	$835	2.67	$652	1.98	$672	1.98	$649	1.88	$659	1.88
4. User Support	$837	2.00	$830	1.91	$796	1.73	$818	1.71	$772	1.66	$770	1.66
5. Management	$655	2.21	$686	2.21	$696	2.17	$717	2.17	$733	2.16	$748	2.16
Total	**$4,400**	**13.82**	**$4,498**	**13.38**	**$4,499**	**12.48**	**$4,499**	**11.98**	**$4,498**	**11.49**	**$4,499**	**11.49**

III. FY25 - FY30 '5-way' Breakdown ($K+FTE/WYE) **for SMD-funded work**

	FY25		FY26		FY27		FY28		FY29		FY30	
	Total Cost	FTE/WYE	Total Cost	FTE/WYE	Total Cost	FTE/WYE	Total Cost	FTE/WYE	Total Cost	FTE/WYE	Total Cost	FTE/WYE
1. Operations and Infrastructure	$365	0.59	$343	0.60	$370	0.40	$324	0.40	$388	0.40	$297	0.40
2. Software Dev. & Maintenance	$849	4.04	$934	4.09	$834	3.19	$857	3.01	$797	2.58	$832	2.58
3. Data Curation & Preservation	$928	4.66	$931	4.31	$929	3.70	$878	3.52	$855	2.93	$884	2.93
4. User Support	$1,108	3.56	$1,038	3.18	$1,069	3.10	$1,111	3.06	$1,097	2.94	$1,101	2.94
5. Management	$750	2.88	$753	2.76	$800	2.79	$831	2.79	$864	2.81	$884	2.81
Total	**$4,000**	**15.74**	**$4,001**	**14.93**	**$4,002**	**13.18**	**$4,000**	**12.78**	**$4,002**	**11.66**	**$3,999**	**11.66**

IV. FY25 - FY30 '5-way' Breakdown for Augmented Budgets ($K+FTE/WYE)												
	FY25		FY26		FY27		FY28		FY29		FY30	
	Total Cost	FTE/WYE	Total Cost	FTE/WYE	Total Cost	FTE/WYE	Total Cost	FTE/WYE	Total Cost	FTE/WYE	Total Cost	FTE/WYE
1. Operations and Infrastructure	$0	0.00	$0	0.00	$40	0.20	$44	0.20	$47	0.20	$133	0.20
2. Software Dev. & Maintenance	$0	0.00	$2	0.00	$188	0.90	$389	1.56	$645	2.31	$674	2.31
3. Data Curation & Preservation	$0	0.00	$193	0.83	$485	2.13	$638	2.31	$795	3.00	$838	3.00
4. User Support	$0	0.00	$187	0.56	$197	0.63	$215	0.69	$347	0.85	$391	0.85
5. Management	$0	0.00	$39	0.15	$43	0.15	$48	0.15	$52	0.15	$53	0.15
Total	**$0**	**0.00**	**$421**	**1.54**	**$954**	**4.00**	**$1,334**	**4.90**	**$1,886**	**6.51**	**$2,089**	**6.51**

Table 4: 5-way breakdown of budgets. Part I: NASA guidelines for AP/SMD funded work. Part II: AP-funded baseline budget. Part III: SMD-funded baseline budget. Part IV: augmented budget (both AP and SMD funded) required to maintain a constant effort throughout the proposal years. The breakdown of the budget categories includes the following activities:

1. ***Operations and Infrastructure:*** *compute & storage on-premises & cloud, network, licenses, security, privacy, etc.*
2. ***Software Development & Maintenance:*** *API support, user interface, tools & applications, search & database technologies, data processing pipelines, etc.*
3. ***Data Curation & Preservation****: content selection & ingest; data harvesting & collection; data extraction, transformation, loading; metadata enrichment*
4. ***User Support:*** *documentation and tutorials; helpdesks; outreach such as workshops, webinars and conference booths; user groups*
5. ***Management:*** *partner relations, hiring, budget, schedules, reviewing, reporting*

References

Accomazzi, A., Novacescu, J., Frey, K., & Protopapas, P. (2022) “Unified Astronomy Thesaurus (UAT) Integration in ADS Search and Discovery,” ADS Blog. https://www.scixplorer.org/scixblog/uat-integration

Accomazzi, A. (2024) “Decades of Transformation: Evolution of the NASA Astrophysics Data System's Infrastructure.” arXiv:2401.09685

Alkan, A. K., Grouin, C., Schussler, F., & Zweigenbaum, P. (2022) “TDAC, The First Corpus in Time-Domain Astrophysics: Analysis and First Experiments on Named Entity Recognition.” In Proceedings of the first Workshop on Information Extraction from Scientific Publications. https://aclanthology.org/2022.wiesp-1.15/

Allen, Thomas (2023) “SciX Models and Datasets,” SciX blog. https://scixplorer.org/scixblog/ads-models-and-datasets

Blanco-Cuaresma, S., Ciucă, I., Accomazzi, A., et al. (2023) “Experimenting with Large Language Models and vector embeddings in NASA SciX,” arXiv:2312.14211

CIA. (2023) “The World Factbook 2023.” Washington, DC: Central Intelligence Agency. https://www.cia.gov/the-world-factbook/

Cox SJD, Gonzalez-Beltran AN, Magagna B, Marinescu M-C (2021) “Ten simple rules for making a vocabulary FAIR.” PLoS Computational Biology 17(6): e1009041. https://doi.org/10.1371/journal.pcbi.1009041

Cayrel, R., et al (1974) “The Bibliographic Star Index.” Bulletin d'Information du Centre de Données Stellaires, vol. 6, p.24 https://scixplorer.org/abs/1974BICDS...6...24C/abstract

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020) “The citation advantage of linking publications to research data,” PLoS ONE, vol. 15, issue 4, p. e0230416. https://doi.org/10.1371/journal.pone.0230416

Dar, G., Geva, M., Gupta, A., and Berant, J. (2022) “Analyzing Transformers in Embedding Space.” arXiv:2209.02535

de Solla Price, D. J. (1961) “Science Since Babylon.” Yale University Press, New Haven, Connecticut. http://derekdesollaprice.org/science-since-babylon/

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018) “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv:1810.04805

Dorch, B. F., Drachen, T. M., & Ellegaard, O. (2016) “The data sharing advantage in astrophysics.” Astronomy in Focus, as presented at the IAU XXIX General Assembly, 2015. Proceedings of the IAU, Volume 29A, pp. 172-175. arXiv:1511.02512

Frey, K. and Accomazzi, A. (2018) “The Unified Astronomy Thesaurus: Semantic Metadata for Astronomy and Astrophysics.” The Astrophysical Journal Supplement Series, vol. 236, 24. https://doi.org/10.3847/1538-4365/aab760

Global Change Master Directory (GCMD). (2023). GCMD Keywords, Version 17.2, Greenbelt, MD: Earth Science Data and Information System, Earth Science Projects Division, Goddard Space Flight Center, NASA. https://forum.earthdata.nasa.gov/app.php/tag/GCMD+Keywords

Grant, C. and Templeton, M. (2020) “Affiliation searches: the Why, What, and How of our Canonical Affiliation Feature.” ADS Blog. https://www.scixplorer.org/blog/affiliations-feature

Grezes, F., Blanco-Cuaresma, S., Accomazzi, A., et al. (2021) “Building astroBERT, a language model for Astronomy & Astrophysics” arXiv:2112.00590

Grezes, F., Blanco-Cuaresma, S., Thomas, A., Ghosal, T. (2022) “Overview of the First Shared Task on Detecting Entities in the Astrophysics Literature (DEAL).” First Workshop on Information Extraction from Scientific Publications, held online November 20, 2022. https://aclanthology.org/2022.wiesp-1.1/

GROBID (2008-2023). Software Package. https://github.com/kermitt2/grobid

Henneken, Edwin (2023) “The ADS Curation Model,” ADS Blog. https://www.scixplorer.org/scixblog/curation-model

Henneken, E., et al. (2007) “E-prints and journal articles in astronomy: a productive co-existence” Learned Publishing, Volume 20, page 16 https://scixplorer.org/abs/2007LePub..20...16H/abstract

Henneken, E. & Accomazzi, A. (2012) “Linking to data: Effect on citation rates in astronomy.” ASP Conference Series, 461, 763-766. https://scixplorer.org/abs/2012ASPC..461..763H/abstract

Hogan, A., Blomqvist, E., Cochez, M., et al. (2020) "Knowledge Graphs." arXiv:2003.02320

INSPIRE project (2022). “Connecting ORCID to your INSPIRE author profile.” INSPIRE help pages. https://help.inspirehep.net/knowledge-base/connect_orcid_author_profile/

J. Paul Getty Trust, The. (2017). Getty Thesaurus of Geographic Names. https://www.getty.edu/research/tools/vocabularies/tgn/index.html

King, D. & Feldman S. (2021) “S2AND: An Improved Author Disambiguation System for Semantic Scholar.” AI2 Blog. https://blog.allenai.org/s2and-an-improved-author-disambiguation-system-for-semantic-scholar-d09380da30e6

Koch, J., Shapurian, G., Grant, C., Thompson, D. (2022) “ADS Docmatcher.” ADS Blog. https://scixplorer.org/scixblog/docmatcher

Kurtz, M. J. (2020). “NASA and Open Access Publishing,” ADS Blog. https://www.scixplorer.org/scixblog/nasa-open-access

Kurtz, M. J., et al. (1993) “Intelligent Text Retrieval in the NASA Astrophysics Data System” Astronomical Data Analysis Software and Systems II, A.S.P. Conference Series, Vol. 52, p. 132. https://scixplorer.org/abs/1993ASPC...52..132K/abstract

Kurtz, M. J., Accomazzi, A., & Henneken, E. (2018) “Merging the Astrophysics and Planetary Science Information Systems.” https://scixplorer.org/abs/2018arXiv180303598K/abstract

Kurtz, M. J., Accomazzi, A., & Henneken, E. (2021) “Enabling Synergy: Improving the Information Infrastructure for Planetary Science” Bulletin of the American Astronomical Society, Vol. 53, Issue 4, e-id. 470 https://scixplorer.org/abs/2021BAAS...53d.470K/abstract

Kurtz, M. J., et al. (2024) “Institution Based Changes in Astronomy Research 1997-2023.” Astrophysical Journal Supplement Series, in preparation.

Kurtz, M. J. & Accomazzi, A. (2019) “From Dark Energy to Exolife: Improving the Digital Information Infrastructure for Astrophysics” Astro2020: Decadal Survey on Astronomy and Astrophysics, science white papers, no. 17; Bulletin of the American Astronomical Society, Vol. 51, Issue 3, id. 17 https://scixplorer.org/abs/2019BAAS...51c..17K/abstract

Kurtz, M. J. & Henneken, E. (2018) “ Citations to Astronomy Journals 1: The growth of interdisciplinarity,” ADS Blog. https://ui.adsabs.harvard.edu/blog/citations-journals

Kurtz, M. J. & Bollen, J. (2010) “Usage Bibliometrics" Annual Reviews of Information Science abd Technology, v. 44, p 3-64.

Lockhart, K (2021) “Introducing ADS’s OpenAPI Description and Documentation,” ADS Blog. https://www.scixplorer.org/scixblog/openapi-docs

Mazzarella, J.M. & NED Team. (2017) “Evolution of the NASA/IPAC Extragalactic Database (NED) into a Data Mining Discovery Engine.” Astroinformatics, 325, 379. https://doi.org/10.1017/S1743921316013132

McGibbney, L. J., Whitehead, B., Rueda-Velásquez, C. A., Duerr, R., Keil, J. M., Berg-Cross, G., Rose, K., et al. (2022). “Semantic Web for Earth and Environmental Terminology (SWEET)” (Version 3.5). http://sweetontology.net

Mulcahy, C. (2017) “The Mathematics Genealogy Project Comes of Age at Twenty-one,” Notices of the American Mathematical Society, v. 64, p.466-470. https://www.ams.org/publications/journals/notices/201705/rnoti-p466.pdf

NASA Science Mission Directorate (2019) “Strategy for Data Management and Computing for Groundbreaking Science 2019-2024.” NASA Whitepaper. https://smd-cms.nasa.gov/wp-content/uploads/2023/05/SDMWGStrategy_Final.pdf

Nguyen, T. D., Ting, Y.-S., Ciucă, I., O'Neill, C., Sun, Z.-C., Jabłońska, M., Kruk, S., et al. (2023). “AstroLLaMA: Towards Specialized Foundation Models in Astronomy.” arXiv:2309.06126

OpenAlex project (2023). “Author Disambiguation,” OpenAlex API Documentation. https://docs.openalex.org/api-entities/authors/author-disambiguation

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2023) “Unifying Large Language Models and Knowledge Graphs: A Roadmap.” arXiv:2306.08302

Parsons, M. A., Duerr, R., & Godøy, Ø. (2023) “The Evolution of a Geoscience Standard: An Instructive Tale of Science Keyword Development and Adoption.” Geoscience Frontiers 14 (5): 101400. https://doi.org/10.1016/j.gsf.2022.101400.

Pennington, J., Socher, R., & Manning, C. D. (2014) “GloVe: Global Vectors for Word Representation.” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.3115/v1/D14-1162

Peek, J., V. Desai, R. L. White, R. D'Abrusco, J. M. Mazzarella, C. Grant, J. Novacescu, E. Scire, and S. Winkelman (2019) “Robust Archives Maximize Scientific Accessibility.” Bulletin of the American Astronomical Society, 51, 105. arXiv:1907.06234

Raskin, R. G., & Pan, M. J. (2005) “Knowledge Representation in the Semantic Web for Earth and Environmental Terminology (SWEET).” Computers & Geosciences, Application of XML in the Geosciences, 31 (9): 1119–25. https://doi.org/10.1016/j.cageo.2004.12.004

Rebull, L. M., Desai, V., Teplitz, H., Groom, S., Akeson, R., Berriman, G. B., Helou, G., Imel, D., Mazzarella, J. M., Accomazzi, A., McGlynn, T., Smale, A., & White, R. (2017) “NASA's Long-Term Astrophysics Data Archives.” arXiv:1709.09566

Shapurian, G., Kurtz, M. J., & Accomazzi, A. (2023) “Identifying Planetary Names in Astronomy Papers: A Multi-Step Approach.” arXiv:2312.08579

Shepherd, A., Jones, M. B., Richard, S., Jarboe, N., Vieglais, D., Fils, D., Duerr, R., Verhey, C., Minch, M., Mecum, B., & Bentley., N. (2022). “Science-on-Schema.org v1.3.1.” Zenodo. https://doi.org/10.5281/zenodo.7872383

Tenn, J. S. (2016) “Introducing AstroGen: the Astronomy Genealogy Project.” Journal of Astronomical History and Heritage, 19, 298. arXiv.1612.08908

Templeton, M. And Grant, C. (2021) “Affiliation Data in ADS: A Work in Progress.” ADS Blog. https://www.scixplorer.org/scixblog/affils-update

Timmer, R. C., Scen Khoo, F., Mark, M., Scoczynski Ribeiro Martins, M., Berea, A., Renard, G., & Bugbee, K. (2023) “NASA Science Mission Directorate Knowledge Graph Discovery.” arXiv:2303.10871

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., et al. (2023). “LLaMA: Open and Efficient Foundation Language Models.” arXiv:2302.13971

van Noorden, R. (2015). “Interdisciplinary research by the numbers.” Nature, Volume 525, Issue 7569, pp. 306-307. https://doi.org/10.1038/525306a

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017) “Attention is all you need.” arXiv:1706.03762

Wenger, M., Ochsenbein, F., Egret, D., Dubois, P., Bonnarel, F., Borde, S., Genova, F., Jasniewicz, G., Laloë, S., Lesteven, S., & Monier, R. (2000) “The SIMBAD astronomical database. The CDS reference database for astronomical objects.” Astronomy and Astrophysics Supplement, v.143, p.9-22. https://scixplorer.org/abs/2000A&AS..143....9W/abstract

White, R. L., A. Accomazzi, G. B. Berriman, G. Fabbiano, B. F. Madore, J. M. Mazzarella, A. Rots, A. P. Smale, L. Storrie-Lombardi, & S. Winkelman (2009) “The High Impact of Astronomical Data Archives.” astro2010: The Astronomy and Astrophysics Decadal Survey, 2010, P64. https://scixplorer.org/abs/2009astro2010P..64W/abstract

Wilkenson, M. D., et al (2016) “The FAIR Guiding Principles for scientific data management and stewardship” Scientific Data, Volume 3, id. 160018. https://scixplorer.org/abs/2016NatSD...360018W/abstract

Zhang S., Roller S., Goyal N., Artetxe M., Chen M., Chen S., Dewan C., et al. (2022) “OPT: Open Pre-trained Transformer Language Models.” arXiv:2205.01068

Appendix A - Acronyms and Abbreviations

AAS	American Astronomical Society
ADS	Astrophysics Data System
AGU	American Geophysical Union
AWS	Amazon Web Services
AWS IAM	AWS Identity and Access Management
AHED	Astrobiology Resource Metadata Standard Keywords
AI	Artificial Intelligence
AP	Astrophysics
API	Application Programming Interface
ARC/SS	NASA Ames Space Science & Astrobiology
ASCL	Astrophysics Source Code Library
Astromat	NASA Astromaterials Data System
BPS	Biological and Physical Sciences
CRESST	Center for Research and Exploration in Space Science and Technology
CIA	Central Intelligence Agency
DAAC	Distributed Active Archive Center
DOE	Department of Energy
DOI	Digital Object Identifier
DPS	Division for Planetary Sciences
DR	Disaster Recovery
ES	Earth Sciences
ESA	European Space Agency
FAIR	Findable Accessible Interoperable Reusable
GCMD	Global Change Master Directory
GES DISC	NASA Goddard Earth Sciences Data & Info. Services Center
GSFC	NASA Goddard Space Flight Center
HEASARC	High Energy Astrophysics Science Archive Research Center
HP	Heliophysics
IAU	International Astronomical Union
IRSA	Infrared Science Archive
KG	Knowledge Graphs

LLM	Large Language Model
MAST	Barbara A. Mikulski Archive for Space Telescopes
ML	Machine Learning
NASA	National Aeronautics and Space Administration
NEA	NASA Exoplanet Archive
NED	NASA/IPAC Extragalactic Database
NER	Named Entity Recognition
NIH	National Institutes of Health
NLP	Natural Language Processing
NOAA	National Oceanic and Atmospheric Administration
NSIDC	National Snow & Ice Data Center
NTRS	NASA Technical Reports Server
OA	Open Access
ORCID	Open Researcher and Contributor ID
ORNL	Oak Ridge National Laboratory
OSSI	Open Source Science Initiative
OWL	Web Ontology Language
PDF	Portable Data Format
PDS	Planetary Data System
PMC	PubMed Central
PS	Planetary Science
UI	User Interface
USGS	US Geological Survey
UX	User Experience
ROR	Research Organization Registry
SciX	NASA Science Explorer
SciX22	SciX 2022 NASA Proposal
SCoPE	SMD Community of Practice for Education
SED	NASA Goddard Sciences and Exploration Directorate
SEDAC	NASA Socioeconomic Data and Applications Center
SIMBAD	Set of Identifications, Measurements, and Bibliography for Astro. Data
SMD	Science Mission Directorate
SOSO	Science on Schema.org
SPASE	Space Physics Archive Search and Extract

STI	NASA Scientific and Technical Information
STI/NTRS	NASA PubSpace
SWEET	Semantic Web for Earth and Environmental Terminology
TGN	Getty Thesaurus of Geographic Names
UAT	Unified Astronomy Thesaurus
WCAG	Web Content Accessibility Guidelines
WIESP	Workshop on Information Extraction from Scientific Papers