

ADS Expansion

Alberto Accomazzi and the ADS Team

ADS Users Group Meeting, 9-10 Nov. 2022



SMD Strategy for Data Management and Computing 2019-2024



Science Mission Directorate's
Strategy for Data Management and Computing for Groundbreaking Science 2019-2024

Prepared by the Strategic Data Management Working Group

Approved by:

A blue ink signature of Thomas H. Zurbuchen.

12/17/18

Thomas H. Zurbuchen, Ph.D.
Associate Administrator,
Science Mission Directorate

Vision: To enable **transformational open science** through continuous evolution of science data and computing systems for NASA's Science Mission Directorate.

Mission:

- Lead an **innovative and sustainable program** supporting NASA's unique science missions with academic, international and commercial partners to **enable groundbreaking discoveries with open science.**
- **Continually evolve systems** to ensure they are usable and support the latest analysis techniques while protecting scientific integrity.

Goal 1: Develop and Implement Capabilities to Enable Open Science

Goal 2: Continuous Evolution of Data and Computing Systems

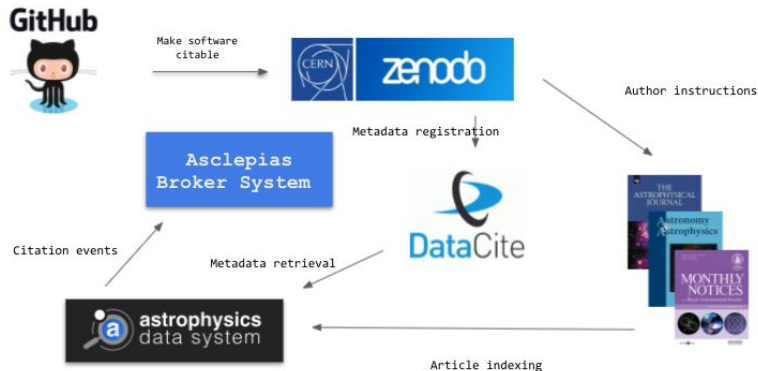
Goal 3: Harness the Community and Strategic Partnerships for Innovation

SMD Strategy for Data Management and Computing 2019-2024

Goal 1: Develop and Implement Capabilities to Enable Open Science		Goal 2: Continuous Evolution of Data and Computing Systems		Goal 3: Harness the Community and Strategic Partnerships for Innovation	
1.1	Develop and implement a consistent open data and software policy tailored for SMD	2.1	Establish standardized approaches for all new missions and sponsored research that encourage the adoption of advanced techniques	3.1	Develop community of practice and standards group
1.2	Upgrade capabilities at existing archives to support machine readable data access using open formats and data services	2.2	Integrate investment decisions in High-End Computing with the strategic needs of the research communities	3.2	Partner with academic, commercial, governmental and international organizations
1.3	Develop and implement a SMD data catalog to support discovery and access to complex scientific data across divisions	2.3	Invest in capabilities to use commercial cloud environments for open science	3.3	Promote opportunities for continuous learning as the field evolves through collaboration
1.4	Increase transparency into how science data are being used through a free and open unified journal server	2.4	Invest in the tools and training necessary to enable breakthrough science through application of AI/ML		

Why ADS - Support for Open Science

- ADS Facilitates discovery and dissemination of Open Access publications by aggregating and linking to OA versions
- ADS indexes and exposes links to data repositories
- ADS integrates astronomical object search and access
- ADS Indexes software records from the Astrophysics Source Code Library and software records cited via DOIs



SIMBAD OBJECTS	
<input checked="" type="checkbox"/> Other	19
<input type="checkbox"/> K2-18b	19
<input type="checkbox"/> K2-3b	7
<input type="checkbox"/> K2-3d	6
<input type="checkbox"/> K2-3c	5
<input type="checkbox"/> K2-9b	5
more	
<input type="checkbox"/> Star	18
<input type="checkbox"/> Galaxy	1
<input type="checkbox"/> Nebula	1

FULL TEXT SOURCES

- My Institution
- Publisher
- arXiv

DATA PRODUCTS

SIMBAD (8)	NED (3)
MAST (1)	IRSA (1)
Gemini (1)	ESA (1)
Chandra (1)	CDS (1)

[Add paper to library](#)

GRAPHICS



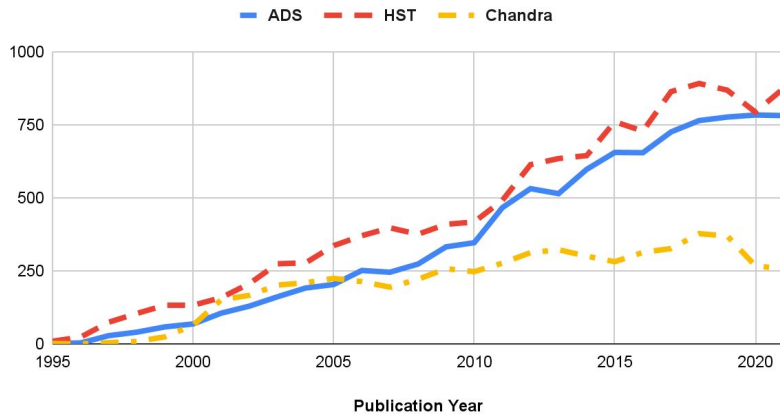
ASSOCIATED WORKS (2)

- [Catalog Description](#)
- [Source Paper](#)

Why ADS - Impact

ADS is recognized as one of NASA's most successful initiatives to support research in Astrophysics. Other NASA disciplines have now taken notice.

Acknowledgments in the literature



SLA PAM Division (2001): development of ADS
“represents an unparalleled shift in the propagation of the literature of astronomy”

AAS (2001): “ADS has revolutionized for over a decade the speed and thoroughness in which astronomers now can search and access the vast and still growing technical literature.”

NAS (2001): ADS “has vastly increased the accessibility of the scientific literature for astronomers.”

CfA Visiting Committee (2002): “ADS is probably the most valuable single contribution to astronomy research that the CfA has made in its lifetime.”

NASA AP (2008): “ADS is so extensively used by the entire professional astronomy community that it is hard to imagine existing without it”

The NASA Science Explorer (SciX)

NASA SciX will be a literature-based, open digital information system covering and unifying the fields of Astrophysics, Planetary Science, Heliophysics, and Earth Science. It will also cover NASA funded research in Biological and Physical Sciences.

NASA SciX will combine a scalable, discipline-agnostic core with a set of discipline specific knowledge centers which will curate and enrich its content using the deep subject matter expertise which has been crucial to the success of the ADS.

The screenshot shows the NASA Science Explorer (SciX) website. At the top, there is a navigation bar with the SciX logo, a dropdown menu for "General Science", and links for "ORCID", "About", and "Account". Below the navigation bar is a header with the SciX logo and the text "NASA Science Explorer". The main content area features a search bar with a "QUICK FIELD:" dropdown menu containing options like "author", "first author", "abstract", "year", and "fulltext". A search input field with a "Search..." placeholder and a search button is also present. Below the search bar is a "Search Examples" section with a grid of search terms and their corresponding field names.

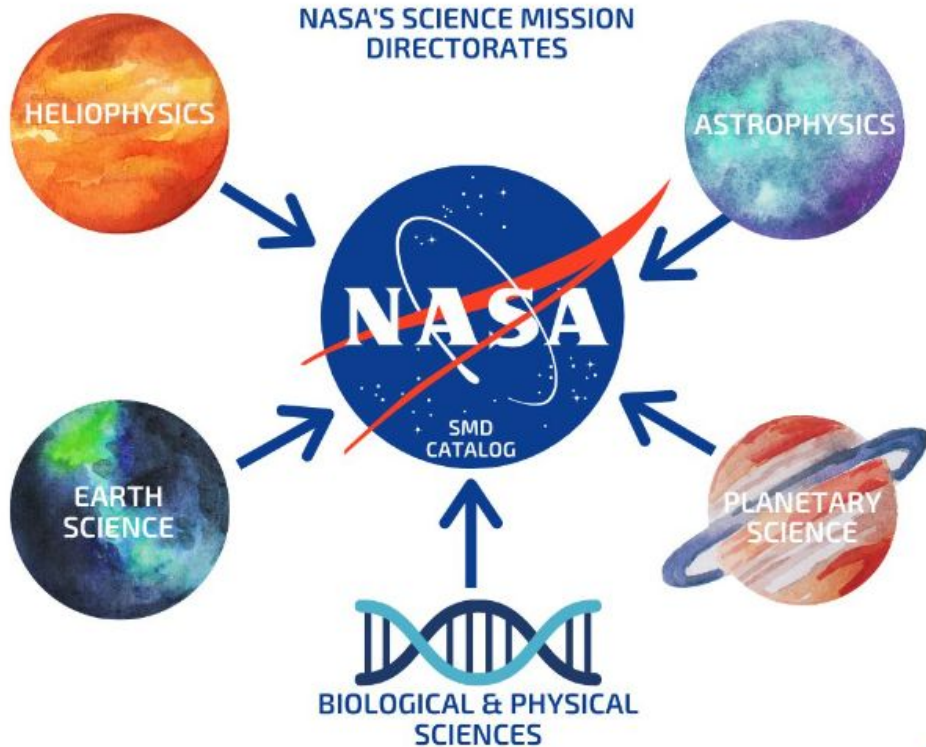
Search Examples			
author	author:"penrose, roger"	citations	citations(abstract:JWST)
first author	author:"^penrose, roger"	refereed	property:refereed
abstract+title	abs:"black hole"	collection	collection:astronomy
year	year:2000	exact search	=body:"reproducibility"
year range	year:2000-2005	institution	inst:NASA
full text	full:"black hole"	record type	doctype:software
publication	bibstem:ApJ		

Scope of the Expansion

ADS will extend its coverage and services into the disciplines represented in NASA's Science Mission Directorate (SMD).

Note: ADS was originally designed to serve NASA Astrophysics, but has already expanded its coverage to Planetary Science and Heliophysics.

Of all the disciplines in SMD, Earth Science is by far the largest and the one which will require the most attention on our part.



Assumptions and Context

Expansion proposal relies on the following observations and assumptions:

1. Preliminary ADS request was for an effort that would double our budget; NASA guideline funding is 15% below this level, which lead to slower system development
2. We can leverage economies of scale and current ADS infrastructure to scale up; we assume a gain of 2x efficiency in data ingest and curation for mainstream content
3. Disciplinary knowledge is required for the expansion but data science, AI/ML knowledge and staff are already present within the team and being developed further
4. Expansion is part of larger NASA Open Science effort, which includes mandates for opening access to articles and publishing data/software, all of which will benefit SciX
5. We will benefit from a growing ecosystem of scholarly initiatives and data brokers (e.g. CrossRef and Datacite) which make it easier for us to get relevant content
6. The system we are building is here to stay and will need indefinite support from NASA

SciX: a Digital Library for NASA Science

Goal is to build a discovery platform in which:

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 in a platform that is designed to scale without sacrificing accuracy and flexibility;
4. Digital collections are enriched with links to other research objects such as datasets, software, notebooks, and funding information;
5. Discipline-specific capabilities and analytic services are exposed to the relevant research communities;
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.

What the Expansion Entails

- Content Selection: complete coverage of PS, HP, add ES and NASA BPS
 - We estimate to grow by a factor of 2.5 in records indexed
 - We will ingest records for data products
- Partnerships and Collection Development: develop collaborations & outreach
 - Interface with NASA divisions, new publishers, and archives
 - Develop an outreach strategy in new disciplines
- Data Ingestion and Curation: improve efficiency
 - Improve ingest pipelines, text mining activities, ML-based metadata enrichment
 - Select collections of data products to index/link
 - Provide releases of Open Access content, training data, and models
- System Development and support: improve discovery
 - Improve Semantic Search capabilities (synonyms, taxonomies)
 - Improve author name disambiguation and ORCID integration
 - Develop UI enhancements tailored to specific disciplines

Semantic Search and the Disambiguation Challenge

Discipline-specific knowledge is used to improve discovery

- Synonyms - *exoplanet(s)*, *exoplanetary*, *extrasolar planet(s)*
- Acronyms - *ADS, Astrophysics Data System, Anti de Sitter Space*
- Taxonomies - *exoplanet astronomy > exoplanets > hot jupiters*
- Integration of data services - *object name normalization and indexing*


SciX will extend this approach to new disciplines

The image shows a search interface with a dropdown menu for the term 'jupiter'. The dropdown list includes the following items: 'Hot Jupiters', 'Epistellar jovians (Hot Jupiters)', 'Pegasean planets (Hot Jupiters)', 'Pegasids (Hot Jupiters)', 'Roaster planets (Hot Jupiters)', 'Moons of Jupiter (Jovian satellites)', 'Jupiter's satellites (Jovian satellites)', 'Jupiter's moons (Jovian satellites)', 'Jupiter', and 'Jupiter troians'. Below this, there is a SciX search bar with a dropdown menu for 'Earth Science'. The SciX search bar also includes a magnifying glass icon and the text 'SciX'. The dropdown menu for 'Earth Science' lists the following categories: 'General Science', 'Astrophysics', 'Heliophysics', 'Planetary Science', 'Earth Science' (highlighted), and 'Biological & Physical Science'. In the background, there is a NASA logo and the text 'author abstract'.

Data Releases and ML Support

ADS and SciX will create and release multiple data products for the SMD disciplines they cover

- An Open Corpus collection (metadata and full-text)
- An Open Annotated Dataset for use in M/L projects
- Discipline-specific language models (X-BERT)
- AI/ML Data Challenges, starting with the 2022 Workshop on Information Extraction from Scientific Publications

 WIESP 2022

WIESP @ ACL-IJCNLP 2022

The first Workshop on Information Extraction from Scientific Publications will be held at the [ACL-IJCNLP 2022](#) and it will feature:

- Paper Presentations
- Keynote talks
- Shared task presentations
- Panel discussion
- Invited paper presentations
- A virtual social cum poster presentation

Submission Site

Submission Link: <https://softconf.com/acl2022/WIESP/>

Shared Tasks

WIESP includes one shared task where we invite teams (individuals and groups) to come up with a system to tackle a Name Entity Recognition (NER) challenge:

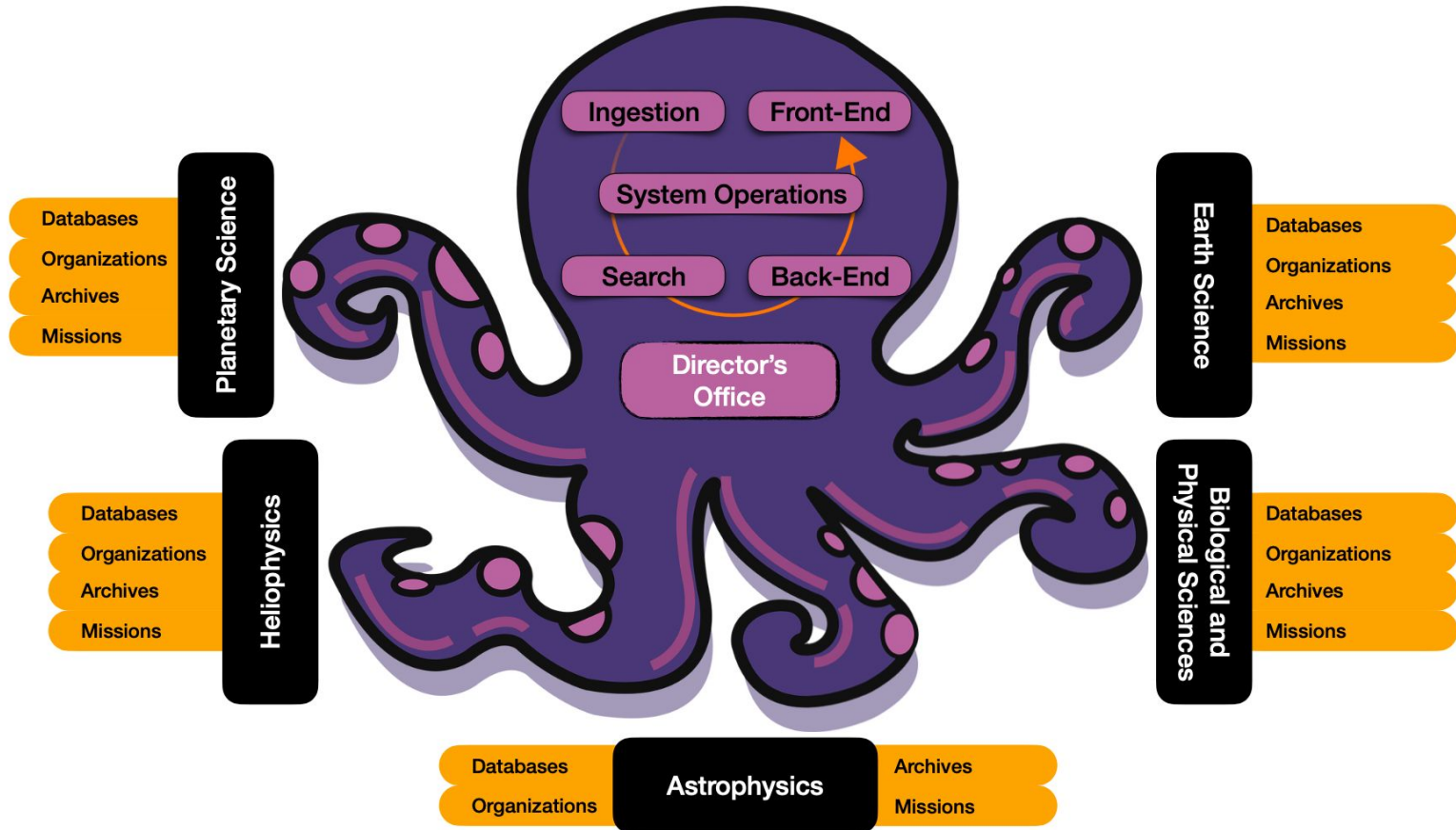
- [DEAL: Detecting Entities in the Astrophysics Literature](#)

Participants will have the opportunity to present their findings during the workshop and write a short paper. The best performer or interesting approaches might be invited to further collaborate with the [NASA Astrophysical Data System](#).

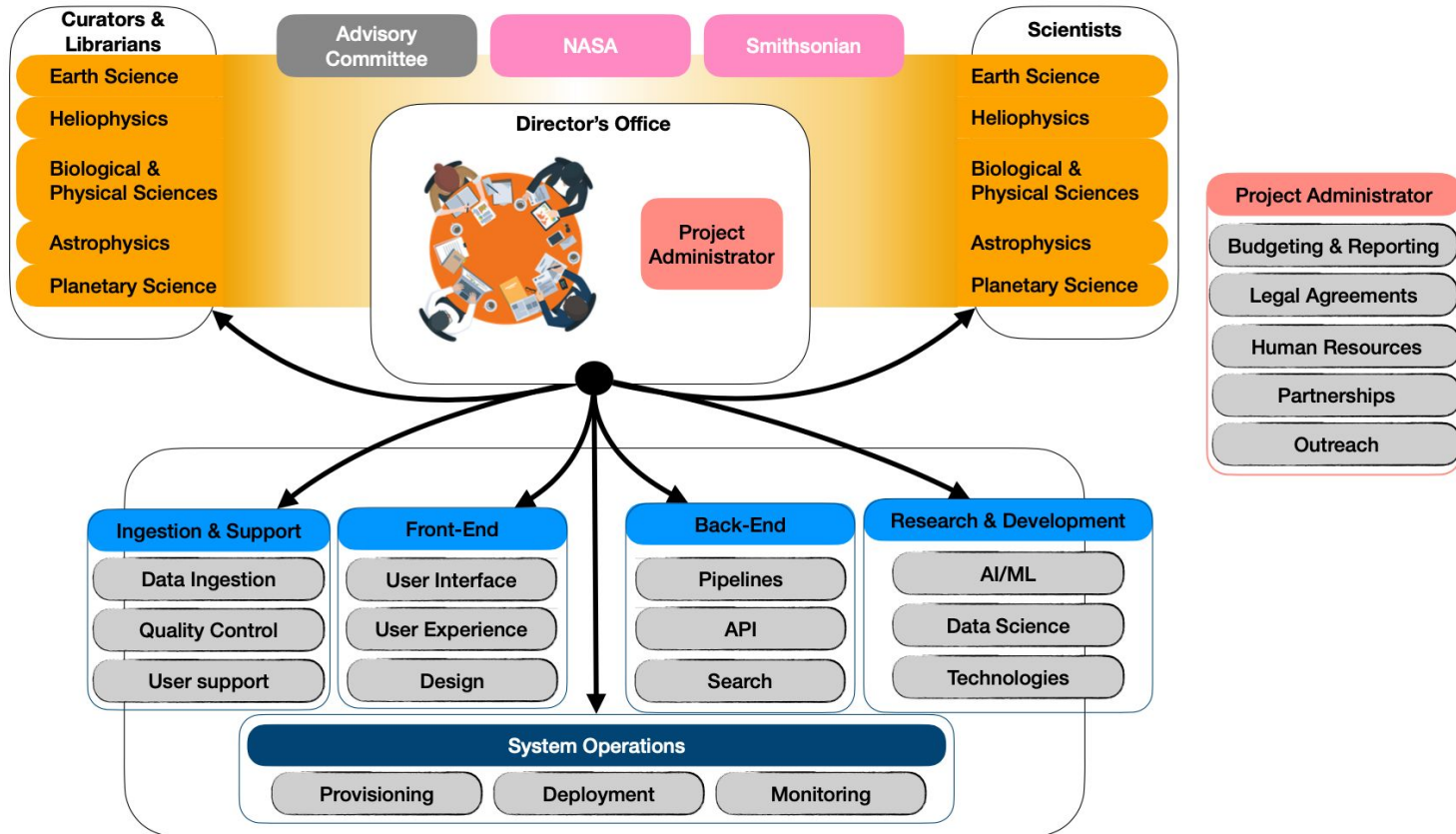
Major Milestones

When	What	How
Now - 2/23	Project restructuring, hire PS+ES Scientists, Admin, Developer Complete PS+HP refereed, census of ES, collaborate with SMD archives Set up Advisory Board for expanded scope	+4 FTEs
3/23 - 2/24	Hire ES curator, Project Scientist Ingest ES literature, preprints, NASA STI, launch SciX at AGU meeting Updated search, author name lookup, metadata enrichment, open datasets	+2 FTEs
3/24 - 2/25	Hire ES/BPS librarian, curator, devops engineer Census of BPS, ES gray literature, data indexing Metadata enrichment for ES, search author profiles, PS object search	+3 FTEs
3/25 - 2/26	Hire R&D developer, back-end developer (interoperability) Ingest BPS literature, cited ES literature, preprints Public author profiles, authorship suggestions, filtering via taxonomies, language models	+2 FTEs

Organizational Structure (high-level)



Organizational Structure (low-level)



Team Growth and Interactions with Community

- When fully staffed, the expansion adds 16 FTEs to the original ADS team, doubling our current size
- For most of the new positions we will offer the option of remote work, but will decide on a case-by-case basis, e.g. Project Scientist positions
- We have budgeted for 3x year team meetings in Cambridge. The meetings are for team building, program planning and execution
- We will have a new advisory group, with participants drawn from ES, BPS, and scholarly communication fields
- We will have some new conferences to attend to, with effort mostly lead by disciplinary scientists which will also lead Ambassador programs

Why are we doing this?

- To better support Open Science initiatives
- To have a larger impact to a larger research community
- To be true to the mission our institutions (Harvard, SI, NASA)

It's good for Science!

Why are we doing this?

- To better support Open Science initiatives
- To have a larger impact to a larger research community
- To be true to the mission our institutions (Harvard, SI, NASA)

It's good for Science!

- To raise visibility on our efforts worldwide
- To remain relevant in an age of increased access to information
- To thrive, not just survive

It's good for the Project!

Why are we doing this?

- To better support Open Science initiatives
- To have a larger impact to a larger research community
- To be true to the mission our institutions (Harvard, SI, NASA)

It's good for Science!

- To raise visibility on our efforts worldwide
- To remain relevant in an age of increased access to information
- To thrive, not just survive

It's good for the Project!

- Have more resources to invest in common infrastructure and functionality
- Better integration with a larger research community
- A more strategic role within NASA

It's good for Astronomy!

Feedback sought from ADSUG Panel

- Future of ADSUG meetings
 - Should we keep meetings virtual or should we consider in-person?
- Composition of expansion advisory group vs. ADSUG
 - Where should Planetary and Heliophysics go? What level of representation?
 - What level of cross-membership and communication is appropriate?
- Validation of organizational structure and staff management
 - Endorsement of flexible work conditions and schedules, including remote work
 - Advice on our envisioned organizational structure
- Support for selection of upcoming staff and advisory group members
 - We need help in filling the positions of scientists in our orgchart:
 - Planetary Science and Earth Science Deputies next in line
 - We need help selecting members of the expansion advisory group
- Revisit community outreach activities
 - Is #AstroTwitter still the place? Why is @adsabs engagement low?

Backup Slides

Feedback received from Expansion Review Panel

Expansion plan is ambitious given time frame and scale of disciplines

Engaging with different communities important, and financial support for Ambassador program should be considered

Members had several questions about how the system would implement features useful to the different disciplines, e.g.:

- Why is the proposed advisory group focused on Earth and Biological and Physical sciences rather than all five divisions?
- How will contributions from Ambassadors program, librarians, data stewards, be represented in the expanded ADS?
- How will discipline-specific portals will enable cross-disciplinary research?
- How will UI/UX design be conducted to meet the information behavioral needs of the different disciplines?

What ADS Indexes

- Literature: bibliographic metadata and full-text from multiple sources (arXiv and publishers) - 14.8M records
- High-level data products appearing in journal articles (mostly catalogs, data behind the plots) - 20K records
- Observing and Funding proposals - 46K records
- Curated collections (“bibliographic groups”) - 400K records
- Data links - 430K records
- Software entries from the Astrophysics Source Code Library (ASCL) and cited software products (Zenodo) - 15.6K records

Approach

- Bring together ADS's established curation model and new M/L efforts
- Obtain and index full-text for all the core literature
- Use citation analysis for detecting related/missing content, ingest corresponding metadata
- Use text mining for metadata enrichment efforts via Natural Language Processing techniques
- Leverage citation, usage for identifying connected research areas and promote interdisciplinary research

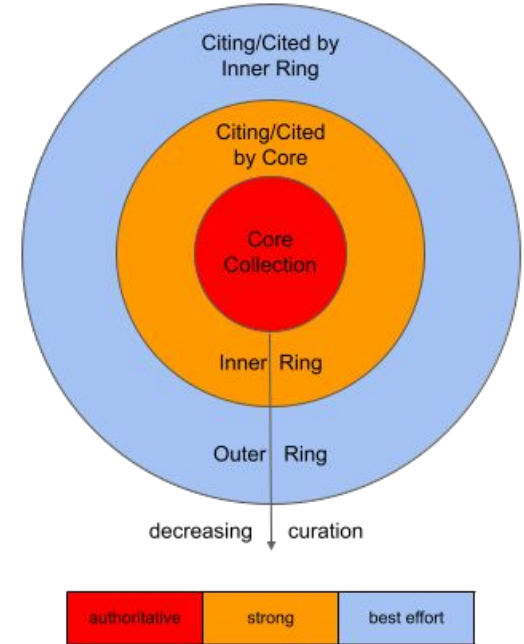


Figure 5. ADS's tiered curation model. The core collection represents disciplines where its curation is strongest and its coverage is authoritative. The surrounding tiers are connected to the core via the citation citation network.

Knowledge Management: Search Semantics

Discipline-specific knowledge is used to improve discovery

- Synonyms
- Acronyms
- Taxonomies
- Integration of data services

Knowledge Management: Search Semantics

Discipline-specific knowledge is used to improve discovery

- Synonyms - *exoplanet(s)*, *exoplanetary*, *extrasolar planet(s)*
- Acronyms
- Taxonomies
- Integration of data services

The screenshot shows a search interface with the query 'exoplanet' in a search bar. Below the search bar, three search results are displayed, each with a document icon, a list icon, and a menu icon. The first result is from 2012MNRAS.421.2498G, dated 2012/04, with 32 citations. The title is 'A lucky imaging multiplicity study of exoplanet host stars' by Ginski, C.; Mugrauer, M.; Seeliger, M. and 1 more. The abstract snippet reads: 'A lucky imaging multiplicity study of exoplanet host stars it is necessary to determine the fraction of multiple stellar systems amongst the known extrasolar planet'. The second result is from 2011EPJWC..1602004J, dated 2011/07. The title is 'The Calan-Hertfordshire extrasolar planet search' by Jenkins, J. S.; Jones, H. R. A.; Gozdziewski, K. and 7 more. The abstract snippet reads: 'The Calan-Hertfordshire extrasolar planet search The detailed study of the exoplanetary systems HD189733 and HD209458 has given rise to a wealth'. The third result is from 2011AAS...21821106D, dated 2011/05. The title is 'Validation and characterization of Kepler exoplanet candidates with Warm Spitzer' by Desert, Jean-Michel; Charbonneau, D.; Kepler Science Team. The abstract snippet reads: 'Validation and characterization of Kepler exoplanet candidates with Warm Spitzer Space Telescope to gather near-infrared photometric measurements of transiting extrasolar planet candidates detected'.

Knowledge Management: Search Semantics

Discipline-specific knowledge is used to improve discovery

- Synonyms - *exoplanet(s)*, *exoplanetary*, *extrasolar planet(s)*
- Acronyms - *ADS*, *Astrophysics Data System*, *Anti de Sitter Space*
- Taxonomies
- Integration of data services

The screenshot displays a search interface for the Astrophysics Data System (ADS). At the top, a search bar contains the text "ADS" and a magnifying glass icon. Below the search bar, three search results are listed, each with a checkbox, a document ID, a date, and a title. The first result is for document 2022PhyA.60427965H, dated 2022/10, with the title "Specified QoS based networked observer and PI controller design with disturbance and noise rejection under random packet dropout". The second result is for document 2022arXiv220904460N, dated 2022/09, with the title "Figure and Figure Caption Extraction for Mixed Raster and Vector PDFs: Digitization of Astronomical Literature with OCR Features". The third result is for document 2022arXiv220902709A, dated 2022/09, with the title "New recursions for tree-level correlators in (Anti) de Sitter space". Each result includes a snippet of text from the document, with the acronym "ADS" highlighted in blue. The interface also includes icons for document, list, and menu functions.

ADS

- 1 2022PhyA.60427965H 2022/10
Specified QoS based networked observer and PI controller design with disturbance and noise rejection under random packet dropout
Halder, Kaushik; Panda, Deepak Kumar; Das, Saptarshi and 2 more
has been derived using an asynchronous dynamical system (ADS) approach as a linear matrix inequality (LMI)
- 2 2022arXiv220904460N 2022/09
Figure and Figure Caption Extraction for Mixed Raster and Vector PDFs: Digitization of Astronomical Literature with OCR Features
Naiman, J. P.; Williams, Peter K. G.; Goodman, Alyssa
to the astrophysics literature holdings of the Astrophysics Data System (ADS), we find F1 scores of 90.9% (92.2%)
- 3 2022arXiv220902709A 2022/09
New recursions for tree-level correlators in (Anti) de Sitter space
Armstrong, Connor; Gomez, Humberto; Lipinski Jusinskas, Renann and 2 more
New recursions for tree-level correlators in (Anti) de Sitter space
We present for the first time classical multiparticle solutions in Anti de Sitter space (AdS)

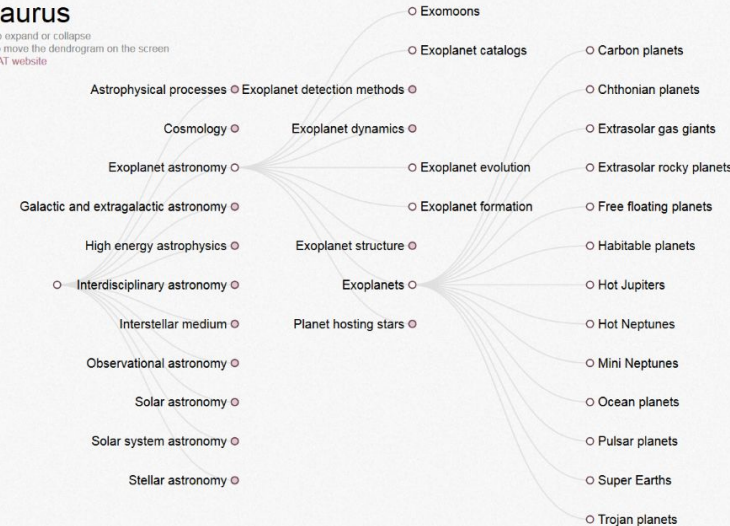
Knowledge Management: Search Semantics

Discipline-specific knowledge is used to improve discovery

- Synonyms - *exoplanet(s)*, *exoplanetary*, *extrasolar planet(s)*
- Acronyms - *ADS, Astrophysics Data System, Anti de Sitter Space*
- Taxonomies - *exoplanet astronomy > exoplanets > hot jupiters*
- Integration of data services

Unified Astronomy Thesaurus

click a node to expand or collapse
click & drag to move the dendrogram on the screen
back to the UAT website



Knowledge Management: Search Semantics

Discipline-specific knowledge is used to improve discovery

- Synonyms - *exoplanet(s)*, *exoplanetary*, *extrasolar planet(s)*
- Acronyms - *ADS*, *Astrophysics Data System*, *Anti de Sitter Space*
- Taxonomies - *exoplanet astronomy* > *exoplanets* > *hot jupiters*
- Integration of data services - *object name normalization and indexing*

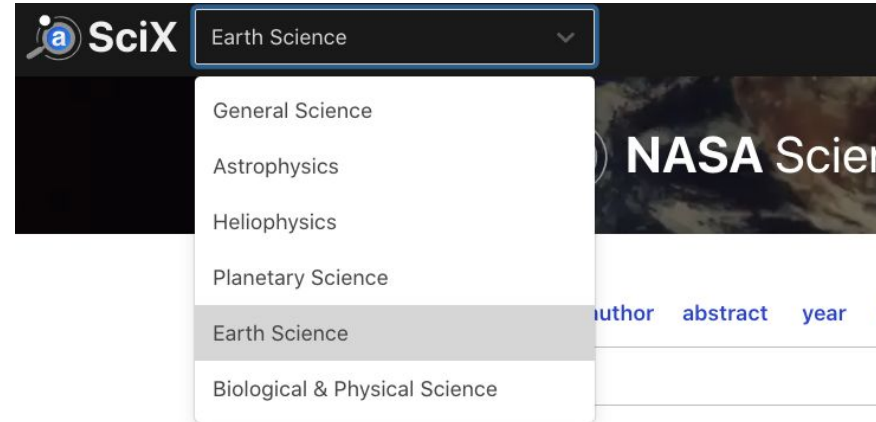
The screenshot shows a search interface with a search bar containing 'object:Andromeda' and a magnifying glass icon. Below the search bar are two panels of results. The left panel is titled 'SIMBAD OBJECTS' and shows a list of categories with their respective counts: Galaxy (12.1k), M 31 (11.8k), LMC (3.5k), Local Group (3.3k), M 33 (3.2k), and SMC (2.3k). A 'more' button is located at the bottom right of this panel. The right panel is titled 'NED OBJECTS' and shows a list of categories with their respective counts: Galaxy (7.9k), MESSIER 031 (4.5k), MESSIER 033 (1.3k), Large Magellanic Cloud (1.1k), MESSIER 081 (938), and Small Magellanic Cloud (789). A 'more' button is located at the bottom right of this panel.

Database	Category	Count
SIMBAD OBJECTS	Galaxy	12.1k
	M 31	11.8k
	LMC	3.5k
	Local Group	3.3k
	M 33	3.2k
	SMC	2.3k
NED OBJECTS	Galaxy	7.9k
	MESSIER 031	4.5k
	MESSIER 033	1.3k
	Large Magellanic Cloud	1.1k
	MESSIER 081	938
	Small Magellanic Cloud	789

Knowledge Management: Search Semantics

Discipline-specific knowledge is used to improve discovery

- Synonyms - *exoplanet(s), exoplanetary, extrasolar planet(s)*
- Acronyms - *ADS, Astrophysics Data System, Anti de Sitter Space*
- Taxonomies - *exoplanet astronomy > exoplanets > hot jupiters*
- Integration of data services - *object name normalization and indexing*



SciX will extend this approach to new disciplines

Knowledge Management: Data Indexing & Linking

ADS is indexing *some* data products:

- Astronomical data catalogs
- Cited software packages
- **New:** PDS datasets

ADS is linking papers to other datasets

- Curated “telescope” bibliographies
- **New:** links mined from DAS in papers

SciX will expand on these efforts, both via a bottom-up approach (text mining) and top-down (curated bibliographies).

The image shows two overlapping screenshots of the ADS (Astrophysics Data System) interface. The top-left screenshot displays search results for 'FULL TEXT SOURCES' and 'RELATED MATERIALS (14)'. The 'RELATED MATERIALS' section lists 'Software Source' with versions 0.8.0, 0.9.11, 0.9.12, 0.9.13, 0.9.14, 0.9.15, and 0.9.3. The bottom-right screenshot shows search results for 'FULL TEXT SOURCES' and 'DATA PRODUCTS'. The 'DATA PRODUCTS' section lists SIMBAD (8), MAST (1), Gemini (1), Chandra (1), NED (3), IRSA (1), ESA (1), and CDS (1). Both screenshots include a sidebar with 'My Institution' and 'Publisher' options, and a 'Full Text Sources' icon.

TBD for SciX: what records should be *indexed* vs. *linked*?

Indexing vs. Linking Research Data Products

Indexed (an actual database record, searchable)

- The scholarly literature of interest to Astronomers
- Peer reviewed catalogs, IVOA standards, observing and funding proposals
- Software products: ASCL records, software packages cited via DOI
- Soon: cited data products, other research objects such as notebooks

Indexed records are scholarly research objects.

They are discoverable and citable via ADS, and their metrics are tracked

Linked (resource accessible from a record via a link)

- Data Products hosted by external collaborators (Archives, SIMBAD, NED)

**Linked data collections can be used as a filter in ADS,
and to evaluate impact of linked data products**

ADS Expansion in PS & HP

Effort

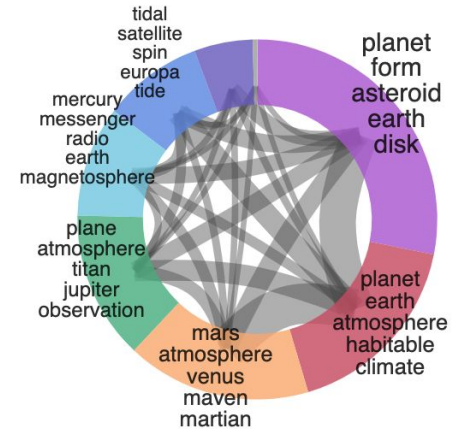
- Goal is for ADS to be as useful to PS and HP as it is to AP, providing completeness in coverage and richness of features
- Effort started 1 yr ago, ongoing

Literature

- 2,500 new journal articles
- 1,800 book series
- 2,500 Elsevier monographs
- 1,400 conferences/gray literature

Software and data products

- Added 600 datasets from PDS SBN
- Added data links to 5.6K AGU journal articles
- Added links to 480 software packages cited in 513 AGU articles



data:PDS jupiter

JGR Planets

Research Article | Full Access

Jupiter's Great Red Spot: Strong Interactions Incoming Anticyclones in 2019

A. Sánchez-Lavega, A. Anguiano-Arteaga, P. Iñurrigarro, E. García-Me Hueso, J. F. Sanz-Requena, S. Pérez-Hoyos, I. Mendikoa, M. Soria, ... See

First published: 17 March 2021 |

<https://doi-org.ezp-prod1.hul.harvard.edu/10.1029/2020J006686> | Citations: 1

SECTIONS

PDF

TOOLS

SHARE

FULL TEXT SOURCES

My Institution

Publisher

DATA PRODUCTS

Zenodo (3)

MAST (1)

ESA (1)

PDS (3)

Figshare (1)