ADS Expansion

Alberto Accomazzi, Michael J. Kurtz and the ADS Team

ADS Users Group Meeting, 15-16 Nov. 2021

Background

- Jan. 2018: ADSUG recommends expansion of ADS coverage to the entirety of the Exoplanet Literature and to study an expansion into Planetary Science
- Mar. 2018: ADS submits WP to Committee on Exoplanet Science Strategy on merging Astrophysics and Planetary Science information systems
- Nov. 2018: ADSUG "strongly encourages" NASA to find a way to fund Planetary Science expansion
- May 2019: ADS submits WP to Astro2020 Decadal Survey on interdisciplinary research
- Nov. 2019: ADSUG "strongly supports" Planetary Science expansion
- Dec. 2019: NASA SMD publishes "Strategy for Data Management and Computing for Groundbreaking Science 2019-2024," which includes a plan to provide unified access to SMD literature
- Sep. 2020: ADS submits WP to Planetary Science & Astrobiology Decadal Survey on Improving Information Infrastructure in Planetary Science
- Oct. 2020: NASA requests proposal for extending coverage to Planetary Science and Heliophysics
- May 2021: NASA funds Year 1 of Planetary and Heliophysics expansion, development of proposal for Earth Sciences and Biological and Physical Sciences

Context and Rationale

NASA SMD organized into 5 divisions:

- Astrophysics, Planetary Sciences, Heliophysics, Earth Sciences, Biosciences
- ADS has served the first one superbly, the next two adequately, last two poorly

Interdisciplinary research requires expertise across subject boundaries

- Exoplanets: Astrophysics, Planetary, Geophysics, Astrochemistry
- Multi-Messenger: Astrophysics, HEP, Computer Science, Instrumentation

Literature can be seen as central, organizing point to navigate research fields

- Big challenges require communities of experts from different fields
- Interdisciplinary research fields are connected by citations, topics

Support for Open Science goals

- Links between archives crucial for making data more discoverable and shared
- FAIR data and software principles require integration with literature

Goals of Further Expansion

- Development of a project plan for extending ADS following the SMD Strategy for data and computing over next 3-5 years
- Include publications from NTRS (NASA Technical Report Server)
- Index NASA awards in ADS and link to publications via DOIs
- Harvest and Index of NASA data sets, integrate in ADS via DOIs
- Include publications from all SMD Divisions: Planetary Science, Heliophysics, Earth Sciences, Biological & Physical Sciences
- Support for AI/ML initiatives using NASA publications
- Support/usage of cloud-based technologies

ADS Perspective - Goals (1/2)

- No “dumbing down”: The level of service for all SMD disciplines should be equivalent to the services provided to Astrophysics by the ADS, both currently and into the future
- Discovery Platform: we envision building and maintaining not just a simple search engine, rather a sophisticated cross-disciplinary discovery system. Integrating discipline-specific knowledge and resources is essential
- Character: maintain those features which have made the ADS successful. Key to this is the project’s personnel and culture. Finding individuals who fit in and who have a background in the SMD disciplines will be challenging

ADS Perspective - Goals (2/2)

- Community: ADS has enjoyed enormous support from its users; in order to be successful, the expansion needs to be seen as a NASA-wide effort to benefit the entirety of the research community supported by SMD
- Size: Astrophysics, Planetary and Heliophysics account for under 50% of publications in all five SMD disciplines; leveraging existing infrastructure and partnerships we should be able to achieve our goals through a doubling of ADS's current size
- Steps: A three year ramp-up time following a re-organization of the project structure is ambitious but should be doable, provided changes are made to properly absorb and manage new personnel

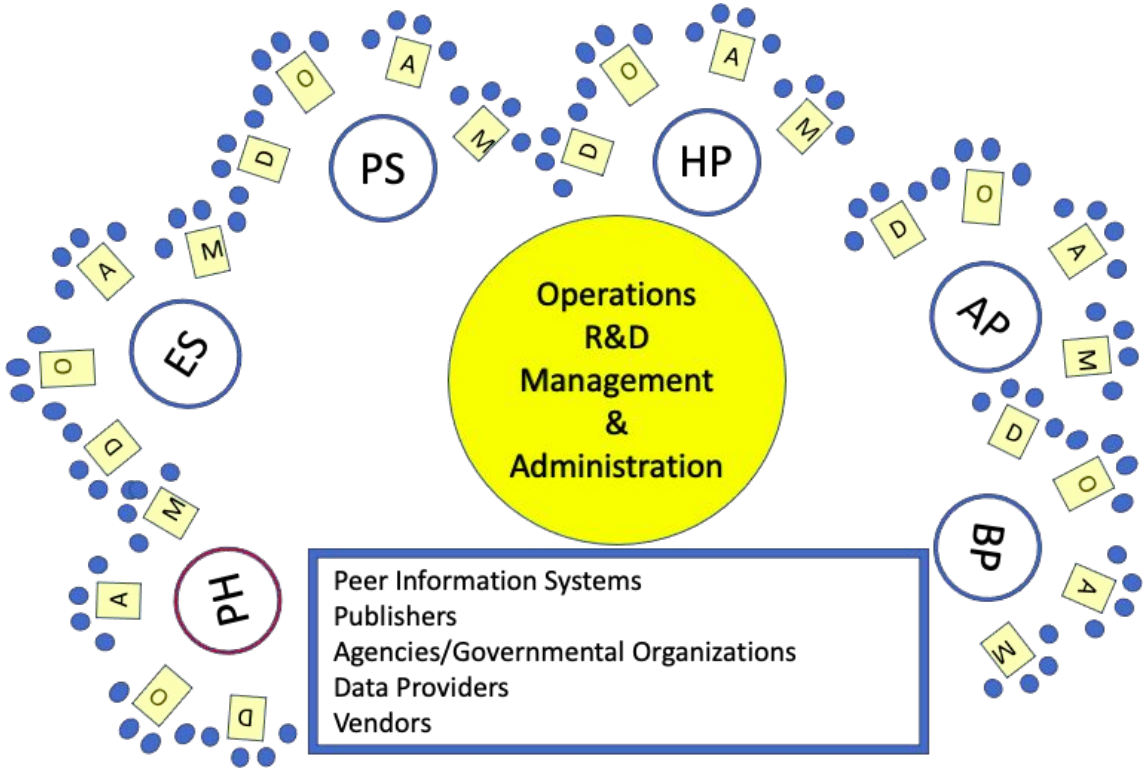
Proposed OrgChart

Core (operations and infrastructure):
- Management
- Operations
- R&D
- Administration
- Partnerships

Key personnel: PI, PS, PM, developers, engineers, curators

Nodes (discipline knowledge):
- Curation
- Interoperability
- User Support
- Outreach

Key personnel: discipline scientists and curators



D: databases
O: organizations
A: archives
M: missions

Roles and Responsibilities - Core

Management (Leadership)

- Principal Investigator: Strategy, Budget, Face of the Project
- Project Manager: Operations, interface with discipline curators
    - Deputy PM: Admin, Partnerships
- Project Scientist: R&D, Discipline scientists, Publications
    - System Architect: system design, IT R&D, operations

Operations

- Ingest/Workflows
- Search/API services
- Infrastructure/DevOps
- UI/UX
- User Support/Outreach

Research and Development

- Technology development (e.g. storage layer, search, workflows)
- Data Science research & publishing (Open Access/Open Source/Open data)
- Data & workflows integration

Administration

- HR and institutional interface
- Budgeting and Reporting
- MOUs and Legal agreements

Partnerships and Interoperability

- Peer Systems
- Publishers
- Agencies/Funders
- Data Providers
- Vendors

Roles and Responsibilities - Nodes

Discipline Scientists

- Act as a spokesperson for the project within the corresponding community
- Participate to relevant conferences, giving talks and demonstrations
- Perform research in the corresponding disciplines (articles, blogs)
- Work with development and operations to develop new and improved services

Discipline Curators

- Maintain discipline-specific knowledge bases supporting search and discovery
- Liaise with discipline specific archives and partners to create connections
- Work with core operation team to implement discipline-specific features

Organizational Changes

1. No longer a single Project Scientist who is an active researcher in Astrophysics, but rather a supervisory scientist working with a number of disciplinary scientists
2. Hiring of a Project Manager and Deputy PM to support system operations, administration and partnerships
3. Restructure organization into teams, with leads reporting to PM/PS
4. Discipline knowledge concentrated in nodes which interface with relevant information providers and provide requirements to core development team

What we need help/feedback on

1. Organizational structure
2. Recruiting talent
3. Running remote teams
4. Role of Astrophysics within larger organization
5. Risk Assessment

Backup Slides

The ADS Content Model

Core: Astrophysics

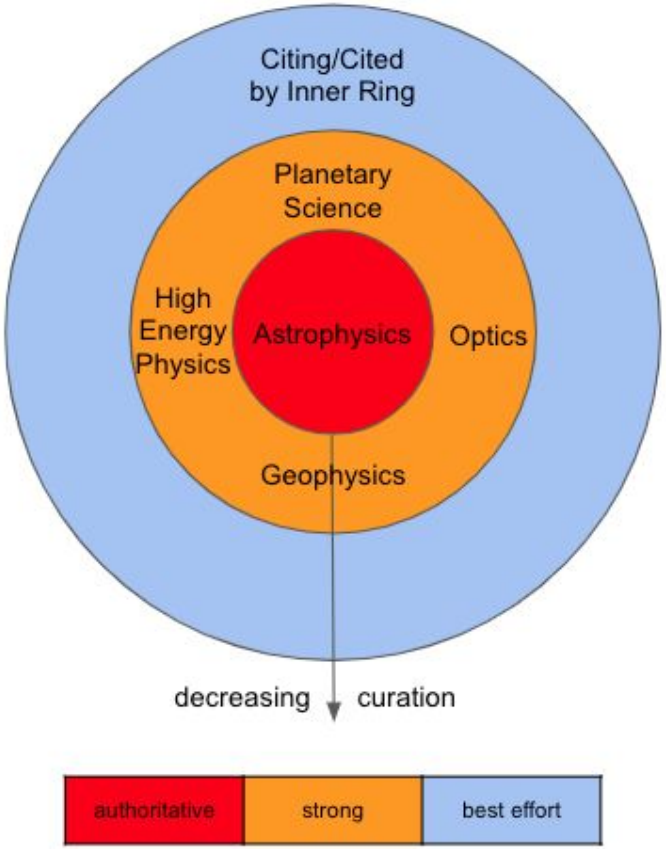
- Complete, authoritative coverage of the literature
- High-level data products and software indexed
- Links to datasets and archives (SIMBAD, NED)

Inner Ring: closely related disciplines

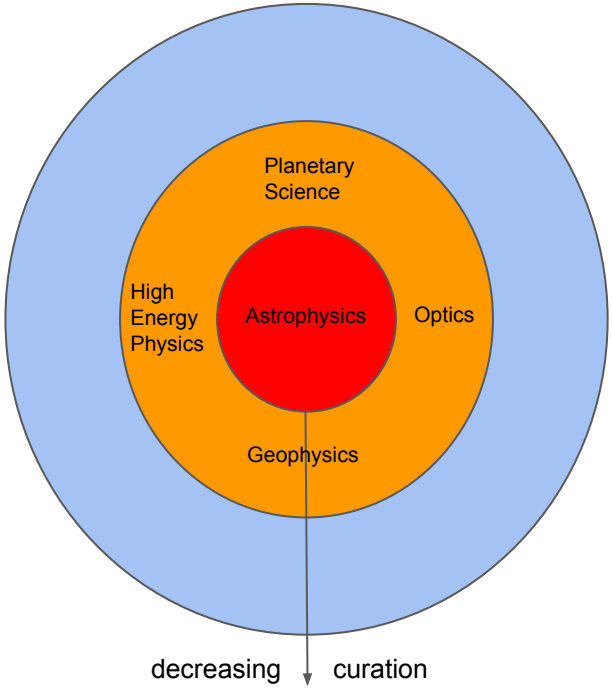
- Most refereed journals covered
- Some conf proceedings, some gray literature
- Citation and fulltext coverage incomplete

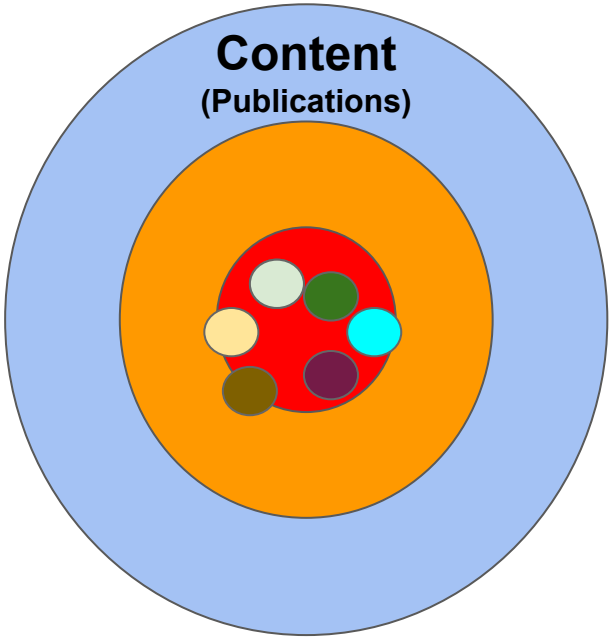
Outer Ring: connected to inner content

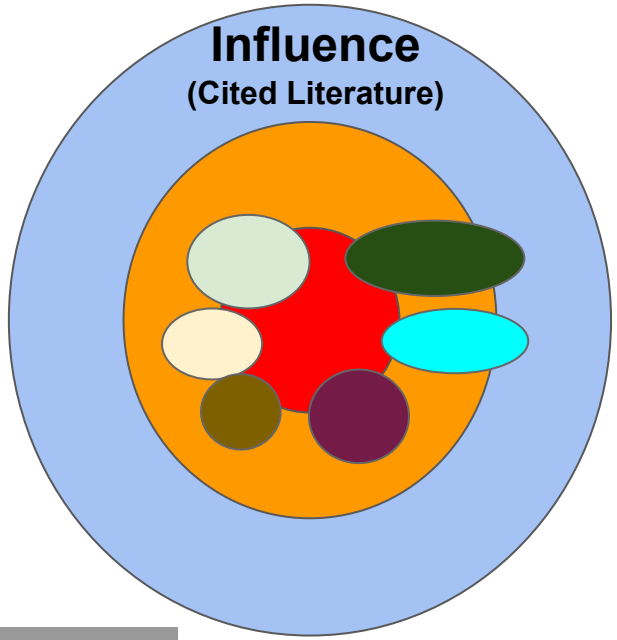
- Incomplete coverage of the literature because outside of disciplinary scope
- Includes content from multidisciplinary journals (e.g. Nature) or repositories (arXiv)
- No curation applied other than basic error corrections



A Closer look to sub-disciplines







Stellar dynamics
Laboratory Astrophysics
Astrobiology
Astrophotonics
Asteroids/Meteoroids
Planet Formation

Strategy

1. Identify relevant content which is not already in the system through a survey of existing online scholarly literature sources and user engagement at community meetings.
2. Perform in-depth analysis of disciplinary content in the system through citation and topic analysis in order to identify publications currently being missed from reference lists in a particular discipline. Prioritize list of content to be ingested based on previous step and source availability.
3. Identify additional partners and alternate data sources which should be incorporated in the database based on the analysis described above. Engage with the relevant parties when no existing agreement exists (e.g. American Chemical Society, Royal Chemical Society).
4. Evaluate relevant content, seeking the highest quality data source (ideally the peer reviewed paper provided by a publisher). For each publication, items of interest includes metadata, references, and full-text, which require partnership with publishers as the publicly available metadata from aggregators such as Crossref is not enough for research needs.

Content

1. Refereed Journal articles: typically the “core” contribution of a scholarly bibliographic system, refereed publications represent the high-impact research output in the field.
2. Gray literature: PhD theses, as well as non-refereed journals, conference proceedings, and meeting abstracts.
3. Eprints: currently the ADS indexes the arXiv e-print server at Cornell University and cross-matches these publications with the published literature; an expansion of scope means additional eprint archives such as ESSOAr (https://www.essoar.org/) will need to be integrated.
4. Data and Software: extend workflow which ingests software and data formally cited in the literature via a DOI to additional communities and disciplines
5. Other scholarly artifacts: NASA Awards proposals, Observing proposals, Astronomy data catalogs, and ASCL software entries are indexed in ADS. An SMD-wide approach to publishing awarded proposals and other scholarly products would increase the scope of the current effort

Development

1. Collection development: develop workflows and pipelines to integrate content from repositories such as NTRS, ESSOAr, EarthArXiv, and other disciplinary archives
2. Document classification: develop a robust methodology to classify records upon ingestion for new sources of content so that we can automatically assign new papers to a particular collection
3. Citation processing: improve the resolution of citations in the newly added literature
4. Text mining: implement a robust pipeline to identify, index and/or link data products mentioned in the available fulltext papers
5. Semantic search: develop lists of synonyms and acronyms in use by the SMD disciplines, and apply them to improve search results
6. UI enhancements: provide proper portal pages for searches tailored to a specific discipline
7. Entity recognition and linking: identify and extract from the text known entities corresponding to concepts, objects, and artifacts to uniquely identify relevant disciplinary knowledge

Infrastructure

1. Data model updates: adjust metadata model to account for features or relationships not currently implemented, such as taxonomies, new identifier schema, or disciplinary metadata
2. Legacy ingest workflows: accelerate the design and implementation of a new modern back-office ingestion pipeline to eliminate the dependency on legacy ADS classic ingestion process
3. Citation network: improve existing mechanisms that maintain the citation network required to provide ADS's advanced search features
4. Content indexing: speed up stages of data processing by moving source data and pipelines closer to the final data store, thus moving key pieces of ADS back-office pipelines to the cloud infrastructure
5. Disaster recovery: increase on-site data storage capacity to meet future needs and ensure proper redundancy, as well as expand and revisit cloud-based backup systems and recovery procedures.