ADS Expansion

Alberto Accomazzi and the ADS Team
@aaccomazzi

ADS Users Group Meeting, 19-20 Nov. 2020

Background (1/2)

- 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
- Apr. 2018: NASA asks estimate for an expansion into Planetary Science
- Aug. 2018: ADS and other NASA Archives participate in NASA Science Mission Directorate (SMD) Processing and Data Exploitation Meeting
- Nov. 2018: ADSUG “strongly encourages” NASA to consider way to fund Planetary Science expansion
- Apr. 2019: NASA asks estimate for an expansion into Heliophysics

Background (2/2)

- May 2019: ADS submits WP to Astro2020 Decadal Survey on improving infrastructure for 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: "*SMD should create a free and open, unified journal server along the lines of PubSpace, ADS or ERS to make science papers more accessible*"
- Sep. 2020: ADS submits WP to Planetary Science & Astrobiology Decadal Survey on Improving Information Infrastructure in Planetary Science
- Oct. 2020: NASA SMD contacts ADS to explore the possibility of extending the project to cover other divisions and including other NASA datasets

Context and Rationale

NASA SMD organized into 5 divisions:

- Astrophysics, Planetary Sciences, Heliophysics, Earth Sciences, Biosciences
- ADS serves the first one superbly, the next two adequately, the last two poorly
- No such distinctions in Europe, all “Astronomy” research funded together

Interdisciplinary research requires expertise across subject boundaries

- Exoplanets: Astrophysics, Planetary, Geophysics, Atmospheres
- Multi-Messenger: Astrophysics, HEP, Computer Science, Instrumentation
- NASA recognizes need to foster cross-disciplinary efforts

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

- Big challenges require communities of experts from different fields
- As interdisciplinary research develops, different fields become organically connected and discoverable through topics, citations, co-readership
- Links between archives crucial for making data more discoverable and shared

The ADS Content Model

Core: Astrophysics

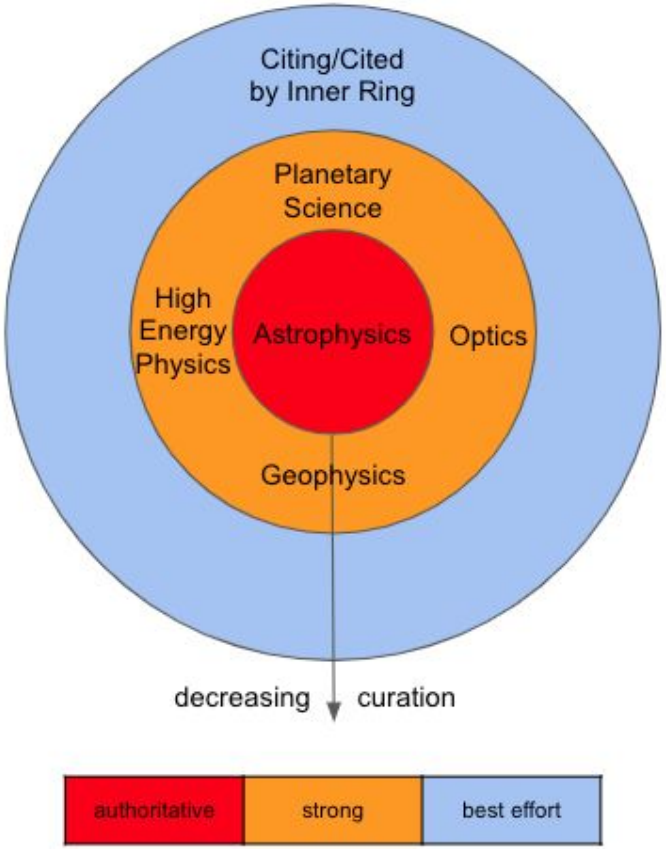
- Complete, authoritative coverage of the literature
- High-level data products and software indexed
- Links to datasets and archives (SIMBAD, NED)

Inner Ring: closely related disciplines

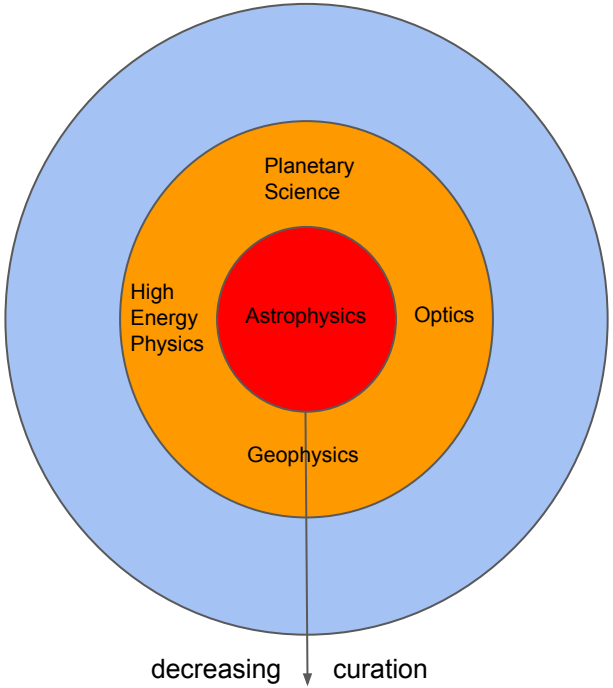
- Most refereed journals covered
- Some conf proceedings, some gray literature
- Citation and fulltext coverage incomplete

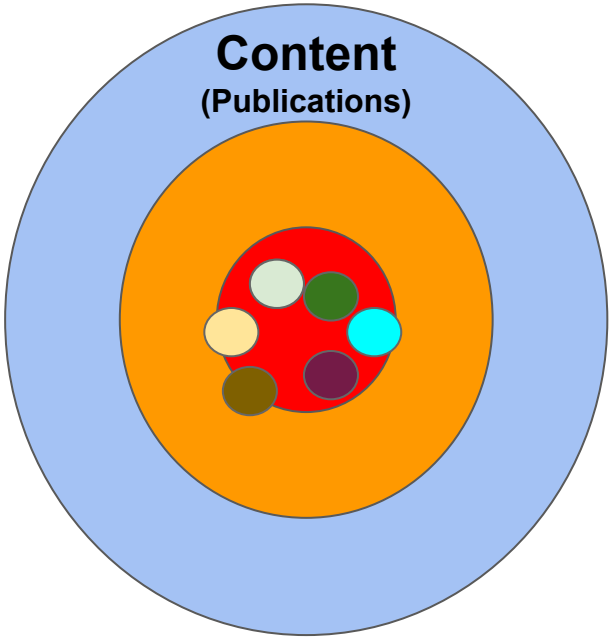
Outer Ring: connected to inner content

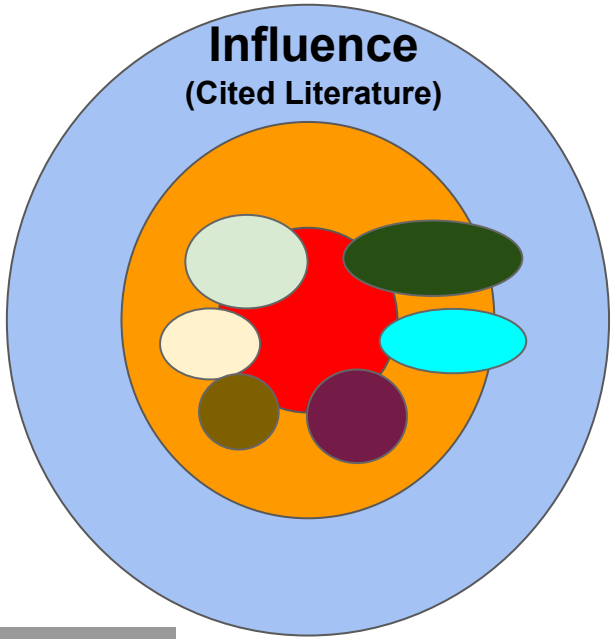
- Incomplete coverage of the literature because outside of disciplinary scope
- Includes content from multidisciplinary journals (e.g. Nature) or repositories (arXiv)
- No curation applied other than basic error corrections



A Closer look to sub-disciplines







authoritative	strong	best effort

Stellar dynamics
Laboratory Astrophysics
Astrobiology
Astrophotonics
Asteroids/Meteoroids
Planet Formation

NASA Exploratory Request from Oct. 2020

- Development of a project plan for extending ADS following the SMD Strategy for data and computing over next 3-5 years
- Implement integration of ADS data with PubSpace (NASA funded articles)
- 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 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 is 20% the size of SMD, yet we can leverage existing infrastructure and partnerships and may be able to achieve our goals through a doubling of ADS's current size (to be confirmed)
- Steps: A three year ramp-up time, 5 new FTEs per year (plus attrition) is ambitious, but may be doable via the current funding vehicle being used for ADS. Additional changes will be required to properly absorb and manage new personnel

ADS Perspective - Process

- Strategy: Identify stakeholders, define goals, develop a plan, forge alliances with data providers and initiatives
- Content: Identify what is in scope, prioritize its ingestion, apply appropriate levels of curation to it
- Development: Implement workflows and procedures to support ingest, curation, indexing and searching of new content
- Infrastructure: Enhance system to support new workflows, increased usage, interoperability with external resources and archives

ADS4SMD Proposal - Year 1

1. Expand coverage of Planetary Science and Heliophysics content in the ADS
2. Integrate content from PubSpace and NTRS in ADS, index and/or link data products identified via DOIs
3. Improve existing workflows and infrastructure required by the expansion of content in step 1
4. Develop a detailed project plan for extending the project over a 5 year period to cover additional disciplines in accordance with the SMD strategy for data and computing

Questions

1. What should ADS4SMD look like? What kind of scientific guidance should it have from the different disciplines?
2. What should be the structure of a larger team? What should be the relationship between Project Scientist(s), Project Manager(s) and the PI?
3. How can ADS preserve its internal culture and its successful relationships with the outside world?
4. What are the threats and risks for the project? How do we avoid or mitigate them?

Backup Slides

PubSpace

PubSpace is an archive of full-text journal articles produced by NASA-funded research and available online without a fee. PubSpace is available from a collaboration between the National Institutes of Health (NIH) and NASA to allow wider access to the results of federally-funded research. Articles collected under the Public Access Policy are archived on PubSpace, which is hosted by NIH on their PubMed Central (PMC) repository.

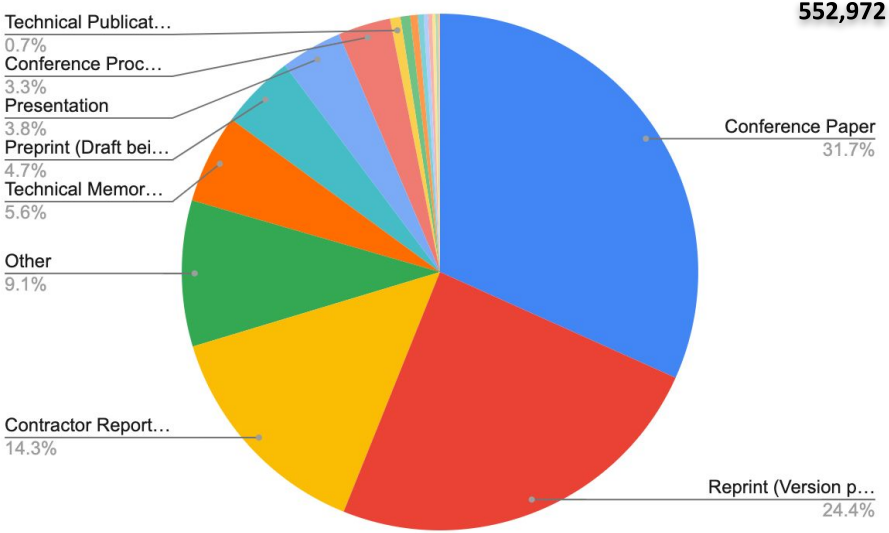
Access: OAI-PMH or Python module `PyMed`

9,310 records

NTRS

NASA's Technical Reports Server (NTRS) provides access to aerospace-related citations, full-text online documents, and images and videos. Content includes: conference papers, journal articles, meeting papers, patents, research reports, images, movies, and technical videos - scientific and technical information (STI) created or funded by NASA.

Access: API



552,972 records

NTRS and PubSpace have complementary collections. Many of the articles that are archived in PubSpace are also available from NTRS.

Note: With the overhaul of the Research Access program, at this time NTRS is not submitting any new intramural (authored by NASA civil servant or NASA contractors) publications to PubSpace. However, extramurals (NASA grantee authors) are still submitting theirs directly to PubSpace; not through NTRS.

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, PubSpace, 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.