Content & Curation

Edwin Henneken, Carolyn Grant, Matthew Templeton, Donna Thompson, Jenny Koch, Daniel Chivvis, Harry Blom and the ADS Team

ADS Users Group Meeting, 16-17 Nov. 2023

Content & Curation

- Content & Curation Roadmap Update (Edwin)
- Content update 2023 (Carolyn/Daniel)
  - Introducing the Earth Science collection
  - Journals update (selection & agreements)
  - Data links & citations (Edwin)
- Integration of NASA bibliographies (Jenny)
- Journals Database & Completion Statistics and Affiliations (Matthew)

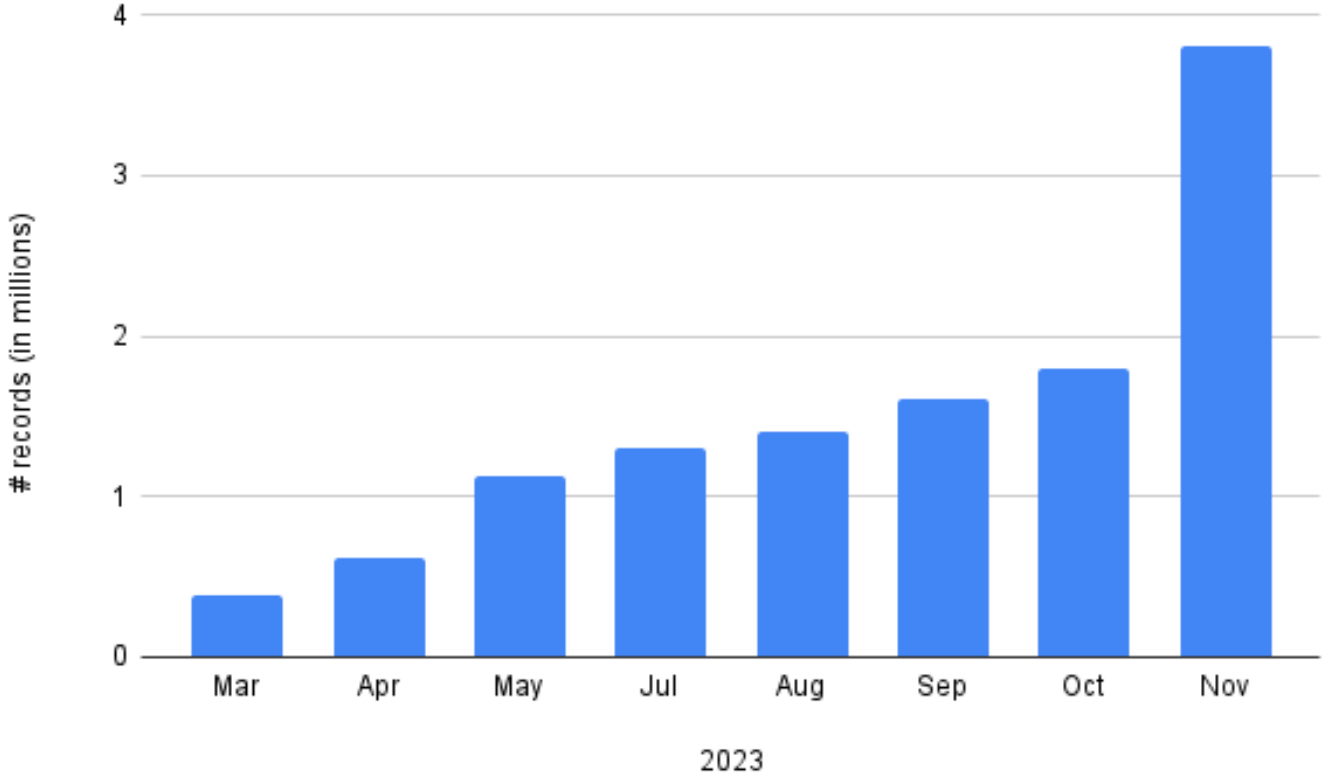
Content & Curation Roadmap

- Identify missing content (citation analysis, internal reporting tools, external bibliography databases)
- Harvest data from external bibliography databases to create ADS bibliographic groups (bibgroups) - e.g. NASA PubSpace
- Testing of the new Reference Pipeline (replacement for Classic reference matching workflow)
- Mining of full text for data links in Data Availability sections
- Use OpenAlex to improve our metadata coverage, like Open Access and licensing information
- New sources of preprints (ESSOAr, EarthArXiv)
- New preprint/journal article matching workflow
- Improve gray literature coverage by harvesting Harvard library catalog
- RFC writing and discussions for replacement of the ADS bibliographic identifier (bibcode)
- The ADS Journals Database

2023 Content update

- 10% growth in abstracts (20.3m) and citations (181m); 1.28m new fulltext records
- Weekly updates
  - 2,500 abstracts/week (AST) (up from 2100)
  - 12,000 abstracts/week (PHY) (up from 10,000)
  - 380,000 citations/week (up from 250,000)
- Software and Curated Datasets (PDS and JAXA)
  - ~3400 software records; < 100 dataset records
- Planetary/Heliophysics additions
  - MPEC weekly updates, EMAC collaboration, Heliophysics decadal papers
  - Updates to existing records (references and fulltext)
  - Planetary Feature Names Project
- Earth Science collection
  - 3.8 million since March 2023
  - ~2000 new journals

Earth Science Collection



Earth Science Collection - Journal Identification

- Initially, over 25,000 earth science journals were identified.
- Many challenges faced (e.g., messy data).
    - To address these issues, a new methodology was implemented:
        - Direct identification of journals from publisher data sources.
        - Verifying journal metadata via databases and indices.
            - Reliance on Elsevier's Scopus for bibliometric data and subject prioritization.

Subject Category Identification & ASJC Codes

- New earth science ontology created.
- All Science Journal Classification (ASJC) codes used to classify journals in earth science and related fields:
  - **Minimum Scope**: Basic earth and planetary sciences categories.
  - **Median Scope**: Expands to include agricultural, biological sciences.
  - **Maximum Scope**: Broadens further to encompass engineering, materials science, physics, chemistry, and other related categories.

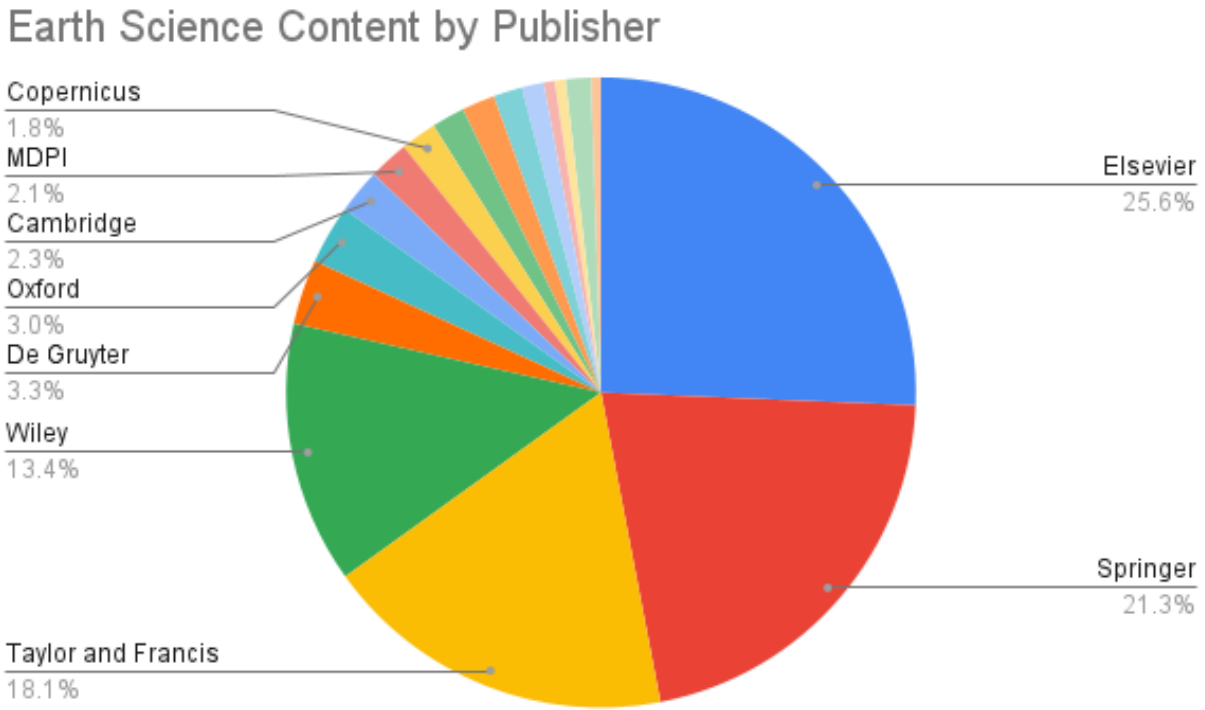
Earth Science Collection - Journals Update

- Journals classified as “minimum scope” from all major and minor publishers are set for curation.
- The curation process included verification and error-checking of the automatically identified journals.
- Newly curated “minimum-minimum” scope lists are enhanced with additional metadata and categories.

    - New “Minimum-Minimum” Scope Categories:
        - Environmental Science
        - Environmental Studies
        - Energy / Sustainability / Green Technologies
        - Ecology

Journals Update (cont.)

- ~2000 journals identified as minimum scope



Earth Science Collection - Ingesting the Data

- Publisher liaison reworking existing agreements and generating new agreements
  - Fulltext indexing legalities very fluid with surge in AI/ML
  - Some publishers asking for fee per journal
- CrossRef feed for journals not covered by those agreements
- Some new publisher feeds, especially from open source publishers (MDPI, IUCr, Pensoft)
- CrossRef data as placeholder until publisher content delivered

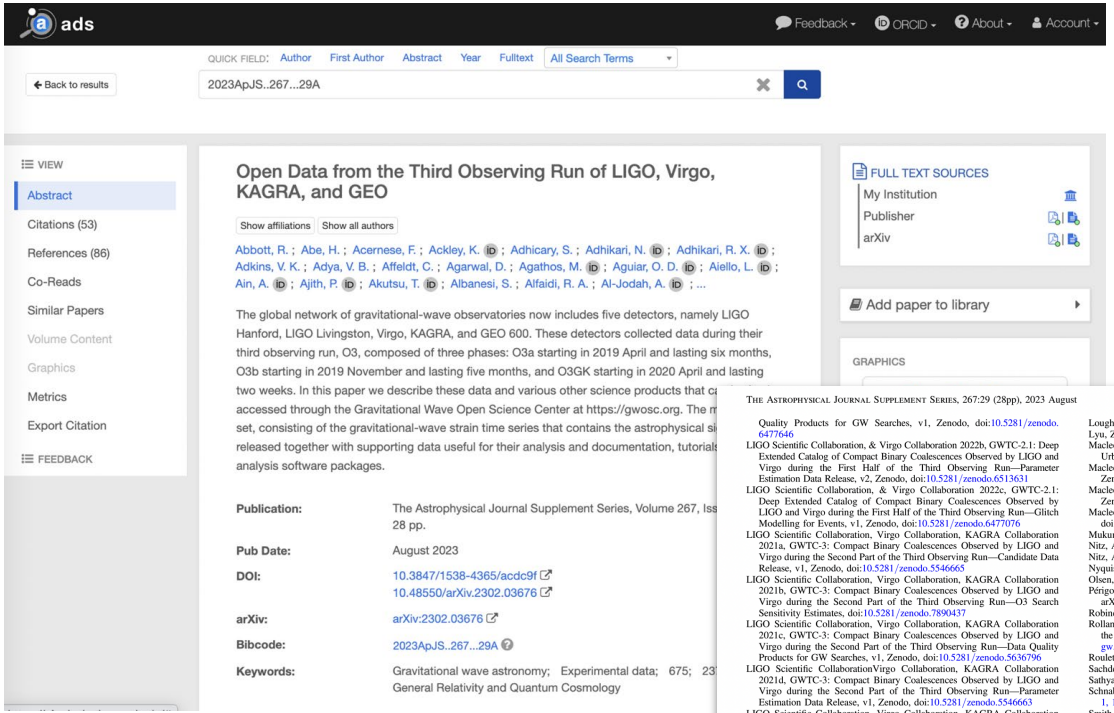
Earth Science Collection - next steps

- Update content with publisher data, including full text
- Review additional journals (earth scientist)
- Classify existing content (multi-disciplinary)
- Identify and acquire grey literature
- Use citation analysis to identify additional missing content
- Metadata enrichment (linking out to data, software, etc.)
- GeoRef comparison/analysis

Data links & citations

We are currently scanning the Data Availability Statement sections for 50 journals (e.g. all AGU, AAS and AMS journals). So far, 80k data links have been extracted.

Data links & citations



Based on DOIs found in our reference data files, we currently have ~125k citations for data sets (based on 30 DOI prefixes). These are currently not assigned, unless we happen to have a record (= mostly just PDS data sets)

Acceptance criteria for dataset indexing

Acceptance criteria for dataset indexing

General decision tree:

1. Publication contains a link to a dataset in reference section
    a. ADS already has a record? → create data link & assign citation
    b. No existing record in ADS:
        i. Is it a high-level/curated dataset? → **create record*** & create data link & assign citation
        ii. Otherwise ("reuse" criterion):
            1. Has the dataset been cited 2 times? → create/link/cite
            2. Record citation & link

* Descriptive metadata for a dataset has to meet criteria laid down by the ADS curation team. For one, descriptive metadata (title, abstract, keywords) must be sufficiently descriptive to be similar to basic metadata found in scholarly publications.

- We don't want to overwhelm our index with records for data sets
- This is something we need to collaborate on with Kaylin Bugbee's team (Science Discovery Engine), at least for NASA data

NASA Bibliographies

Goals:

- Establish collaborative relationships with NASA stakeholders to support increased use of ADS/SciX
- Connect NASA data products with ADS/SciX; learn what publications contain use of NASA data

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 and Information 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

Journals Database

Primary data store for journals data in ADS, to better enable us to interface with and collect metadata from a variety of sources

- Public API available since early 2023.
- Journals API integrated into backoffice data processing and analysis (indexing, completeness reporting, journal- and publisher-specific parsing pipelines)
- Curation is an ongoing process (CSG + curators)

Completeness data

- ADS holdings versus Crossref data
- 367 journals currently tracked, > 6.5M records
- Majority of ref journals are > 95% complete; astro typically > 98%
- Missing content: multidisciplinary (*Science*, *Nature*) some historical (e.g. *Nachrichten*)
- By-journal / by-volume statistics available from API (when we have it)

Affiliations

We assign identifiers to publisher-supplied affiliation strings
=> **facet searches on institutions**

- Refereed Astronomy > 95% matched
- Refereed Physics > 90% matched
- Refereed Earth Science > 70% matched (*without having done any new assignments*)

Backup slides

Data links - top 10

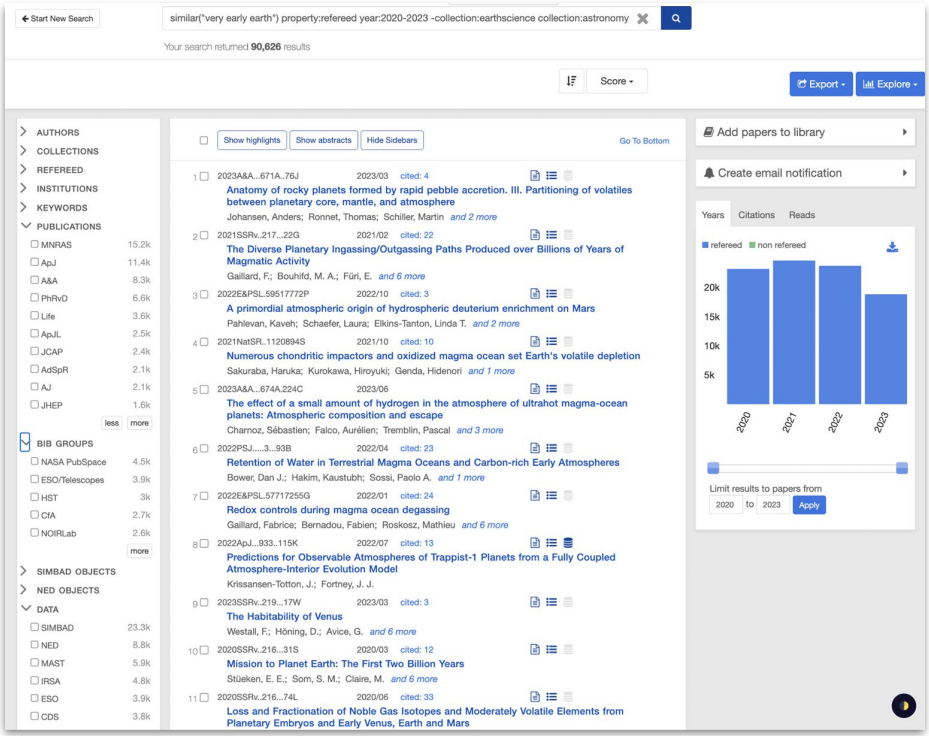
Source	Count
Geophysical Research Letters	15367
Nature	8817
JGR Atmospheres	7064
JGR Space Phys.	6179
JGR Solid Earth	4801
JGR Oceans	4580
Water Resources Research	3981
Journal of Climate	2745

Data citations - top 10

Source	Count
arXiv	24092
Nature Communications	4915
Geoscientific Model Development	3582
Scientific Data	3580
Atmospheric Chemistry & Physics	3494
The Astrophysical Journal	3210
Geophysical Research Letters	2831
Earth System Science Data	2751

Related to Very Early Earth - Recent, Refereed Literature with Associated Data Different Perspectives

Astronomy Centric



Earth Science Centric