System Development: Data Enrichment
Named Entities & Language Model

Felix Grezes, Thomas Allen and the ADS Team

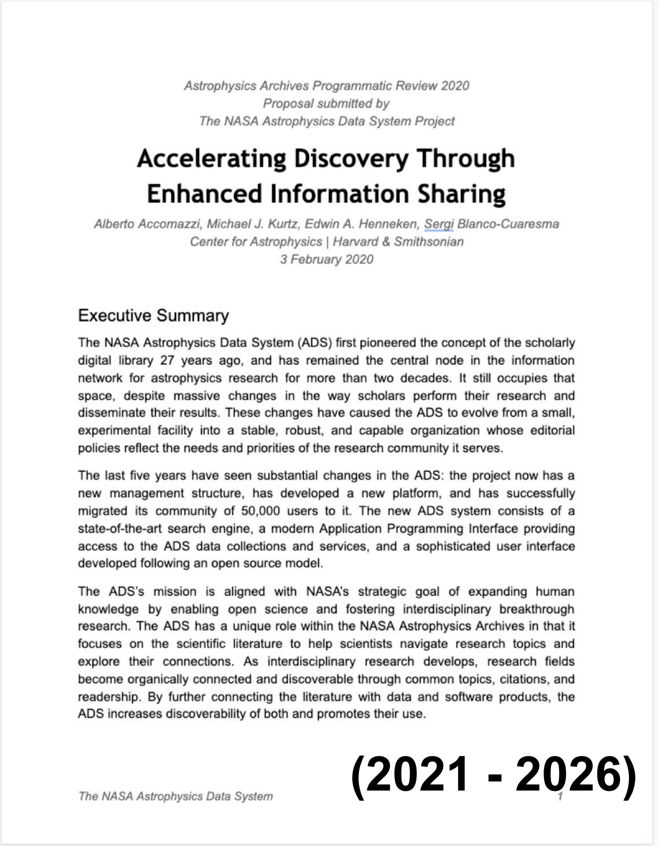
ADSUG Meeting, November 2022

Smithsonian Institution

astrophysics data system

NASA

Why Use Machine Learning/AI at ADS



Astrophysics Archives Programmatic Review 2020
Proposal submitted by
The NASA Astrophysics Data System Project

Accelerating Discovery Through Enhanced Information Sharing

Alberto Accomazzi, Michael J. Kurtz, Edwin A. Henneken, Sergi Blanco-Cuaresma
Center for Astrophysics | Harvard & Smithsonian
3 February 2020

Executive Summary

The NASA Astrophysics Data System (ADS) first pioneered the concept of the scholarly digital library 27 years ago, and has remained the central node in the information network for astrophysics research for more than two decades. It still occupies that space, despite massive changes in the way scholars perform their research and disseminate their results. These changes have caused the ADS to evolve from a small, experimental facility into a stable, robust, and capable organization whose editorial policies reflect the needs and priorities of the research community it serves.

The last five years have seen substantial changes in the ADS: the project now has a new management structure, has developed a new platform, and has successfully migrated its community of 50,000 users to it. The new ADS system consists of a state-of-the-art search engine, a modern Application Programming Interface providing access to the ADS data collections and services, and a sophisticated user interface developed following an open source model.

The ADS's mission is aligned with NASA's strategic goal of expanding human knowledge by enabling open science and fostering interdisciplinary breakthrough research. The ADS has a unique role within the NASA Astrophysics Archives in that it focuses on the scientific literature to help scientists navigate research topics and explore their connections. As interdisciplinary research develops, research fields become organically connected and discoverable through common topics, citations, and readership. By further connecting the literature with data and software products, the ADS increases discoverability of both and promotes their use.

(2021 - 2026)

- What we are accomplishing
  - Text enrichment:
    - Identify Entities in the Literature (e.g., observatories, instruments)
    - Identify Planetary Feature Names
  - Build an Astronomy specific language model to automate the enrichment
  - Provide models and datasets for other researchers in the field

Entities of Interest to the Astronomical Community

Traditional NER

- Person
- Organization
- Location
- EntityOfFutureInterest

Data/Software

- Software
- Model
- Dataset

Facilities

- Observatory
- Telescope
- Instrument
- Wavelength/Filter
- Archive
- Mission
- Collaboration
- Survey
- Database
- ComputingFacility

Awards

- Grant
- Fellowship
- Proposal

Astronomical Objects

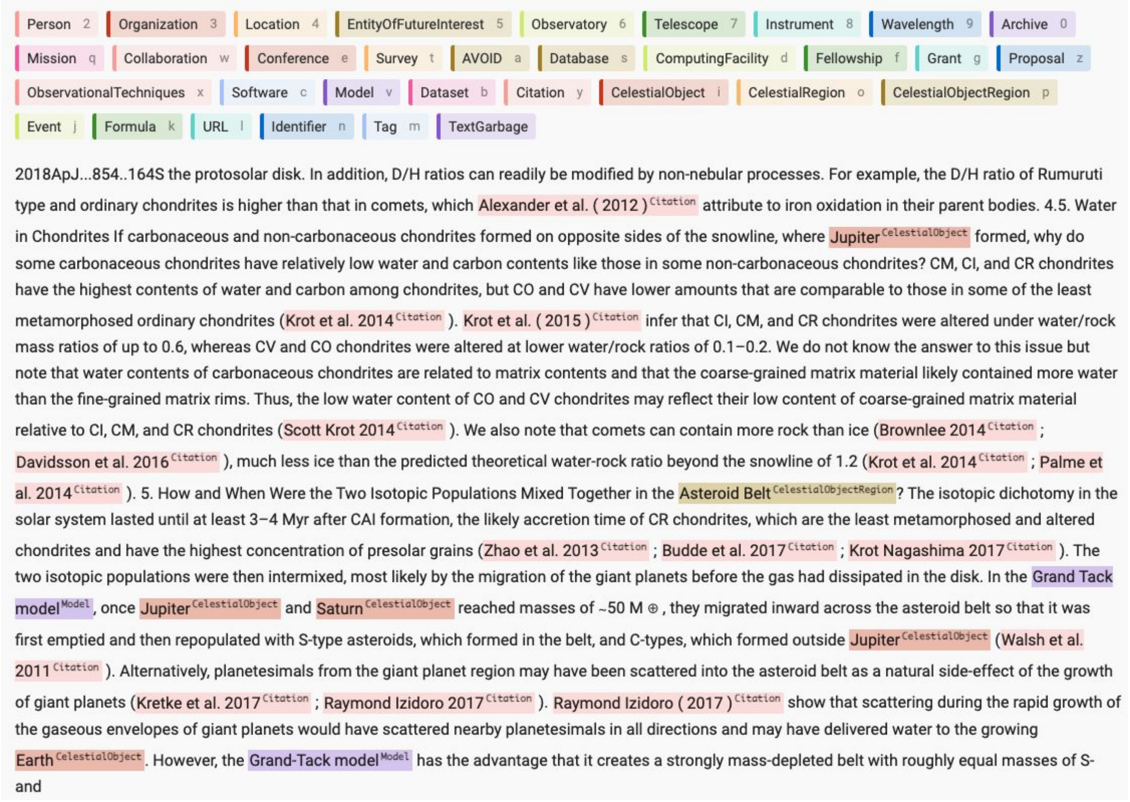
- CelestialObject
- CelestialRegion
- CelestialObjectRegion

Other

- Citation
- Event
- Formula
- URL
- Identifier (e.g., DOIs or other document identifiers)
- Tag (e.g., **<whatever>**some valuable text**</whatever>**)
- TextGarbage

Labeled Entities Astronomy Dataset

- Astrophysical Literature
  - Full-text - 3009 snippets
    71631 labelled entities
  - Acknowledgements - 3004 snippets
    76230 labelled entities
  - ApJ, A&A, MNRAS
  - Years between 2015 and 2020
- Working to automate labeling process

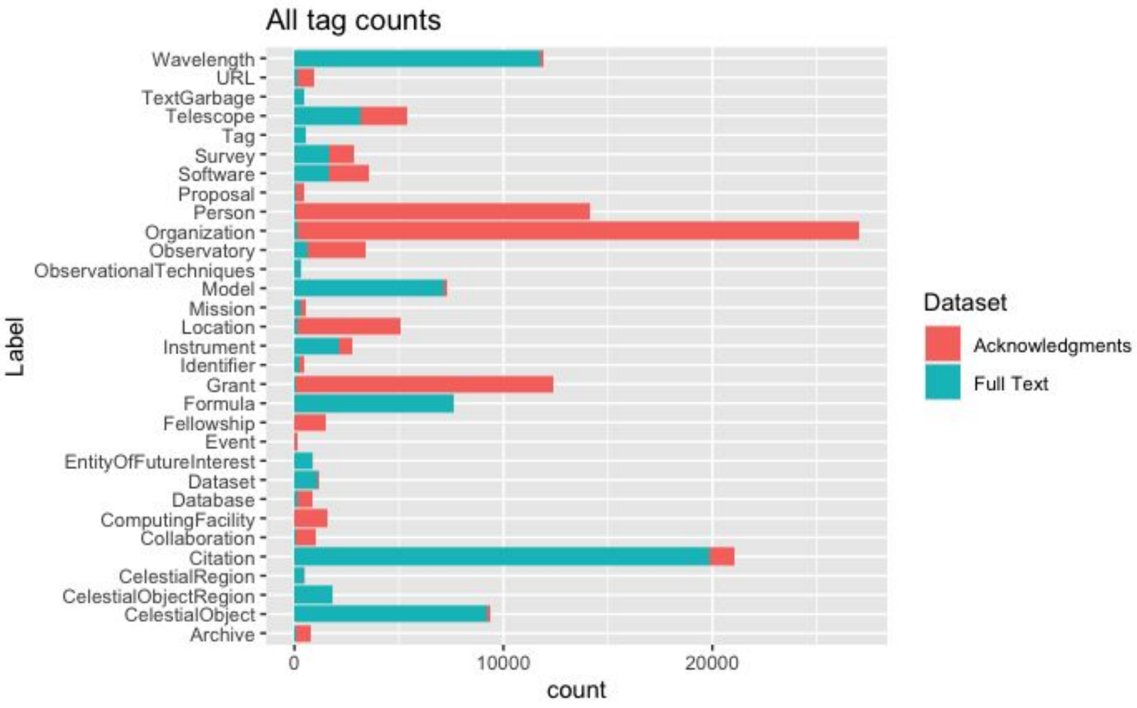


Labeled Entities Astronomy Dataset

- Astrophysical Literature
  - Full-text - 3009 snippets

    71631 labelled entities
  - Acknowledgements - 3004 snippets

    76230 labelled entities
  - ApJ, A&A, MNRAS
  - Years between 2015 and 2020
- Working to automate labeling process

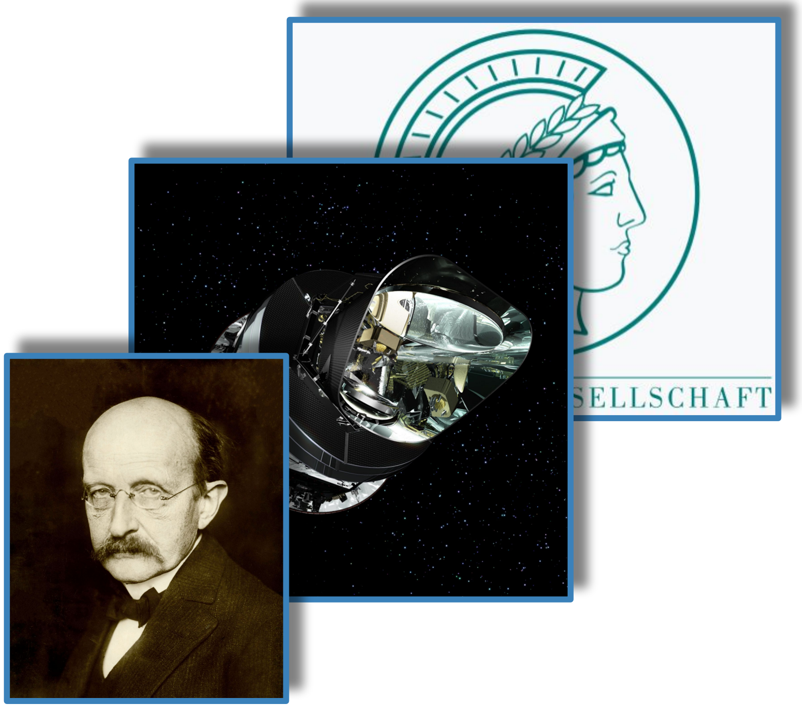


automating the enrichment: astroBERT

astroBERT aims to provide automated enrichment services, similar to SIMBAD and NED but with broader scope.

Under the hood, astroBERT is language model tailored to astrophysics.

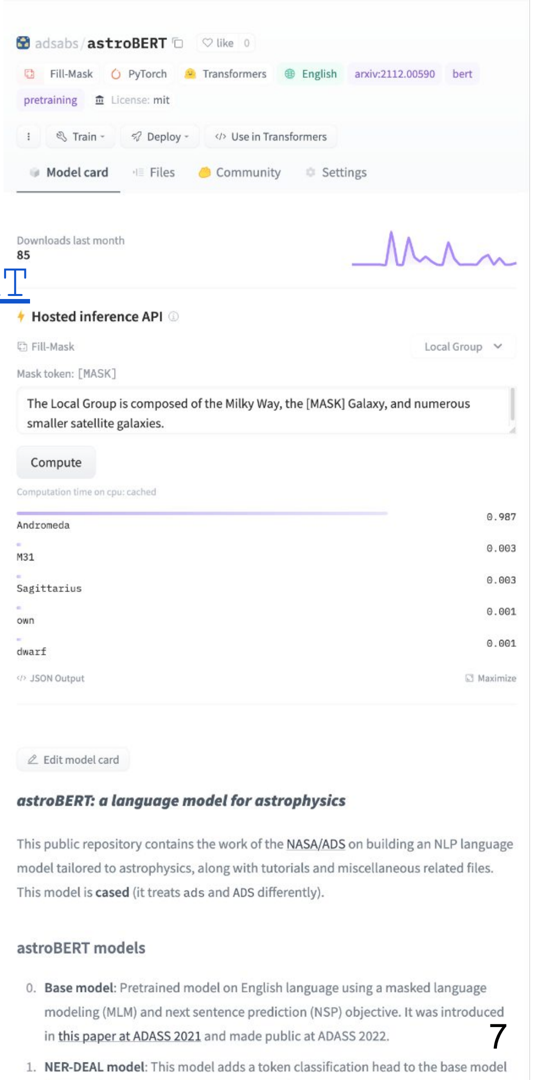
- statistical model that captures the ambiguity of the text used by astrophysicists
- based on the proven technology of BERT and SciBERT
- core component that can be applied to many other tasks, not just labelling entities



astroBERT details

- astroBERT is now publicly available for all
  - 🤗https://huggingface.co/adsabs/astroBERT
  - Includes multiple versions, tutorials, and a demo

- creating astroBERT
  - ~400K astronomy documents
  - ~50 days of computation on dual Nvidia V100 GPUs
  - all open source technologies (Tensorflow, Numpy…)



Evaluating astroBERT

- To evaluate astroBERT, we created the DEAL challenge (Detecting Entities in the Astrophysics Literature)
    - part of a workshop[1] at the AACL-IJCNLP[2] 2022 conference
    - challenged participants to build labeling systems
    - using our dataset of labeled entities
    - measured against BERT, SciBERT, and astroBERT baselines

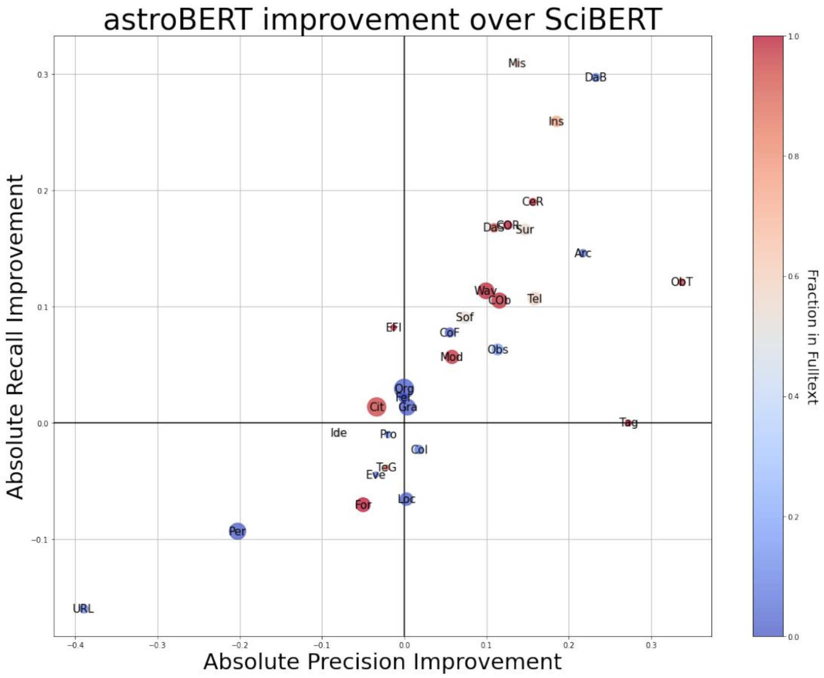
Each baseline model was finetuned for DEAL i.e. the core language models were adapted and retrained for ~12 hours each.

[1] WIESP: The first Workshop on Information Extraction from Scientific Publications
[2] The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing

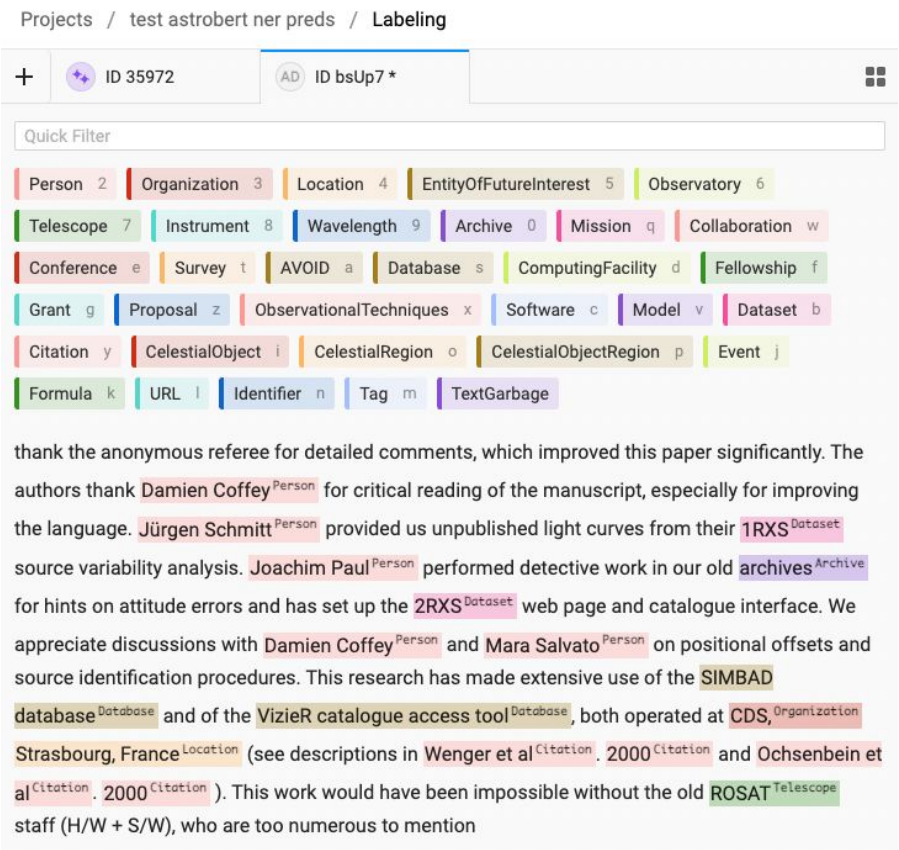
astroBERT on DEAL

astroBERT outperforms BERT and SciBERT on DEAL

Model	Metric / Split	MCC	overall F1 score	overall precision	overall recall	overall accuracy
Random	train	0.1037	0.0170	0.0122	0.0278	0.7146
	val	0.1083	0.0166	0.0119	0.0273	0.7059
	test	0.1057	0.0162	0.0116	0.0269	0.6876
BERT	train	0.7542	0.4920	0.4995	0.4848	0.9256
	val	0.7405	0.4739	0.4780	0.4698	0.9188
	test	0.7229	0.4513	0.4622	0.4409	0.9094
SciBERT	train	0.8159	0.5867	0.5753	0.5986	0.9430
	val	0.8019	0.5601	0.5463	0.5745	0.9366
	test	0.7844	0.5355	0.5313	0.5398	0.9280
astroBERT (WIESP)	train	0.8296	0.6138	0.5889	0.6409	0.9468
	val	0.8104	0.5779	0.5508	0.6077	0.9389
	test	0.7939	0.5561	0.5387	0.5746	0.9308
astroBERT (public release)	train	0.8250	0.5995	0.5701	0.6319	0.9442
	val	0.8194	0.5907	0.5575	0.6282	0.9405
	test	**0.8302**	**0.6093**	**0.5846**	**0.6362**	**0.9418**



Labeling Results



Plans to improve astroBERT

Current astroBERT is good but not production ready

- Improve DEAL performance
  - by using ideas from WIESP: Conditional Random Fields (CRF)
  - with iterative training and labeling
- Improve core language model
  - by training with a new semi-supervised task: Semantic Textual Similarity (STS)
- Evaluate astroBERT on new downstream tasks: UAT Concept Extraction

astroBERT Recap

- We are building astroBERT, a core language model for astrophysics.
- We plan on using it internally for text enrichment, a task in which it already outperforms other language models.
- We publicly released both astroBERT and a dataset of enriched text to the astrophysics research community.