

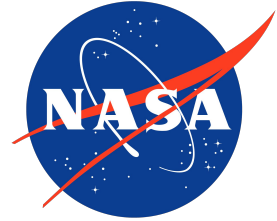
# System Development: Data Enrichment Named Entities & Language Model

*Felix Grezes, Thomas Allen and the ADS Team*

ADSUG Meeting, November 2022



# 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)**

*The NASA Astrophysics Data System*

- **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

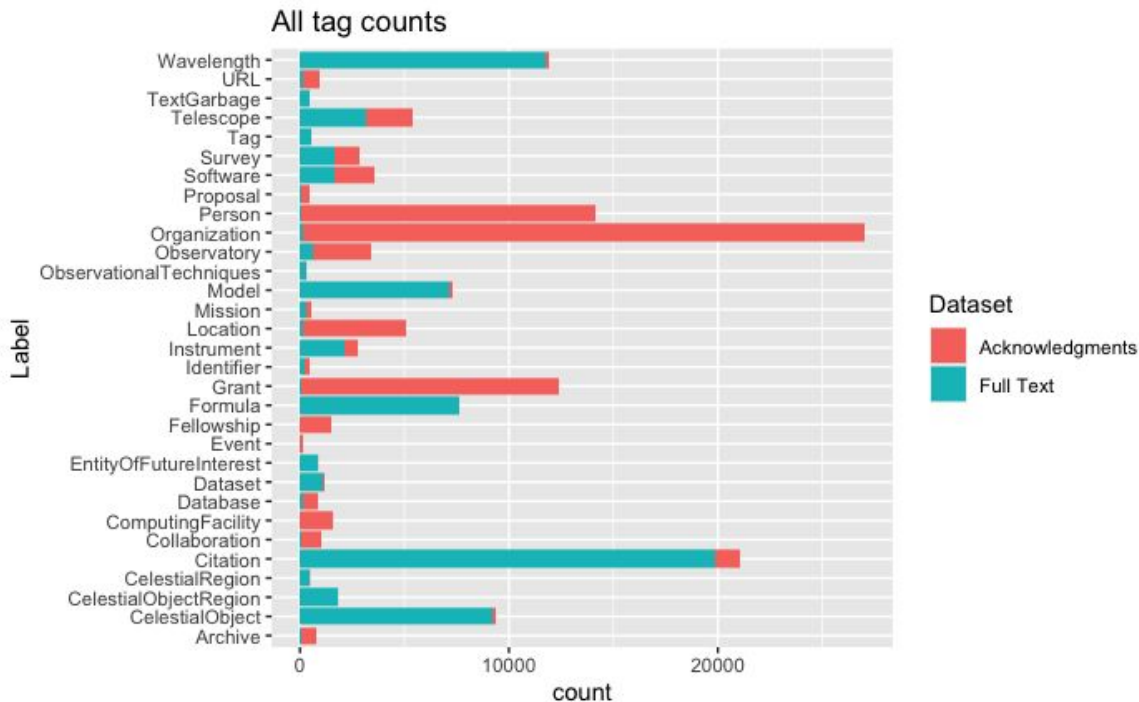
- Working to automate labeling  
process

Person 2 | Organization 3 | Location 4 | EntityOfFutureInterest 5 | Observatory 6 | Telescope 7 | Instrument 8 | Wavelength 9 | Archive 0  
Mission q | Collaboration w | Conference e | Survey t | AVOID a | Database s | ComputingFacility d | Fellowship f | Grant g | Proposal z  
ObservationalTechniques x | Software c | Model v | Dataset b | Citation y | CelestialObject i | CelestialRegion o | CelestialObjectRegion p  
Event j | Formula k | URL l | Identifier n | Tag m | TextGarbage

2018ApJ...854..164S the protosolar disk. In addition, D/H ratios can readily be modified by non-nebular processes. For example, the D/H ratio of Rumuruti type and ordinary chondrites is higher than that in comets, which Alexander et al. (2012)<sup>Citation</sup> attribute to iron oxidation in their parent bodies. 4.5. Water in Chondrites If carbonaceous and non-carbonaceous chondrites formed on opposite sides of the snowline, where Jupiter<sup>CelestialObject</sup> formed, why do some carbonaceous chondrites have relatively low water and carbon contents like those in some non-carbonaceous chondrites? CM, CI, and CR chondrites have the highest contents of water and carbon among chondrites, but CO and CV have lower amounts that are comparable to those in some of the least metamorphosed ordinary chondrites (Krot et al. 2014<sup>Citation</sup>). Krot et al. (2015)<sup>Citation</sup> infer that CI, CM, and CR chondrites were altered under water/rock mass ratios of up to 0.6, whereas CV and CO chondrites were altered at lower water/rock ratios of 0.1–0.2. We do not know the answer to this issue but note that water contents of carbonaceous chondrites are related to matrix contents and that the coarse-grained matrix material likely contained more water than the fine-grained matrix rims. Thus, the low water content of CO and CV chondrites may reflect their low content of coarse-grained matrix material relative to CI, CM, and CR chondrites (Scott Krot 2014<sup>Citation</sup>). We also note that comets can contain more rock than ice (Brownlee 2014<sup>Citation</sup>; Davidsson et al. 2016<sup>Citation</sup>), much less ice than the predicted theoretical water-rock ratio beyond the snowline of 1.2 (Krot et al. 2014<sup>Citation</sup>; Palme et al. 2014<sup>Citation</sup>). 5. How and When Were the Two Isotopic Populations Mixed Together in the Asteroid Belt<sup>CelestialObjectRegion</sup>? The isotopic dichotomy in the solar system lasted until at least 3–4 Myr after CAI formation, the likely accretion time of CR chondrites, which are the least metamorphosed and altered chondrites and have the highest concentration of presolar grains (Zhao et al. 2013<sup>Citation</sup>; Budde et al. 2017<sup>Citation</sup>; Krot Nagashima 2017<sup>Citation</sup>). The two isotopic populations were then intermixed, most likely by the migration of the giant planets before the gas had dissipated in the disk. In the Grand Tack<sup>Model</sup> model, once Jupiter<sup>CelestialObject</sup> and Saturn<sup>CelestialObject</sup> reached masses of ~50 M<sub>⊕</sub>, they migrated inward across the asteroid belt so that it was first emptied and then repopulated with S-type asteroids, which formed in the belt, and C-types, which formed outside Jupiter<sup>CelestialObject</sup> (Walsh et al. 2011<sup>Citation</sup>). Alternatively, planetesimals from the giant planet region may have been scattered into the asteroid belt as a natural side-effect of the growth of giant planets (Kretke et al. 2017<sup>Citation</sup>; Raymond Izidoro 2017<sup>Citation</sup>). Raymond Izidoro (2017)<sup>Citation</sup> show that scattering during the rapid growth of the gaseous envelopes of giant planets would have scattered nearby planetesimals in all directions and may have delivered water to the growing Earth<sup>CelestialObject</sup>. However, the Grand-Tack model<sup>Model</sup> has the advantage that it creates a strongly mass-depleted belt with roughly equal masses of S- and

# 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

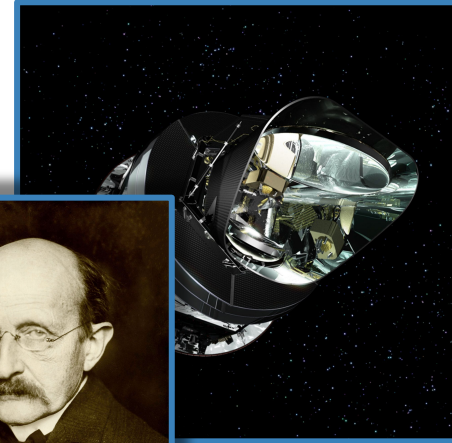
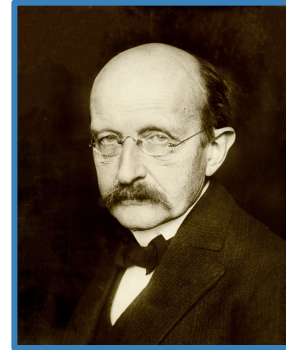


# automating the enrichment: astroBERT

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

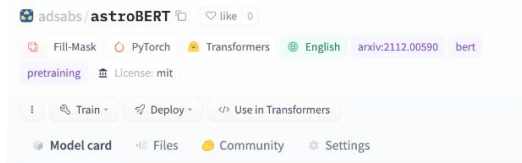
Under the hood, astroBERT is language model tailored to astrophysics.

- 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...)



Downloads last month  
85 

## Hosted inference API

Fill-Mask

Local Group

Mask token: [MASK]

The Local Group is composed of the Milky Way, the [MASK] Galaxy, and numerous smaller satellite galaxies.

Compute

Computation time on cpu: cached

Andromeda	0.987
M31	0.003
Sagittarius	0.003
own	0.001
dwarf	0.001
JSON Output	Maximize

Edit model card

## astroBERT: a language model for astrophysics

This public repository contains the work of the NASA/ADS on building an NLP language model tailored to astrophysics, along with tutorials and miscellaneous related files. This model is **cased** (it treats ads and ADS differently).

## astroBERT models

0. **Base model:** Pretrained model on English language using a masked language modeling (MLM) and next sentence prediction (NSP) objective. It was introduced in [this paper at ADASS 2021](#) and made public at ADASS 2022.

1. **NER-DEAL model:** This model adds a token classification head to the base model

# Evaluating astroBERT

- To evaluate astroBERT, we created the DEAL challenge (Detecting Entities in the Astrophysics Literature)
  - part of a workshop<sup>[1]</sup> at the ACL-IJCNLP<sup>[2]</sup> 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

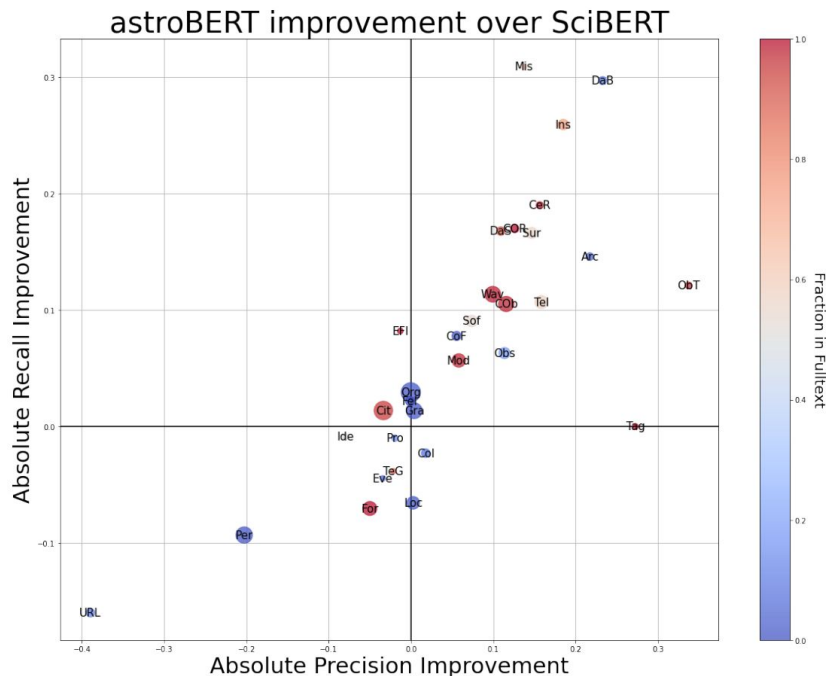
[2] The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing



# astroBERT on DEAL

*astroBERT outperforms BERT and SciBERT on DEAL*

Model	Split	Metric	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		<b>0.8302</b>	<b>0.6093</b>	<b>0.5846</b>	<b>0.6362</b>	<b>0.9418</b>



# Labeling Results

Projects / test astrobert ner preds / Labeling

+ ID 35972 AD ID bsUp7\*

Quick Filter

Person 2 | Organization 3 | Location 4 | EntityOfFutureInterest 5 | Observatory 6  
Telescope 7 | Instrument 8 | Wavelength 9 | Archive 0 | Mission q | Collaboration w  
Conference e | Survey t | AVOID a | Database s | ComputingFacility d | Fellowship f  
Grant g | Proposal z | ObservationalTechniques x | Software c | Model v | Dataset b  
Citation y | CelestialObject i | CelestialRegion o | CelestialObjectRegion p | Event j  
Formula k | URL l | Identifier n | Tag m | TextGarbage

thank the anonymous referee for detailed comments, which improved this paper significantly. The authors thank **Damien Coffey** Person for critical reading of the manuscript, especially for improving the language. **Jürgen Schmitt** Person provided us unpublished light curves from their **1RXS** Dataset source variability analysis. **Joachim Paul** Person performed detective work in our old **archives** Archive for hints on attitude errors and has set up the **2RXS** Dataset web page and catalogue interface. We appreciate discussions with **Damien Coffey** Person and **Mara Salvato** Person on positional offsets and source identification procedures. This research has made extensive use of the **SIMBAD** Database database and of the **VizieR** catalogue access tool Database, both operated at **CDS**, Organization Strasbourg, France Location (see descriptions in **Wenger et al** Citation. 2000 Citation and **Ochsenbein et al** Citation. 2000 Citation). This work would have been impossible without the old **ROSAT** Telescope staff (H/W + S/W), who are too numerous to mention

Projects / test astrobert ner preds / Labeling

+ ID 35967 AD ID DwBz1\*

Quick Filter

Person 2 | Organization 3 | Location 4 | EntityOfFutureInterest 5 | Observatory 6  
Telescope 7 | Instrument 8 | Wavelength 9 | Archive 0 | Mission q | Collaboration w  
Conference e | Survey t | AVOID a | Database s | ComputingFacility d | Fellowship f  
Grant g | Proposal z | ObservationalTechniques x | Software c | Model v | Dataset b  
Citation y | CelestialObject i | CelestialRegion o | CelestialObjectRegion p | Event j  
Formula k | URL l | Identifier n | Tag m | TextGarbage

; **White van Paradijs** 1996 Citation; **Romani** 1998 Citation) and implies that only of **Galactic** CelestialObject BHTs have been discovered. Our empirical estimate is an order of magnitude lower than the BHTs predicted by **Kiel Hurley** (2006) Citation or **Yungelson et al** Citation. (2006) Citation using population synthesis models. However, we note that our analysis is based on the study of observed systems but limited only to the nine BHTs with reliable distance estimates, which are located in a cylinder of 4 kpc radius centred on the **Sun** CelestialObject. In addition, we have assumed that the solar vertical distribution () can be extrapolated to other regions of the **Galaxy** CelestialObject. However, the bulge contains of the stellar mass of the **Galaxy** CelestialObject, which is confined in a reduced spheroid and it is expected to host a higher concentration of BHTs (**Muno et al** Citation. 2005 Citation). Furthermore, we considered a cylinder with a height defined by **MAXI J1659152** CelestialObject (the object with the highest ) but there could be objects located at higher distances over the plane. Finally, we normalized our estimated value to an average recurrence period of 100 yr, which explicitly does not take into consideration any systems with lower accretion rates or longer recurrence periods, nor does it account for a likely population of intrinsically faint **X** Wavelength-**ray** Wavelength BHTs. Taking in all of the above, we conclude that our crude calculation of the number of BHTs expected in the **Galaxy** CelestialObject is very conservative and sets a lower limit to the hidden population. 4. Physical properties of dynamical BHTs Fig. 7 Histograms of the 17 dynamically confirmed BHs. Left : extinction-corrected absolute -band magnitudes in bins of 2 mag. The black histoaram denotes confirmed IMXBs. Right : orbital periods in a logarithmic scale usina

# 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.