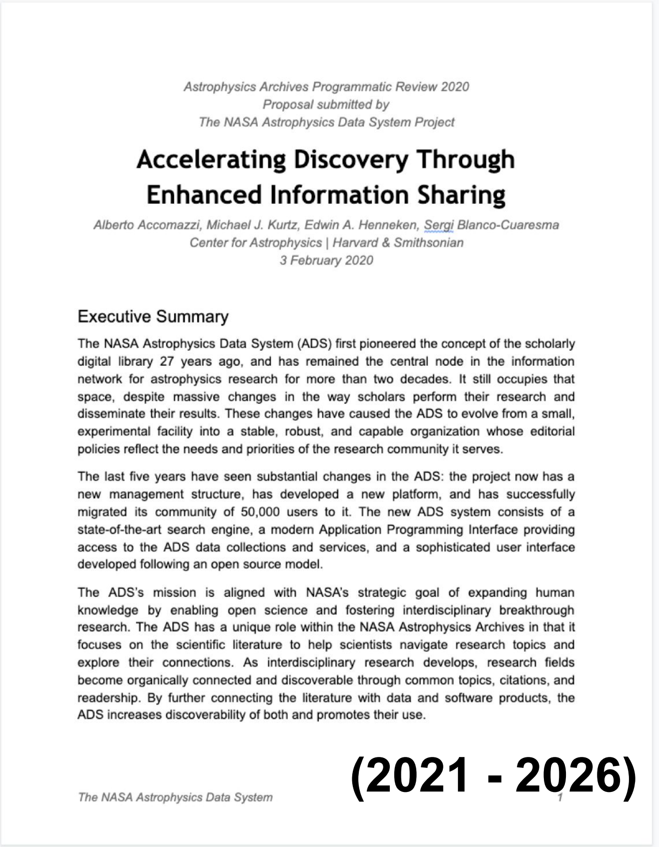
Natural Language Processing (NLP) Overview

Sergi Blanco-Cuaresma, Felix Grezes, Golnaz Shapurian, Thomas Allen and the ADS Team

ADS Users Group Meeting, 15-16 Nov. 2021

Senior review 2020



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

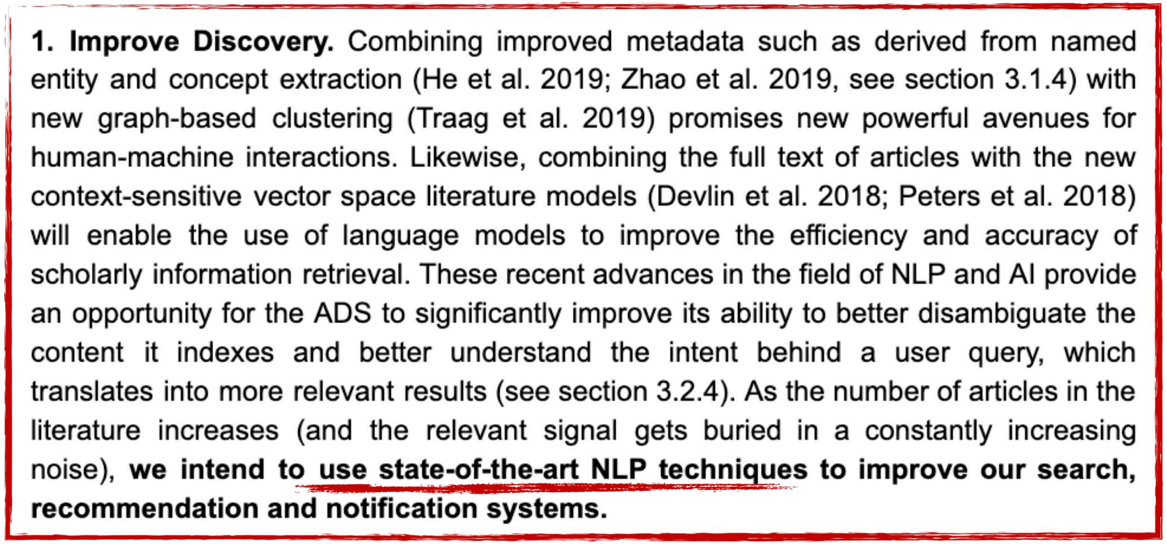
Executive Summary

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.

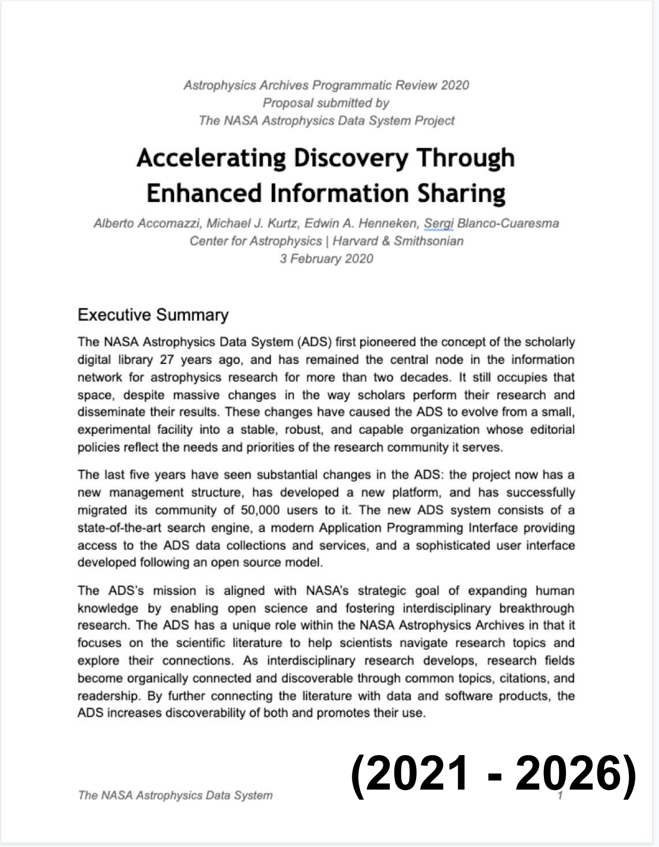
The ADS's mission is aligned with NASA's strategic goal of expanding human knowledge by enabling open science and fostering interdisciplinary breakthrough research. The ADS has a unique role within the NASA Astrophysics Archives in that it focuses on the scientific literature to help scientists navigate 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)



1. Improve Discovery. Combining improved metadata such as derived from named entity and concept extraction (He et al. 2019; Zhao et al. 2019, see section 3.1.4) with new graph-based clustering (Traag et al. 2019) promises new powerful avenues for human-machine interactions. Likewise, combining the full text of articles with the new context-sensitive vector space literature models (Devlin et al. 2018; Peters et al. 2018) will enable the use of language models to improve the efficiency and accuracy of scholarly information retrieval. These recent advances in the field of NLP and AI provide an opportunity for the ADS to significantly improve its ability to better disambiguate the content it indexes and better understand the intent behind a user query, which translates into more relevant results (see section 3.2.4). As the number of articles in the literature increases (and the relevant signal gets buried in a constantly increasing noise), **we intend to use state-of-the-art NLP techniques to improve our search, recommendation and notification systems.**

Senior review 2020



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.

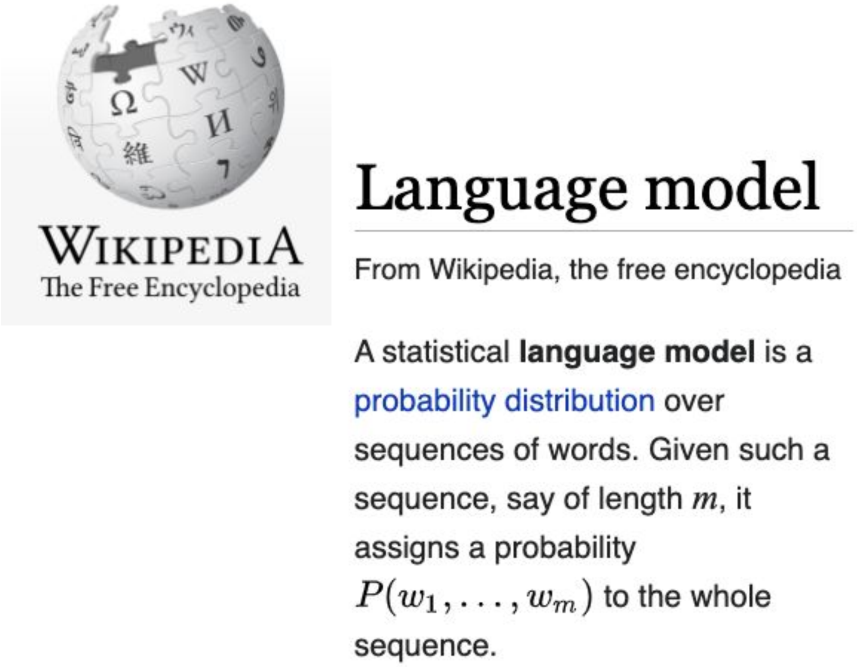
The NASA Astrophysics Data System 1

(2021 - 2026)

- What we want to accomplish
  - ○ Text enrichment:
    - ■ Identify facilities (e.g., observatories, instruments)
    - ■ Connect to Unified Astronomy Thesaurus (UAT) terms
  - ○ Personalized recommendations
  - ○ Disambiguate authors (e.g., Is this your paper?)
  - ○ UI suggestions (e.g., Did you mean this?)
  - ○ Understand Citation Context
  - ○ ...

Language Model

BERT



WIKIPEDIA
The Free Encyclopedia

Language model

From Wikipedia, the free encyclopedia

A statistical **language model** is a probability distribution over sequences of words. Given such a sequence, say of length m, it assigns a probability $P(w_1, \ldots, w_m)$ to the whole sequence.

Language Model

BERT

- Guess the masked word
  - The **[MASK]** medium is the gas and dust between stars
  - The **primary** medium is the gas and dust between stars

Language Model

BERT

- Guess the masked word
  - A solar twin is a star with **[MASK]** parameters and chemical composition very similar to our Sun.
  - A solar twin is a star with **orbital** parameters and chemical composition very similar to our Sun.

Language Model

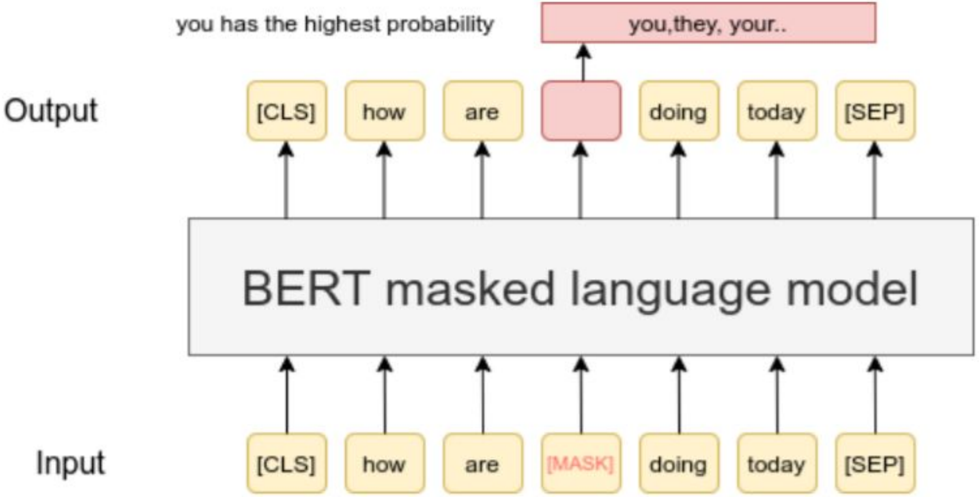
- Guess the masked word
  - ○ Stars are made mostly of **[MASK]**.
  - ○ Stars are made mostly of **gold**.

Language Model

BERT

BERT

Training astroBERT



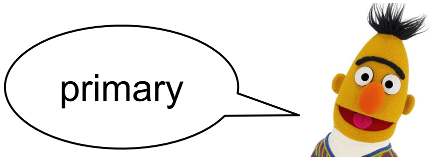
2 x NVIDIA V100 GPUs

Dataset

395,499	papers
121,207,934	sentences
2,977,635,680	words

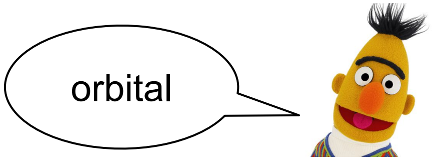
Testing astroBERT

- Guess the masked word
  - ○ The **[MASK]** medium is the gas and dust between stars
  - ○ The **interstellar** medium is the gas and dust between stars



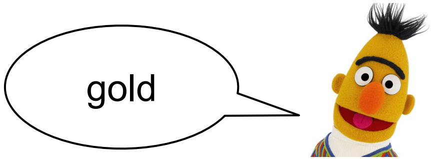
Testing astroBERT

- Guess the masked word
  - A solar twin is a star with **[MASK]** parameters and chemical composition very similar to our Sun.
  - A solar twin is a star with **atmospheric** parameters and chemical composition very similar to our Sun.

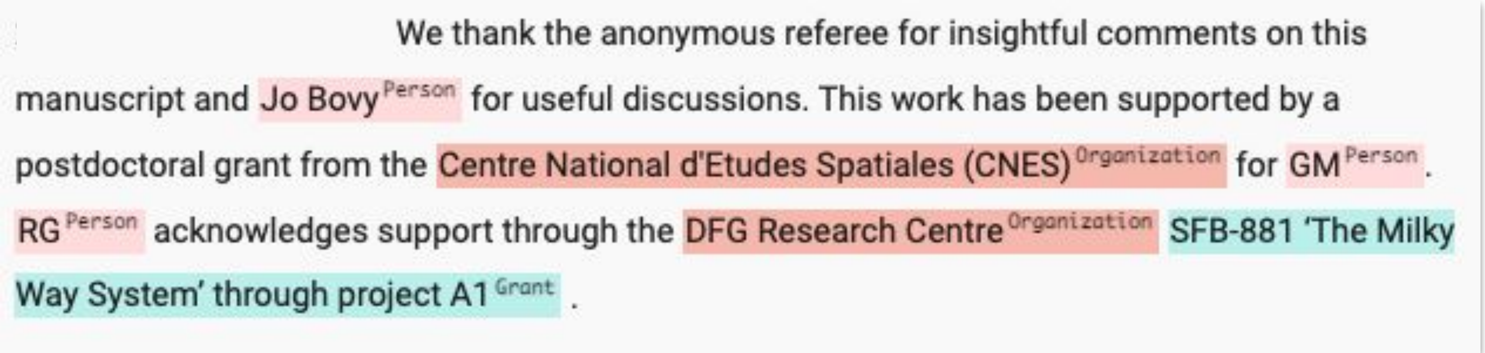


Testing astroBERT

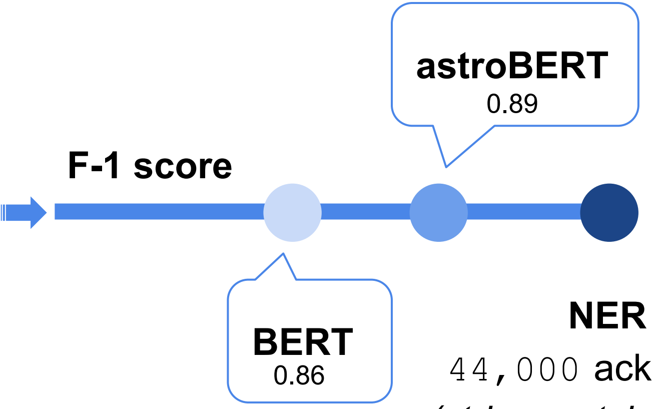
- Guess the masked word
  - Stars are made mostly of **[MASK]**.
  - Stars are made mostly of **stars**.



astroBERT & Named Entity Recognition (NER)



We thank the anonymous referee for insightful comments on this manuscript and Jo Bovy Person for useful discussions. This work has been supported by a postdoctoral grant from the Centre National d'Etudes Spatiales (CNES) Organization for GM Person. RG Person acknowledges support through the DFG Research Centre Organization SFB-881 'The Milky Way System' through project A1 Grant .



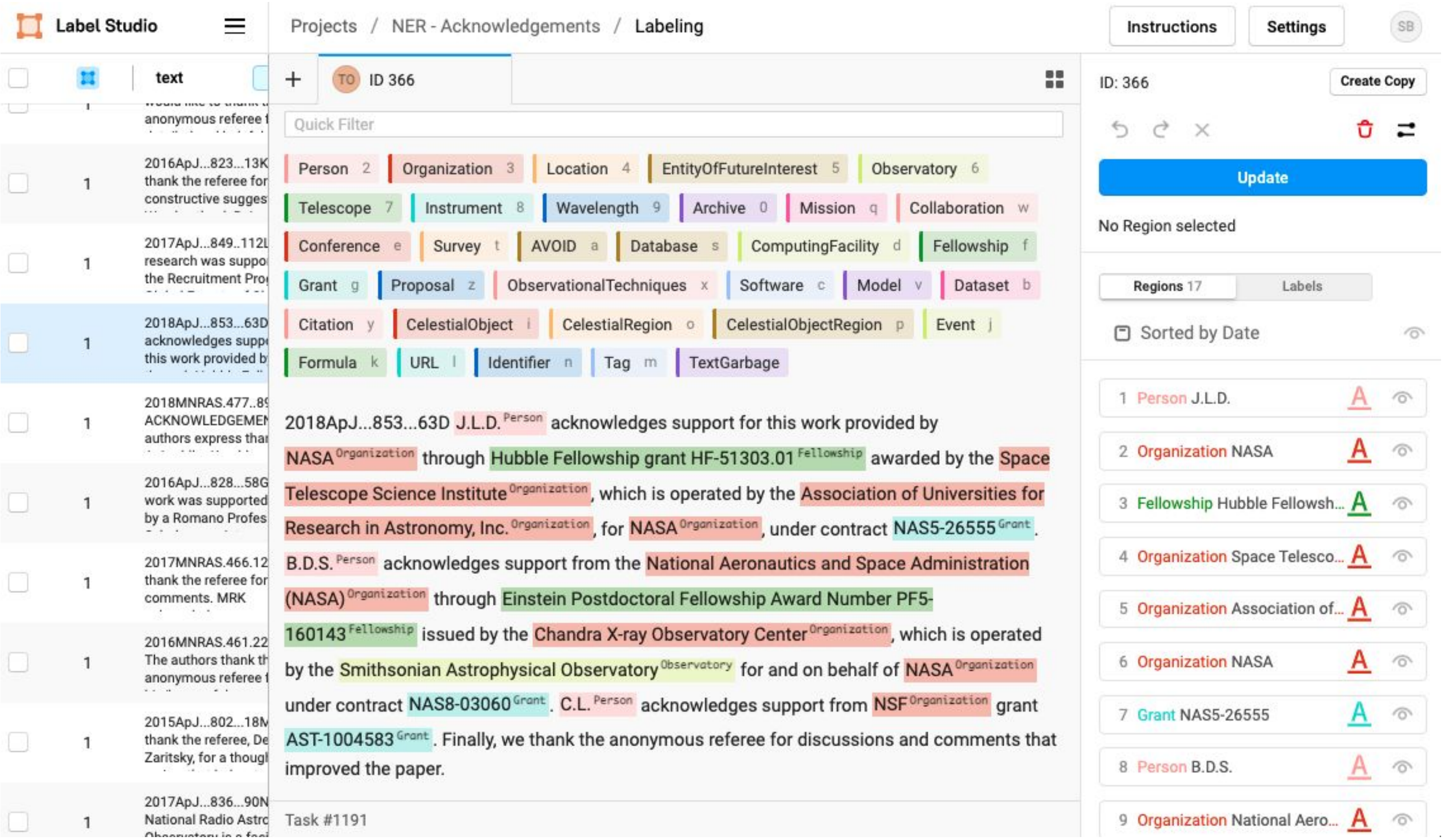
NER Dataset

`44,000` acknowledgements

(string matched <u>organizations</u>)

New NER dataset

Label Studio



2016ApJ...823...13K thank the referee for constructive sugges

2017ApJ...849..112L research was suppo the Recruitment Pro

2018ApJ...853...63D acknowledges supp this work provided b

2018MNRAS.477..8 ACKNOWLEDGEMEN authors express tha

2016ApJ...828...58G work was supported by a Romano Profes

2017MNRAS.466.12 thank the referee for comments. MRK

2016MNRAS.461.22 The authors thank th anonymous referee

2015ApJ...802...18M thank the referee, De Zaritsky, for a thoug

2017ApJ...836...90N National Radio Astro

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...853...63D J.L.D. Person acknowledges support for this work provided by NASA Organization through Hubble Fellowship grant HF-51303.01 Fellowship awarded by the Space Telescope Science Institute Organization, which is operated by the Association of Universities for Research in Astronomy, Inc. Organization, for NASA Organization, under contract NAS5-26555 Grant. B.D.S. Person acknowledges support from the National Aeronautics and Space Administration (NASA) Organization through Einstein Postdoctoral Fellowship Award Number PF5-160143 Fellowship issued by the Chandra X-ray Observatory Center Organization, which is operated by the Smithsonian Astrophysical Observatory Observatory for and on behalf of NASA Organization under contract NAS8-03060 Grant. C.L. Person acknowledges support from NSF Organization grant AST-1004583 Grant. Finally, we thank the anonymous referee for discussions and comments that improved the paper.

Task #1191

No Region selected

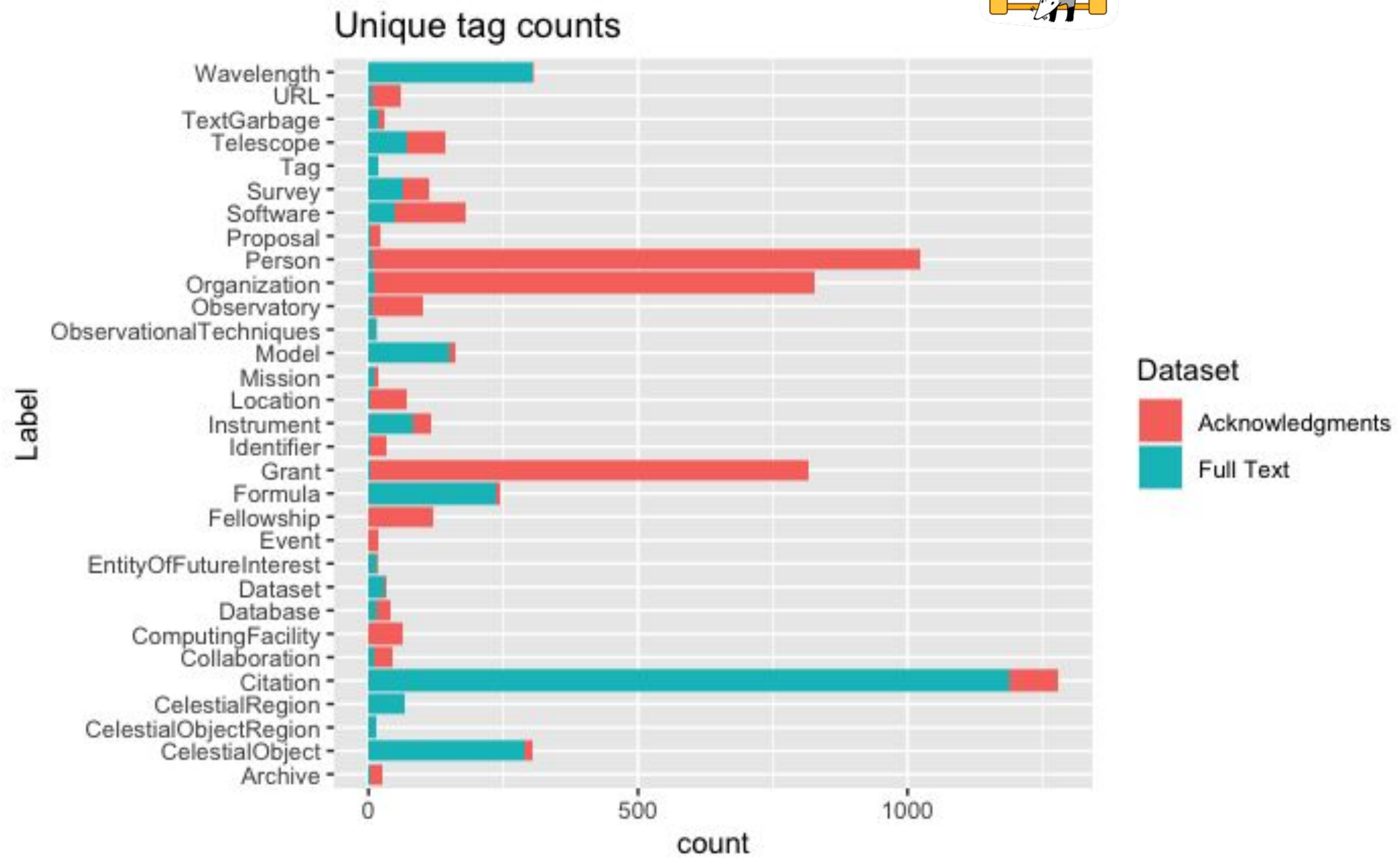
1 Person J.L.D.
2 Organization NASA
3 Fellowship Hubble Fellowsh...
4 Organization Space Telesco...
5 Organization Association of...
6 Organization NASA
7 Grant NAS5-26555
8 Person B.D.S.
9 Organization National Aero...

+280 tagged **acknowledgements** and counting...

+260 tagged **full text fragments** and counting...

New NER dataset

Label Studio



+280 tagged **acknowledgements** and counting...

+260 tagged **full text fragments** and counting...

Summary

- Work completed
  - First prototype of astroBERT
  - First semi-automatic NER dataset
  - First NER test searching for organizations in acknowledgements

- Work ahead
  - Create high-quality astronomical NER dataset (ack+full text fragments)
  - Automatic detection of facilities (observatories, instruments…)
  - Assign Unified Astronomy Thesaurus terms to papers