

Making an Annotated and Curated Corpus of Heliophysics Literature for Named Entity Recognition

Liza Fretel, Sébastien Durna, Baptiste Cecconi

DASH IHDEA Lightning Talk
6th of October, 2026

LIRA  Observatoire
de Paris

PSL 

I. Context (Liza's Ph.D)

1. Reproducibility of heliophysics pipelines

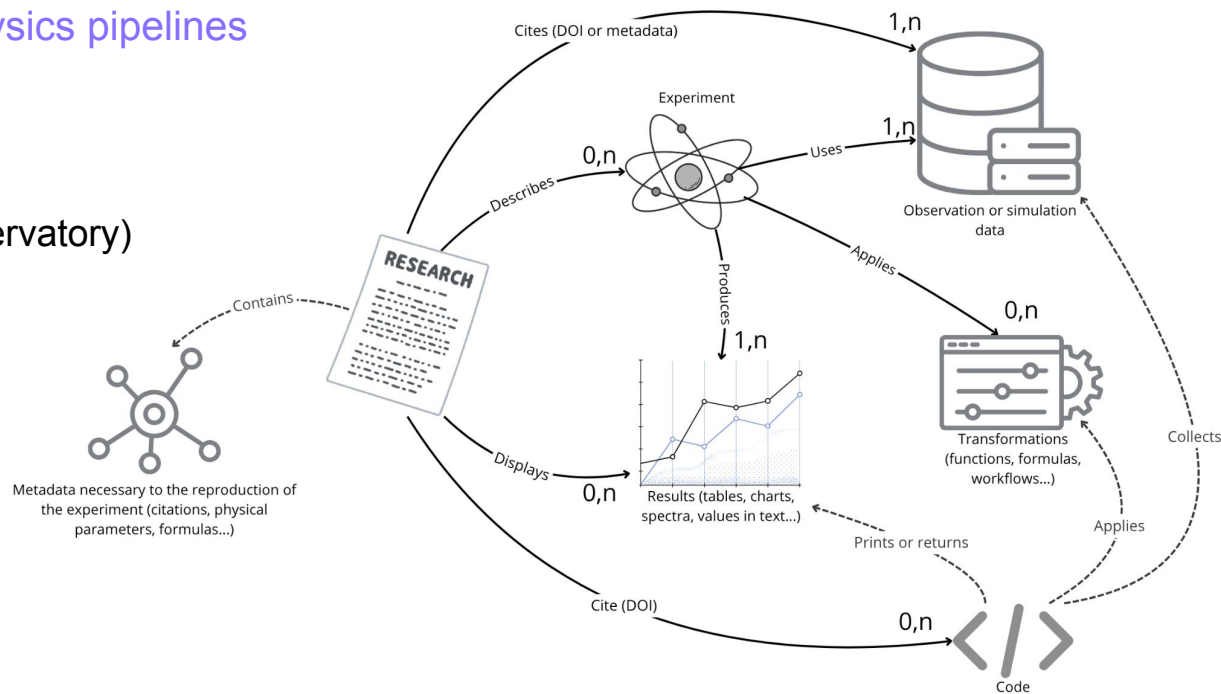
Directors

Baptiste CECCONI (LIRA, Paris Observatory)

Cyril GROUIN (LISN, Paris Saclay)

Research Question

From Heliophysics papers, how to extract information necessary to data access and generate code to reproduce the described experiments?



I. Context (Liza's Ph.D)

1. Reproducibility of heliophysics pipelines

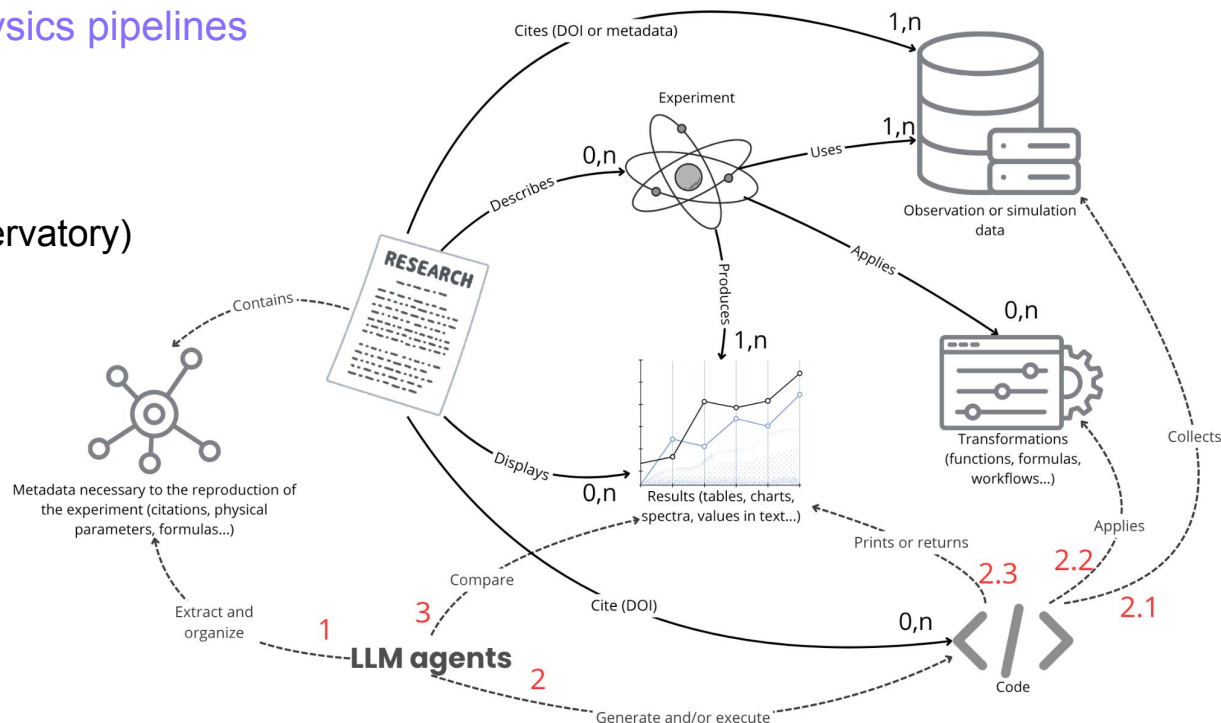
Directors

Baptiste CECCONI (LIRA, Paris Observatory)

Cyril GROUIN (LISN, Paris Saclay)

Research Question

From Heliophysics papers, how to extract information necessary to data access and generate code to reproduce the described experiments?



II. Automatic pre-annotation

1. Tokens pre-annotation

Domain vocabularies (Helio-KNOW)

- Phenomenon => Phenomenon
- Instrument => ObservationalTool
- Model => NumericalTool
- Region => ObjectOfInterest

BERT*-NER* models for astronomy

- adsabs/astroBERT
- adsabs/nasa-smd-ibm-v0.1_NER_DEAL
- atillaalkan/astroNLPy-ner

Three different LLMs*

- openai/gpt-oss-120b
- mistralai/Mistral-Small-3.2-24B-Instruct-2506
- mistralai/Ministral-3-8B-Instruct-2512

String matching (papers' text => terms from the vocabularies)

Run => labels for astronomy => mapping from their schema to our annotation schema => labels for HP

Prompt with the annotation schema and the papers' text

*BERT = encoder architecture

*LLM = decoder architecture

*NER = task of token classification

II. Automatic pre-annotation

1. Tokens pre-annotation

Token classification task

ObservableEntity 1

Phenomenon 2

PhysicalParameter 3

UnitOfMeasure 4

TimeCoverage 5

SpatialCoverage 6

Observatory 7

ObservationalTool 8

ObservationalTechniques 9

NumericalTool 0

Instrument q

InstrumentType w

Database e

Dataset t

Citation a

URL s

Identifier d

SpectralCoverage f

Model g

scintillation timescales via full width at half maximum (FWHM). Applying this chain to multi-channel observations of [Cassiopeia A](#) and [Cygnus A](#), we find systematically higher IPS ratios and narrower FWHM for [Cas A](#), with statistically significant separation confirmed by both two-sample t-tests ($p = 0.0006$) and non-overlapping bootstrap confidence intervals. These results demonstrate that IPS dominates the fast fluctuations in [Cas A](#), while [Cyg A](#) serves as a control for ionospheric scintillation. In agreement with [Fallows et al. \(2020\)](#), we acknowledge possible D-region contributions at low frequencies, but our multi-source comparison rules out ionospheric contamination as the sole explanation. The methodology provides a transparent, transferable standard for IPS detection, applicable to other compact sources and coordinated multi-station campaigns.

[\\(^{1}\\)](#) Institute of Radio Astronomy, National Academy of Sciences of Ukraine, 61002 Kharkiv, Ukraine

[\\(^{2}\\)](#) LESIA, Observatoire de Paris, CNRS, Université PSL, Sorbonne Université, Université Paris Cité, CNRS, 92190 Meudon, France

[\\(^{3}\\)](#) Observatoire Radioastronomique de Nançay (ORN), Obs. Paris, CNRS, PSL, UO, Nançay, France

[\\(^{4}\\)](#) Poltava Gravimetric Observatory, Subbotin Institute of Geophysics, National Academy of Sciences of Ukraine, 36029 Poltava, Ukraine

A. A. Stanislavsky et al.

1 Introduction

Regular observations of ionospheric and interplanetary scintillations in the radiation from discrete cosmic sources have a long history (Rickett, 2001). Often in this case, the four most powerful radio sources in the Northern Hemisphere, namely, [Cassiopeia A](#) ([Cas A](#) in short), [Cygnus A](#) ([Cyg A](#)), [Taurus A](#), and [Virgo A](#), are used. In particular, the [Poker Flat Observatory](#), Alaska (65.1°N, 147.5°W), investigated ionospheric scintillations from these sources with a high-latitude imaging riometer (Beazrodny et al., 2008). The system receives signals through 16×16 pairs of crossed horizontal symmetric dipoles. The phasing antenna array forms 256 narrow beams covering the angular sector about ±75° from the zenith. The instrument is capable of mapping the celestial hemisphere as a multi-beam radio telescope, but only at a single frequency of 38.2 MHz. The measurements indicated the possibility of reliably and simultaneously recording scintillations of several discrete cosmic sources. The aim of our paper is to present wide-band observations of radio source scintillation and their analysis using a

#825

Submit

II. Automatic pre-annotation

2. Annotation schema

References in the PDF

FormulaMention
FigureMention
TableMention

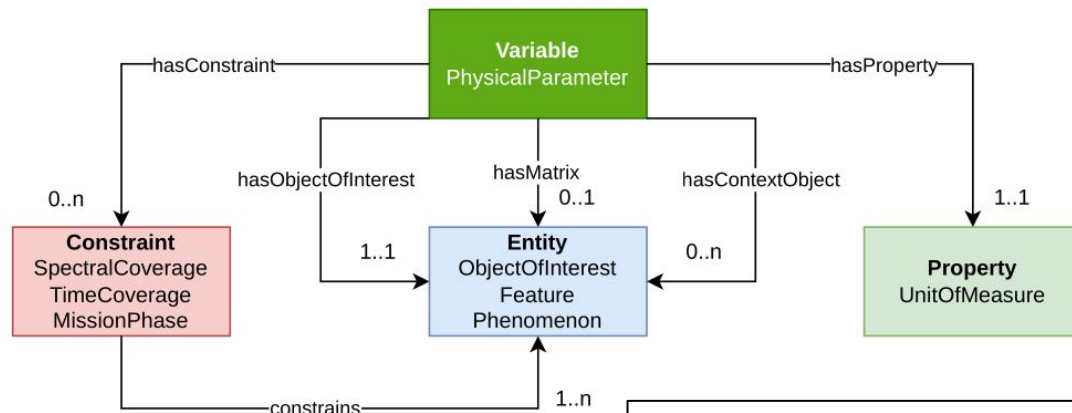
Cross-resource References

URL
Identifier
Citation

Scientific tools

ObservationalTool
NumericalTool
Database
Dataset

Scientific variable description



Thanks for your attention

Bibliography & Sitography

Alkan et al., “HelioBERT-XL: a pre-trained language model for large-scale heliophysical literature mining” (2026)

Grezes et al., “Building astroBERT, a language model for Astronomy & Astrophysics” (2021)

Livathinos et al., “Docling: An Efficient Open-Source Toolkit for AI-driven Document Conversion” (2025)

Yin et al., “Unlimited OCR Works” (2026)

Ye et al., “ReplicationBench: Can AI Agents Replicate Astrophysics Research Papers?” (2025)

<https://hal.science/hal-04285006v1> Génot et. al., “BibHelioTech” (2023)

<https://paper-data.helioanalytics.io/> Paper-Data, Buonomo et. al., paper in writing

<https://github.com/rmcgranaghan/Helio-KNOW> (Commons & vocabularies for HP)

<https://github.com/sapols/PyHC-Chat>

<https://scixplorer.org/>

<https://graph.openaire.eu/>

<https://www.acm.org/publications/policies/artifact-review-and-badging-current> ACM, “Artifact Review and Badging - Current” (2020)