Home

COMET Community Enrichment Project Hub:

Improve Affiliations Parsing of Preprints

Welcome! 👋

This is the Project Hub for the COMET initiative enrichment project 'Improve affiliations parsing of preprints'. Use the tabs on the left to navigate through the sections of this document. This open document aims to keep the community up to date and invites participation and feedback – please comment on anything to ask questions and give your input. You can also reach out to the COMET Lead, Dione Mentis, at dione.mentis@datacite.org.

Announcements 🔊

Updates

Project Updates

Current Phase

Status	Detail
In Progress	A text-based affiliation parsing model has been developed and trained. We are currently assessing its performance on the benchmark dataset, as well as in comparison to alternative methods including GROBID and visual language models.

Next Phase

Status	Detail
Reporting	Publishing a report of the project results

Overview

Project Overview

Project Objective

We're developing improved parsing methods for preprints, using works from arXiv in the pilot, whilst simultaneously evaluating how GROBID, a widely used parsing tool, performs on preprints.

Why This Matters

Preprints represent the cutting edge of research sharing, where scientists first communicate breakthrough findings and emerging theories. Yet the predominant sharing format – PDF – presents challenges to metadata extraction, requiring intelligent PDF parsing tools that can handle significant layout variability. When these tools fail, this creates gaps in author and affiliation information that leave crucial institutional connections incorrectly captured or completely missing.

The Approach

We're conducting a dual-purpose investigation that both evaluates existing systems and develops improved solutions. By systematically sampling different types of OpenAlex records, we can understand both the current state of metadata extraction and the specific challenges that different submission patterns present.

We’re using three strategically sampled OpenAlex datasets of ArXiv preprint submissions as inputs

- works with affiliations
- works without affiliations
- a random general sample to capture the full spectrum of parsing and extraction patterns.

We're measuring success through several key outcomes:

- **Comprehensive Baseline Assessment**: Clear understanding of how well GROBID performs when OpenAlex processes arXiv content, establishing benchmarks for improvement across different types of preprint submissions
- **Improved Parsing Performance**: Development of preprint parsing methods that significantly outperform current approaches in extracting author and affiliation metadata
- **Cost-Effective Processing**: Creation of solutions that provide better results whilst remaining economically viable for processing the continuous flow of new preprints

Expected Results

This project aims to produce:

- Benchmark datasets and strategy performance metrics
- A trained parsing model optimized for arXiv preprints
- Quantitative assessment on the performance of OpenAlex's current GROBID pipeline for arXiv metadata extraction.

Who Benefits

Improved parsing of preprints can have many benefits, including:

- **Preprint Authors** receive better attribution and discoverability for their early research contributions
- **Research Institutions** gain proper recognition for their researchers' preprint activities, enabling better tracking of institutional research output and collaboration patterns
- **Research Libraries and Discovery Systems** can provide more reliable search and filtering capabilities
- **OpenAlex and Similar Aggregation Services** receive detailed feedback on parsing performance that can inform systematic improvements to their metadata extraction processes
- **The Broader Research Community** gains access to more reliable early research discovery and can better understand emerging research trends through comprehensive preprint metadata.

Team & Tasks

Project Team & Tasks

Team

Name (Affiliation)	Role	Location
Dione Mentis (DataCite)	Project management	South Africa
Adam Buttrick (California Digital Library, CDL)	Organizer	United States
Parth Sarin (Stanford University)	Development	United States
Eric Jeangirard (French Ministry of Higher Education and Research)	Advisory	France

Task tracker

Assignee	Title	Date	Status
Adam Buttrick	Evaluate the accuracy of OpenAlex's current GROBID implementation on arXiv papers	Sep 19, 2025	Completed
Adam Buttrick	Compare performance of	Sep 26, 2025	Completed

Task tracker

Assignee	Title	Date	Status
	the developed text-based model against a visual language model.		
Person		Date	Not started

Resources

Project Resources

Resource Name	Type	Description	Link or File
Code repository	Source Code	The Github repository. See READMEs for usage instructions	https://github.com/cometadata/arxiv-preprint-parsing
Trained affiliation parsing model	Model	This model is a fine-tuned version of Qwen/Qwen3-4B trained using Group Relative Policy Optimization (GRPO) for parsing and extracting author affiliations from preprints.	cometadata/affiliation-parsing-lora-Qwen3-4B
Manually annotated author and affiliation metadata form arXiv preprints	Dataset	This dataset contains manually annotated, structured metadata for a random sample of preprints from arXiv	https://huggingface.co/datasets/cometadata/arxiv-author-affiliations

Results

Project Results

[placeholder for results report]

Assessment

Project Assessment

[placeholder for community assessment of project]