Open Metrics for Global AI and Emerging Technology Research

Zachary Arnold, Catherine Aiken, James Dunham, Jennifer Melot
Center for Security and Emerging Technology, Georgetown University
Contact: zachary.arnold@georgetown.edu

Summary

We introduce an open dataset of global research, classified according to key topics in emerging technology and augmented with high-quality country and organization-level affiliation data. Currently in development, the dataset includes (i) metadata for individual publications relevant to several emerging technology topics, including artificial intelligence, NLP, computer vision, semiconductor design, robotics, cybersecurity, and AI safety;[1] (ii) counts of publications for each topic by author institution and country; and (iii) counts of research collaborations (joint publications) for each topic for institutional and country pairs. This dataset will provide accessible, methodologically transparent data to enable novel research on innovation, economic development, and technological diffusion. The dataset builds on ongoing text classification and model-based entity resolution efforts at the Center for Security and Emerging Technology (CSET) at Georgetown University. Data will be disseminated through CSET's Emerging Technology Observatory (ETO) platform and CSET's network of academic and policy collaborators. We would benefit from feedback from the I3 community as we refine the dataset, work toward a complete public version, and integrate it into our suite of open research tools.

Description of the Dataset

Open Metrics for Global AI and Emerging Technology Research is a dataset of global research relevant to emerging technology, featuring high-quality publication metadata and aggregate research output metrics for countries, institutions, and research collaborations. The dataset builds on CSET's extensive bibliometric data resources, data science and NLP expertise, and robust data and compute infrastructure. It will be a regularly updating resource, meant to integrate easily into a wide range of analyses and empower researchers (including those without extensive data science skills or computing infrastructure) to leverage bibliometric data, derive novel insights, and contribute to research on innovation and technological development.

The dataset will include, for 2012 to present:

- Publication-level tags indicating relevance to the following emerging technology topics:
    - Artificial intelligence
    - Robotics
    - Natural language processing

[1] CSET's work to identify these emerging technology research publications is supported in part by the Alfred P. Sloan Foundation under Grant No. G-2023-22358.

    - Large language models
    - Computer vision
    - AI safety
    - Cybersecurity
    - Semiconductor design and manufacturing
- Publication identifiers for open research datasets (e.g., OpenAlex)
- Counts of research publications by author(s) affiliated research institution
- Counts of research publications by author(s) affiliation country
- Counts of research publications by collaborative pairs of research institutions
- Counts of research publications by collaborative pairs of countries

We welcome input from I3 community members about other features to incorporate into the dataset. For example, additional features might include:

- Counts of highly cited research publications
- AI/ML research tasks and methods extracted from AI publications
- Publication-level field relevance scores for specific fields of study within AI, cybersecurity, biotechnology, and semiconductor design

Methodology

The dataset is derived from CSET's Merged Academic Corpus, a general-purpose bibliometric corpus maintained by CSET containing over 250 million journal articles, conference papers, and preprints obtained from open and commercial sources. The MAC currently incorporates and deduplicates data from six sources, and adds original metadata derived from CSET algorithms.

Using original models, we identify and count publications in the Merged Academic Corpus that are relevant to each of the topics covered in our dataset. CSET data scientists use various modeling approaches, including fine-tuning public-domain language models (Dunham et al. 2020, Rahkovsky et al. 2021, Schoeberl et al. 2023, Toney-Wails et al. 2024), analyzing article-topic similarity using topic-specific text embeddings (Toney and Dunham 2022), and developing topic-specific classifiers using weak supervision methods (Zhang et al. 2022). These topic-tailored modeling approaches are scalable, replicable, and can capture the intricate meanings of (and distinctions between) different topics in emerging technology. They also avoid relying on the traditional academic field taxonomies used in many other bibliometric datasets, which typically omit topics in emerging technology. Currently, our models are developed within the Snorkel Flow platform or on Google Compute Engine, with deployment automated using Apache Airflow and Beam. All models are developed and maintained in-house by CSET data scientists and software engineers.

Then, we identify the countries and institutions associated with each publication based on the institutional affiliation of the publication's author(s), and use these to generate counts of research output by country, institution, and collaborative pairs (country-country and institution-institution). Because it integrates multiple bibliometric sources, the MAC includes unusually comprehensive author and author affiliation metadata. CSET has further improved

this metadata by correcting common errors and inconsistencies in author institution metadata fields that can affect analysis, such as misspellings, abbreviations, and name variants. We perform entity resolution on research organizations by matching organizations to ROR IDs where possible, as well as manually resolving variant names for certain important organizations (e.g., highly active universities and government entities) according to defined guidelines. We are also training a series of novel large language models to improve entity resolution for author institution and country affiliations across the corpus and to integrate ROR IDs more accurately in the data. These models will be released openly. Our preliminary results suggest state-of-the-art performance in resolving authors' affiliations.

We welcome methodological feedback from I3 community members.

Value to the Research Community

Our goal with this dataset is to enable researchers beyond CSET to make sense of today's rapidly evolving research and technology landscape. This dataset offers value to the research community by openly providing data sourced, curated, and maintained by a 10+ person data science and engineering team. It will provide an off-the-shelf option for researchers looking to analyze research related to high-profile topics such as AI, cybersecurity, and semiconductor design, or analyze collaborative research on these topics among countries and leading institutions. Critically, the country- and institution-level metrics in our dataset are pre-computed and ready to incorporate into analysis. This makes them especially useful for researchers without the resources, expertise, or technical infrastructure to use large open-source bibliometric datasets or expensive subscription services.

The dataset will enable research into questions of significant interest to scholars of innovation, economic development, and technological diffusion. For instance, researchers may want to understand the economic effects of AI R&D activity, analyze the conditions under which AI models are prone to bias or pose safety risks, explore how AI methods are being employed to advance science and discovery, understand impacts of AI adoption for the workforce and other economic outcomes, or assess how investments in specific emerging technology domains may or may not foster innovation.

Our team is dedicated to providing an open resource developed through reproducible research and interoperable approach. We use peer-reviewed methods and document and publish our methodology. We also structure the dataset to be able to "plug in" to other open resources, such as OpenAlex, ROR, and Semantic Scholar.

Plans for Public Dissemination

The complete dataset will be maintained by CSET's data science team and disseminated through our public ETO platform (https://eto.tech), as well as online data-sharing platforms such

as Zenodo and GitHub, as appropriate. We will publicize the release of the dataset through existing outreach programs and CSET and ETO newsletters and social media feeds.

The dataset will be updated automatically using Apache Airflow pipelines that leverage various components of CSET's Google Cloud infrastructure along with Python data cleaning and normalization scripts. These pipelines will include automated data quality checks that upon failure will halt the update of production datasets and raise an alert for manual review. We also plan to automate nomination of records that need manual cleaning by human annotators using our existing Airflow pipeline templates that automatically ingest records that need annotation and integrate the results into our data warehouse.

In addition to being openly available for researchers, the dataset will provide the basis for enhancements to ETO's open analytic tools, which are widely used by scholars for initial ideation, orientation to new subject matter domains, and higher-level analysis. These tools include the Map of Science, a visualization of the global research literature based on citation networks, and the Country Activity Tracker, a customizable dashboard of international AI research, patent, and investment activity, including international research collaboration.

The dataset will include detailed documentation and methodological descriptions, similar to documentation for ETO's existing data resources. We will also post code related to data pipelines, enrichment, aggregation, and model training and deployment in public GitHub repositories under CSET's standard Creative Commons Attribution-Non Commercial 4.0 International License, with limited exceptions related to third-party vendor content as specified in our Terms of Use.

References

Dunham, James, Jennifer Melot, and Dewey Murdick. "Identifying the Development and Application of Artificial Intelligence in Scientific Text." arXiv, May 28, 2020. http://arxiv.org/abs/2002.07143.

Rahkovsky, Ilya, Autumn Toney, Kevin W. Boyack, Richard Klavans, and Dewey A. Murdick. "AI Research Funding Portfolios and Extreme Growth." *Frontiers in Research Metrics and Analytics* 6 (April 6, 2021): 630124. https://doi.org/10.3389/frma.2021.630124.

Schoeberl, Christian, Autumn Toney, and James Dunham. "Identifying AI Research." CSET Data Brief (July 2023). https://cset.georgetown.edu/publication/identifying-ai-research/.

Toney, Autumn, and James Dunham. "Multi-Label Classification of Scientific Research Documents Across Domains and Languages." In *Proceedings of the Third Workshop on Scholarly Document Processing*, 105–14. Gyeongju, Republic of Korea: Association for Computational Linguistics, 2022. https://aclanthology.org/2022.sdp-1.12.

Toney-Wails, Autumn, Christian Schoeberl, and James Dunham. "AI on AI: Exploring the Utility of GPT as an Expert Annotator of AI Publications." arXiv, March 14, 2024. https://doi.org/10.48550/arXiv.2403.09097.

Zhang, Jieyu, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. “A Survey on Programmatic Weak Supervision.” arXiv, February 14, 2022. https://doi.org/10.48550/arXiv.2202.05433.