Asynchronous and Modular Pipelines for Fast WARC Annotation

Pedro Ortiz Suarez, Thom Vaughan

Motivation

- Libraries, museums, cultural heritage organizations and web archiving communities have massive repositories of archived web data in the form of WARC and ARC files [1, 2]
- The ever growing demand for data from Large Language Models (LLMs) has led to the NLP/ML community taking an increased interest in web archives as the main dataset for pre-training LLMs [3, 4, 5]
- This interest has led NLP researchers to develop pipelines to annotate, filter, and select data from WARC files in order to extract "relevant" web documents and use them in LLM training
- Moreover, language models are now easier and more cost-effective to use, allowing their implementation in data pipelines that handle raw data directly and allowing better data retrieval [6]

Existing Pipelines for WARC Annotation

- **From the NLP/ML community**
  - **CCNet**: n-gram models [7] as data quality signal [8]
  - **OSCAR**: LangID precision as data quality signal [9, 10]
  - **mC4**: uses word blocklists to remove adult content [11]
  - **Refined Web**: focuses on text extraction [12, 13]
  - **DataTrove**: aggregates filters of other pipelines [14]
  - **Dolma**: focuses on fuzzy deduplication [15]
  - **NeMo-Curator**: combines many heuristics and filters [16]
  - **HPLT**: focuses on bilingual/parallel data extraction [17]
  - **MADLAD**: focuses on LangID [18] and multilinguality [19]
  - **FineWeb**: has an educational content classifier [20]
- **From the web archiving community**
  - **ArchiveSpark**: a framework facilitating efficient data processing for archival collections. It focuses on selection, extraction and knowledge graph creation [21]
  - **The Archives Unleashed Toolkit**: builds on top of ArchiveSpark, to decompose scholarly inquiries into four main activities: filter, extract, aggregate, and visualize [22]
  - **Archives Research Compute Hub (ARCH)**: handles non-textual content and provides a user-friendly interface, building on top of The Archives Unleashed Toolkit [23]

Limitations of WARC Annotation Pipelines

- **From the NLP/ML community**
  - Most are based on research code, remaining unmantained, unoptimized, and sometimes unreleased [8, 11, 19]
  - Given their focus on producing training data, they filter instead of annotate
  - Frequently discard all metadata, focusing only on the raw data
- **From the web archiving community**
  - Focus mainly on metadata and existing annotations for the selection and filtering operations
  - Often have multiple dependencies, making them difficult to install

About Common Crawl

Common Crawl is a non-profit organization which regularly crawls a significant sample of the web and makes the data accessible free of charge to everyone interested in running machine-scale analysis on web data.

At present, we crawl up to 3.0 billion web pages every month. The data is hosted in the Amazon cloud as part of the AWS Open Data program.

Contact:
https://commoncrawl.org/
pedro@commoncrawl.org
thom@commoncrawl.org

A First Prototype for a cc-annotator

- We are inspired by existing pipelines from the NLP and the web archiving communities, and develop our own prototype of a pipeline for WARC annotation
- The first experimental prototype aims to be efficient, modular, open-source, and user-friendly, so that little knowledge of large-scale data processing is needed
- This first prototype focuses on organizing and compiling some of the rapid developments of data pipelines in the NLP/ML space and bringing them to the web archiving community
- Our prototype focuses on textual data, as Common Crawl is a text-only archive
- However, we note that the architecture of our prototype is extensible and borrows ideas from NLP pipelines that have already been extended to non-textual archives [10, 24]
- For this first prototype we only implement **text extraction** and **language annotations**. Common Crawl already distributes text extractions as WET (WARC Encapsulated Text) files [25] and language annotations [26], but the NLP community has already developed more robust models for these tasks that we would like to explore [13, 27]

Architecture

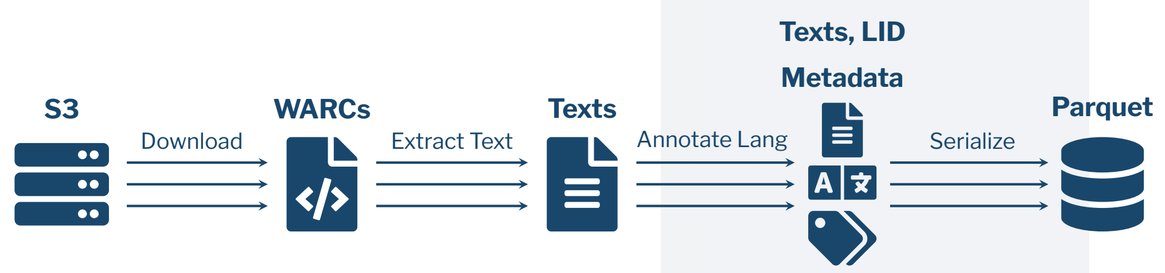

Figure 1. A schema of the first prototype for our pipeline. represents the AWS S3 bucket where the Common Crawl data is stored, represents WARC files, represents text extractions of WARCs, represents language tags, represents metadata coming from WARC headers and represents Parquet files. Finally the arrows represent parallel processing.

- We build our prototype using `Rust` [28]
  - Making the pipeline fast and memory-safe while allowing true parallelism
  - Allowing us to distribute dependency-free binaries for all major platforms
  - Making it possible to easily port code from existing NLP pipelines also built in `Rust` [15, 29]
- We preserve all documents and all metadata contained in the WARC headers
- We use an asynchronous runtime in order to reduce latency and avoid waiting for I/O operations [30]
- Outside AWS, we stream all data using `cc-downloader`, the official Common Crawl download client [31]. Within AWS we use their official SDK to access S3 [32]
- We use modern LangID models, covering 200+ languages [27, 33], compared to the 160 supported by the CLD2 model [34] used by Common Crawl [26]
- We test some of the text extraction algorithms preferred by the NLP community [13, 35, 36, 37]
- We serialize our outputs in Parquet [38], making them compatible with the existing Common Crawl index [39]

Links / Resources

Bibliography, source code and data:
`https://github.com/commoncrawl/wac2025-cc-annotator-poster`

IIPC Web Archiving Conference 2025, 8 – 10 April 2025, Oslo, Norway

COMMON CRAWL

Asynchronous and Modular Pipelines for Fast WARC Annotation

Pedro Ortiz Suarez, Thom Vaughan

References

[1] The Members of the IIPC. *The WARC Format 1.1*. 2025. URL: https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/.

[2] Mike Burner and Brewster Kahle. *Arc File Format*. 1996. URL: https://archive.org/web/researcher/ArcFileFormat.php.

[3] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[4] Niklas Muennighoff et al. "Scaling Data-Constrained Language Models". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 50358–50376. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/9d89448b63ce1e2e8dc7af72c984c196-Paper-Conference.pdf.

[5] Hugo Touvron et al. "LLaMA: Open and Efficient Foundation Language Models". In: *arXiv e-prints*, arXiv:2302.13971 (Feb. 2023), arXiv:2302.13971. DOI: 10.48550/arXiv.2302.13971. arXiv: 2302.13971 [cs.CL].

[6] Rodrigo Nogueira and Kyunghyun Cho. "Passage Re-ranking with BERT". In: *arXiv e-prints*, arXiv:1901.04085 (Jan. 2019), arXiv:1901.04085. DOI: 10.48550/arXiv.1901.04085. arXiv: 1901.04085 [cs.IR].

[7] Kenneth Heafield. "KenLM: Faster and Smaller Language Model Queries". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch et al. Edinburgh, Scotland: Association for Computational Linguistics, July 2011, pp. 187–197. URL: https://aclanthology.org/W11-2123/.

[8] Guillaume Wenzek et al. "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data". eng. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, May 2020, pp. 4003–4012. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.494/.

[9] Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. "A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 1703–1714. DOI: 10.18653/v1/2020.acl-main.156. URL: https://aclanthology.org/2020.acl-main.156/.

[10] Julien Abadji et al. "Towards a Cleaner Document-Oriented Multilingual Crawled Corpus". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 4344–4355. URL: https://aclanthology.org/2022.lrec-1.463/.

[11] Linting Xue et al. "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, June 2021, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41. URL: https://aclanthology.org/2021.naacl-main.41/.

[12] Guilherme Penedo et al. "The RefinedWeb dataset for falcon LLM: outperforming curated corpora with web data only". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23. New Orleans, LA, USA: Curran Associates Inc., 2023.

[13] Adrien Barbaresi. "Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Ed. by Heng Ji, Jong C. Park, and Rui Xia. Online: Association for Computational Linguistics, Aug. 2021, pp. 122–131. DOI: 10.18653/v1/2021.acl-demo.15. URL: https://aclanthology.org/2021.acl-demo.15/.

[14] Hugging Face. *DataTrove Library*. URL: https://github.com/huggingface/datatrove.

[15] Luca Soldaini et al. "Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15725–15788. DOI: 10.18653/v1/2024.acl-long.840. URL: https://aclanthology.org/2024.acl-long.840/.

[16] NVIDIA. *NeMo Curator*. URL: https://github.com/NVIDIA/NeMo-Curator.

[17] Ona de Gibert et al. "A New Massive Multilingual Dataset for High-Performance Language Technologies". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 1116–1128. URL: https://aclanthology.org/2024.lrec-main.100/.

[18] Isaac Caswell et al. "Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 6588–6608. DOI: 10.18653/v1/2020.coling-main.579. URL: https://aclanthology.org/2020.coling-main.579/.

[19] Sneha Kudugunta et al. "MADLAD-400: A Multilingual And Document-Level Large Audited Dataset". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 67284–67296. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/d49042a5d49818711c401d34172f9900-Paper-Datasets_and_Benchmarks.pdf.

[20] Guilherme Penedo et al. "The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale". In: *arXiv e-prints*, arXiv:2406.17557 (June 2024), arXiv:2406.17557. DOI: 10.48550/arXiv.2406.17557. arXiv: 2406.17557 [cs.CL].

[21] Helge Holzmann, Vinay Goel, and Sawood Alam. *ArchiveSpark*. https://github.com/helgeho/ArchiveSpark. 2024.

[22] Nick Ruest et al. "The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives". In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. JCDL '20. Virtual Event, China: Association for Computing Machinery, 2020, pp. 157–166. ISBN: 9781450375856. DOI: 10.1145/3383583.3398513. URL: https://doi.org/10.1145/3383583.3398513.

[23] Helge Holzmann et al. "ABCDEF: the 6 key features behind scalable, multi-tenant web archive processing with ARCH: archive, big data, concurrent, distributed, efficient, flexible". In: *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*. JCDL '22. Cologne, Germany: Association for Computing Machinery, 2022. ISBN: 9781450393454. DOI: 10.1145/3529372.3530916. URL: https://doi.org/10.1145/3529372.3530916.

[24] Matthieu Futeral et al. "mOSCAR: A Large-scale Multilingual and Multimodal Document-level Corpus". In: *arXiv e-prints*, arXiv:2406.08707 (June 2024), arXiv:2406.08707. DOI: 10.48550/arXiv.2406.08707. arXiv: 2406.08707 [cs.CL].

[25] Thom Vaughan. *Web Archiving File Formats Explained*. URL: https://commoncrawl.org/blog/web-archiving-file-formats-explained.

[26] Sebastian Nagel. *August Crawl Archive Introduces Language Annotations*. URL: https://commoncrawl.org/blog/august-2018-crawl-archive-now-available.

[27] Laurie Burchell et al. "An Open Dataset and Model for Language Identification". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 865–879. DOI: 10.18653/v1/2023.acl-short.75. URL: https://aclanthology.org/2023.acl-short.75/.

[28] The Rust Team. *The Rust Programming Language*. 2025. URL: https://www.rust-lang.org.

[29] Julien Abadji et al. "Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus". en. In: *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event)*. Ed. by Harald Lüngen et al. Mannheim: Leibniz-Institut für Deutsche Sprache, 2021, pp. 1–9. DOI: 10.14618/ids-pub-10468. URL: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-104688.

[30] Tokio. *Tokio*. https://github.com/tokio-rs/tokio. 2025.

[31] Common Crawl Foundation and contributors. *cc-downloader*. 2025. URL: https://github.com/commoncrawl/cc-downloader.

[32] Amazon Web Services, Inc. or its affiliate. *Amazon S3 examples using SDK for Rust*. URL: https://docs.aws.amazon.com/sdk-for-rust/latest/dg/rust_s3_code_examples.html.

[33] Amir Hossein Kargaran et al. "GlotLID: Language Identification for Low-Resource Languages". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6155–6218. DOI: 10.18653/v1/2023.findings-emnlp.410. URL: https://aclanthology.org/2023.findings-emnlp.410/.

[34] CLD2 Owners. *Compact Language Detector 2*. https://github.com/CLD2Owners/cld2. 2015.

[35] Janek Bevendorff et al. "Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl". In: *Advances in Information Retrieval. 40th European Conference on IR Research (ECIR 2018)*. Ed. by Leif Azzopardi et al. Lecture Notes in Computer Science. Berlin Heidelberg New York: Springer, Mar. 2018.

[36] Janek Bevendorff, Martin Potthast, and Benno Stein. "FastWARC: Optimizing Large-Scale Web Archive Analytics". In: *3rd International Symposium on Open Search Technology (OSSYM 2021)*. Ed. by Andreas Wagner et al. International Open Search Symposium, Oct. 2021.

[37] Janek Bevendorff et al. "An Empirical Comparison of Web Content Extraction Algorithms". In: *46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*. Ed. by Hsin-Hsi Chen et al. ACM, July 2023, pp. 2594–2603. ISBN: 9781450394086. DOI: 10.1145/3539618.3591920.

[38] Wikipedia contributors. *Apache Parquet — Wikipedia, The Free Encyclopedia*. 2025. URL: https://en.wikipedia.org/wiki/Apache_Parquet.

[39] Sebastian Nagel. *Index to WARC Files and URLs in Columnar Format*. URL: https://commoncrawl.org/blog/index-to-warc-files-and-urls-in-columnar-format.