UNKNOWNDATA: MINING AND CONSOLIDATING RESEARCH DATASET METADATA

LONG ABSTRACT

Keywords Data Discovery · Research Data · FAIR principles

Workshop Theme Research data is vital to facilitate the scientific progress in various domains. Being able to find and share research data paves a concrete way to data reuse, scientific reproduction and verification; meanwhile, the sharing and access of research data are recognized by national and international initiatives, e.g. OPENAIRE[1], Nationale Forschungsdateninfrastructur[2]. Bearing this in mind, UnknownData project[3] aims at mining dataset metadata in computer science and social science domains on the web, and then provide consolidated dataset metadata via integration to DBLP and GESIS search.

UnknownData is decomposed into five work packages: user study pilots (WP1) , web crawling focused information discovery (WP2) , dataset metadata extraction (WP3) , dataset metadata reconciliation and consolidation (WP4) , and integrations to DBLP/GESIS (WP5) . WP1 conducts pilot studiesBeckermann [2023] on GESIS DataSearch and DBLP for social science and computer science dataset metadata, where specific metadata cleaning rules are created to improve metadata quality in social science and data publication interface is incorporated on DBLP. WP2 develops various web crawling strategies, i.e. link-based, content-based judgement and web archivesJäschke [2023], to allow dataset relevant web page discovery on the web. WP3 proposes dataset extraction approachesYounes and Mathiak [2022], Younes and Scherp [2023], Otto et al. [2023] to extract dataset mention and other relevant metadata from the web pages discovered in WP2 and research publications. WP4 disambiguate the identified dataset metadata from the web and research publications from WP3 and consolidate and link the newly identified dataset metadata to existed scholarly knowledge bases. Finally, the consolidated dataset metadata is mapped with ontologies of DBLP and GESIS Search and is provided to enrich the two platforms.

A working prototype of UnknownData has been developed and implemented corresponding to the different work packages described above. The prototype pipeline contains web crawling, dataset extraction, metadata consolidation and linking, and the final integration, including both social science and computer science use cases from the web and research paper setting. We present this prototype to showcase how discovering unknown dataset metadata from the web can be done and how it can be smoothly integrated into current data sharing portals.

Workshop Agenda

00:00 - 00:30	30-min interactive presentation of UnknownData project
00:30 - 00:50	20-min demo of UnknownData prototype
00:50 - 01:20	30-min question and comment collection from the attendees
01:20 - 01:50	30-min group discussions of questions and comments
01:50 - 02:00	10-min conclusion

Contact Information Dr. Brigitte Mathiak, `Brigitte.mathiak@gesis.org`

[1] `https://www.openaire.eu/`

[2] `https://www.akademienunion.de/fileadmin/redaktion/user_upload/Publikationen/Positionspapiere/NFDI-Positionspapier_Akademienunion_01.pdf`

[3] UnknownData is funded by German Research Foundation, program e-Research Technologies, `https://unknowndataproject.github.io/`

References

Bennedikt Beckermann. Dataset publications in dblp, April 2023. URL https://blog.dblp.org/2023/04/27/dataset-publications-in-dblp/.

Robert Jäschke. Tales from the inside: 10 years of growing and maintaining a multi-terabyte longitudinal archive of web pages and tweets, July 2023. URL https://www.fontanearchiv.de/veranstaltung/2023/07/7/do-it-yourself-archives.

Yousef Younes and Brigitte Mathiak. Handling class imbalance when detecting dataset mentions with pre-trained language models. In Mourad Abbas and Abed Alhakim Freihat, editors, *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 79–88, Trento, Italy, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.icnlsp-1.9.

Yousef Younes and Ansgar Scherp. Question answering versus named entity recognition for extracting unknown datasets. *IEEE Access*, 11:92775–92787, 2023. doi:10.1109/ACCESS.2023.3309148.

Wolfgang Otto, Matthäus Zloch, Lu Gan, Saurav Karmakar, and Stefan Dietze. GSAP-NER: A novel task, corpus, and baseline for scholarly entity extraction focused on machine learning models and datasets. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8166–8176, Singapore, December 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.findings-emnlp.548. URL https://aclanthology.org/2023.findings-emnlp.548.