Softalias-KG: Reconciling Software Mentions in Scientific Literature

Esteban González-Guardia[1], Hector Lopez[1] and Daniel Garijo[1]

[1]Ontology Engineering Group, Universidad Politécnica de Madrid, Spain

Abstract

Research software (i.e., the tools and scripts developed as part of a research investigation) are key to support the results described in academic publications. However, current citation practices followed by researchers make it difficult to automatically identify and reconcile the tools used in existing publications (different names are used for the same tool, different citing styles, etc.). In this demo we address this issue by integrating SofCite, a state of the art named entity recognition model designed to find software mentions in the biomedical domain, with Softalias-KG, a Knowledge Graph of software aliases, in order to reconcile software tools in a text.

Demo URL: https://w3id.org/softalias/demo

Code: https://github.com/SoftwareUnderstanding/softalias-rs

Keywords

software alias, software metadata, software reconciliation, knowledge graph

1. Introduction

Research Software [1] is increasingly becoming a first-class citizen research product due to its role in supporting computational results.[1] Open platforms like Papers with Code[2] or OpenAire[3] are dedicated to tracking the links between research articles and code, and paper preprint archives like Arxiv[4] are starting to display such information to readers.

In order to ease software citation, the scientific community has developed guidelines and software citation formats for developers [2]. However, researchers often use different names to refer to the tools they refer to in their work. For example, the "Statistical Package for the Social Sciences"[5] is also known as "SPSS" and "SPSS Statistics" among many other variations. This makes tool reconciliation challenging, making it difficult to provide tool developers their corresponding credit.

In this work introduce Softalias-KG, a Knowledge Graph of software aliases created from a recent analysis and software mention dataset of over 3.8 million papers from PubMed Central [3].

ISWC 2023 Posters and Demos: 22nd International Semantic Web Conference, November 6–10, 2023, Athens, Greece

egonzalez@fi.upm.es (E. González-Guardia); daniel.garijo@upm.es (D. Garijo)

0000-0003-4112-6825 (E. González-Guardia); 0000-0003-0454-7145 (D. Garijo)

[1]https://sfdora.org/read/

[2]https://paperswithcode.com/

[3]https://graph.openaire.eu/

[4]https://arxiv.org/

[5]https://www.ibm.com/products/spss-statistics

Our demo uses Softalias-KG as a reconciliation service for similar tool mentions, which we have integrated with a state of the art Named Entity Recognition (NER) model (Softcite [4]) trained in the biomedical domain to extract software mentions.

2. Softalias-KG: A Knowledge Graph of Software Aliases

We base our work on [3], a recent analysis from the Chan-Zuckerberg Initiative (CZI) where the authors extracted software mentions from over 3.8 million papers in PubMed Central,[6] making their results available online [5]. In the paper, the authors use a clustering algorithm based on Jaro-Winkler distance [6] for grouping similar software mentions, finding the most likely package managers (Pypi,[7] CRAN,[8] Bioconductor[9]) software registries (SciCrunch[10]) and repositories (GitHub[11]) where a cluster of mentions may be linked to. For each potential link, basic metadata of the software mention is downloaded from the corresponding platform (description, package URL, identifier, etc.). Unfortunately, the results are stored in pickle and csv files, designed to be used in notebooks. We have cleaned and transformed these results into the Softalias-KG, focusing on the clustering analysis of software mentions (i.e., software aliases) to facilitate reconciliation through SPARQL queries. We have also expanded the results with software entities from Wikidata [7], in order to generalize the application domain of the KG.

2.1. Representing Aliases and Groups

Softalias-KG includes two main types of entities: software *aliases*, i.e., software mentions as detected in any scientific publication, and *groups*, which represent single software applications (*schema:SoftwareApplication*) grouping a cluster of aliases based on [3]. Groups are described with a canonical name, based on their most frequent alias, and are described with metadata such as URL, license, etc. All metadata are obtained from the information found in external sources (Pypi, CRAN, etc.) and represented using Schema.org [8]. The metadata platform used to enrich all groups is also kept in the KG (*schema:provider*), in order preserve the provenance of each record. Groups are linked with their corresponding aliases with the *alias* property (a software group has one or more aliases).

2.2. RDF Transformation and Cleaning

The CZI dataset [3] consists on two sets of files. A first set of pickle files describe the software mention groups matched to each external platform (Pypi, CRAN, etc.), together with a file assigning an id to each software mention in a publication. A second set of CSV files contain the metadata for each software entry on each external platform.

We unified all pickle files (*.pkl*) in a single JSON representation, counting the number of times a software alias is used in the literature. Next, we removed mentions with null repetitions,

[6] https://www.ncbi.nlm.nih.gov/pmc/
[7] https://pypi.org/
[8] https://cran.r-project.org/
[9] https://contributions.bioconductor.org/index.html
[10] https://scicrunch.org/
[11] https://github.com/

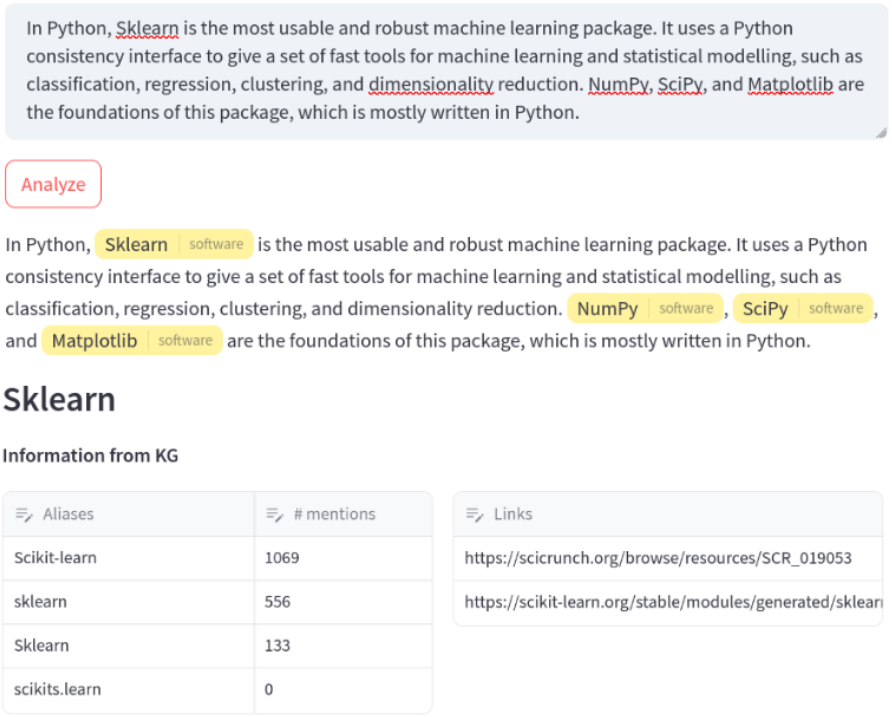

Figure 1: Snippet of our demo application, showing how users may enter text (top), the software mentions found by Softcite (in yellow) and the corresponding URLs found in Softalias-KG.

removed aliases with null ids and groups containing no aliases. Upon further inspection, we also removed the mention links matched to GitHub from the first version of the KG, due to inconsistencies found in the reconciliation. The set of CSV files were also cleaned by removing empty quotes and parenthesis. The final JSON and CSV files were then mapped with YARRRML[12] and converted into RDF using Morph-KGC [9].[13] The final Knowledge Graph contains over 50K aliases which correspond to over 34K unique software application groups.

We have also enriched Softalias-KG with software entities and metadata from Wikidata, one of the largest crowdsourced Knowledge Graphs to date. In particular, we have imported over 8K software applications (free software, *wd:Q341*), with their corresponding 3K alternative labels (*skos:altLabel*).

3. Demonstration: Reconciling Software Mentions

Our demo uses Softalias-KG as a software reconciliation and metadata service. Figure 1 shows an overview of our demo, where users may enter a paragraph of text to search for software mentions (top part of the figure). After clicking on the "Analyze" button, our service runs Softcite [10, 4] (version 0.7.1), a named entity recognition model trained on over 5K open

[12] https://rml.io/yarrrml/
[13] https://github.com/morph-kgc/morph-kgc

research papers in the biomedical and economics domains. Softcite returns a list of candidate software mentions, which are highlighted in yellow in the text below the "Analyze" button.

Each found mention is then used to query the Softalias-KG in order to find the canonical name for that software application. Internally, a SPARQL query retrieves aliases with the mention name, returning the groups that the alias belongs to, as well as additional metadata like the URL where the software may be found (from Pypi, CRAN, SciCrunch or Wikidata). We also retrieve other aliases used for that software application, along with their number of mentions in scientific literature (when available). Our demo,[14] code [11][15] and mappings [12][16] are available online.

4. Conclusions and Future Work

This demo integrates a state of the art named entity recognition model for detecting software mentions in text (SoftCite) with Softalias-KG, a novel Knowledge Graph of software aliases that is used for tool reconciliation. Softalias-KG is based on an existing analysis in the scientific literature over millions of paper in the biomedical domain [3], contains more than 50K unique aliases (belonging to more than 30K unique tools) and is enriched with over 8K tools from Wikidata. Thanks to our demo, we can easily navigate through existing tool aliases, as well as detect of potential potential omissions in the NER model and our Knowledge Graph.

Our next steps will focus on identifying and addressing clustering errors (e.g., by looking into software groups with different tool URLs, joining software groups sharing the same aliases, etc.) and expanding Softalias- KG with Somesci KG [13], another KG of software mentions of smaller size that is already aligned with Wikidata.

Acknowledgments

This work is supported by the Madrid Government (Comunidad de Madrid - Spain) under the Multiannual Agreement with Universidad Politécnica de Madrid in the line Support for R&D projects for Beatriz Galindo researchers, in the context of the VPRICIT, and through the call Research Grants for Young Investigators from Universidad Politécnica de Madrid.

References

[1] N. P. Chue Hong, D. S. Katz, M. Barker, A.-L. Lamprecht, C. Martinez, F. E. Psomopoulos, J. Harrow, L. J. Castro, M. Gruenpeter, P. A. Martinez, et al., FAIR Principles for Research Software (FAIR4RS Principles), 2022. doi:10.15497/RDA00068.

[2] S. Druskat, J. H. Spaaks, N. Chue Hong, R. Haines, J. Baker, S. Bliven, E. Willighagen, D. Pérez-Suárez, O. Konovalov, Citation File Format, 2021. doi:10.5281/zenodo.5171937.

[3] A.-M. Istrate, D. Li, D. Taraborelli, M. Torkar, B. Veytsman, I. Williams, A large dataset of software mentions in the biomedical literature, 2022. arXiv:2209.00693.

[14] Demo: https://w3id.org/softalias/demo

[15] Demo code: https://github.com/SoftwareUnderstanding/softalias-rs

[16] Mappings and transformation scripts: https://github.com/SoftwareUnderstanding/softalias-kg

[4] P. Lopez, C. Du, J. Cohoon, K. Ram, J. Howison, Mining software entities in scientific literature: Document-level NER for an extremely imbalance and large-scale task, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, ACM, New York, USA, 2021, p. 3986–3995. doi:10.1145/3459637.3481936.

[5] A.-M. Istrate, B. Veytsman, D. Li, D. Taraborelli, M. Torkar, I. Williams, CZ Software Mentions: A large dataset of software mentions in the biomedical literature, 2022. doi:10.5061/DRYAD.6WWPZGN2C.

[6] W. Winkler, String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage, Proceedings of the Section on Survey Research Methods (1990). URL: https://eric.ed.gov/?id=ED325505.

[7] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85.

[8] R. V. Guha, D. Brickley, S. Macbeth, Schema. org: evolution of structured data on the web, Communications of the ACM 59 (2016) 44–51.

[9] J. Arenas-Guerrero, D. Chaves-Fraga, J. Toledo, M. S. Pérez, O. Corcho, Morph-KGC: Scalable knowledge graph materialization with mapping partitions, Semantic Web (2022). doi:10.3233/SW-223135.

[10] Softcite software mention recognizer, https://github.com/ourresearch/software-mentions, 2018–2021.

[11] E. González-Guardia, Softwareunderstanding/softalias-rs: Code used for iswc2023 demo, 2023. doi:10.5281/zenodo.8338240.

[12] D. Garijo, H. Lopez, SoftwareUnderstanding/softalias-kg: Softalias-kg: First release of KG transformation scripts, 2023. doi:10.5281/zenodo.8341333.

[13] D. Schindler, F. Bensmann, S. Dietze, F. Krüger, Somesci- a 5 star open data gold standard knowledge graph of software mentions in scientific articles, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, ACM, New York, NY, USA, 2021, p. 4574–4583. doi:10.1145/3459637.3482017.