Meta2KG: Transforming Metadata to Knowledge Graphs

Nora Abdelmageed[1,2,3], Birgitta König-Ries[1,2,3]

[1]Heinz Nixdorf Chair for Distributed Information Systems
[2]Michael Stifel Center Jena
[3]Friedrich Schiller University Jena, Jena, Germany

Abstract
Metadata is used to describe data. It includes information about the who, when, where, how, and why of data collection. Ideally, it should be in a machine-understandable format like RDF. This enables queries using structured query languages like SPARQL and empowers further data usage. In this paper, we investigate metadata as a source for generating Knowledge Graphs (KGs). We introduce a fully automatic approach that transforms raw metadata files into a Knowledge Graph (KG). Our resources and code are publicly available[1].

Keywords
Metadata Analysis, RDF, Matching, Knowledge Graph, Embeddings

1. Introduction

Knowledge Graphs (KGs) are widely used to represent information about entities of interest and their relations [1]. Lately, this includes information encoded in scientific datasets. Often, these datasets are accompanied by metadata describing the who, when, where, how, and why of data collection. Transforming metadata into KGs increases the FAIRness [2] of the data by enhancing its reusability.

Embeddings are a well-established technique that captures the semantics of a given word or sentence. Previous works have shown their significant impact on many Natural Language Processing (NLP) applications [3]. In this work, we transform raw metadata files into a KG using an embedding-based matching technique. We tested our technique on a biodiversity use case; however, we expect our method to be domain-independent.

2. Methodology

Figure 1 shows the four phases of our pipeline. 1) **Data Acquisition** We collected our metadata files from various biodiversity data portals to develop the data model and evaluate our matching technique. 2) **Ontology Development** The data-driven process of crafting our data model

[1]https://github.com/fusion-jena/Meta2KG

Ontology Matching @ISWC 2022
nora.abdelmageed@uni-jena.de (N. Abdelmageed); birgitta.koenig-ries@uni-jena.de (B. König-Ries)
0000-0002-1405-6860 (N. Abdelmageed); 0000-0002-2382-9722 (B. König-Ries)

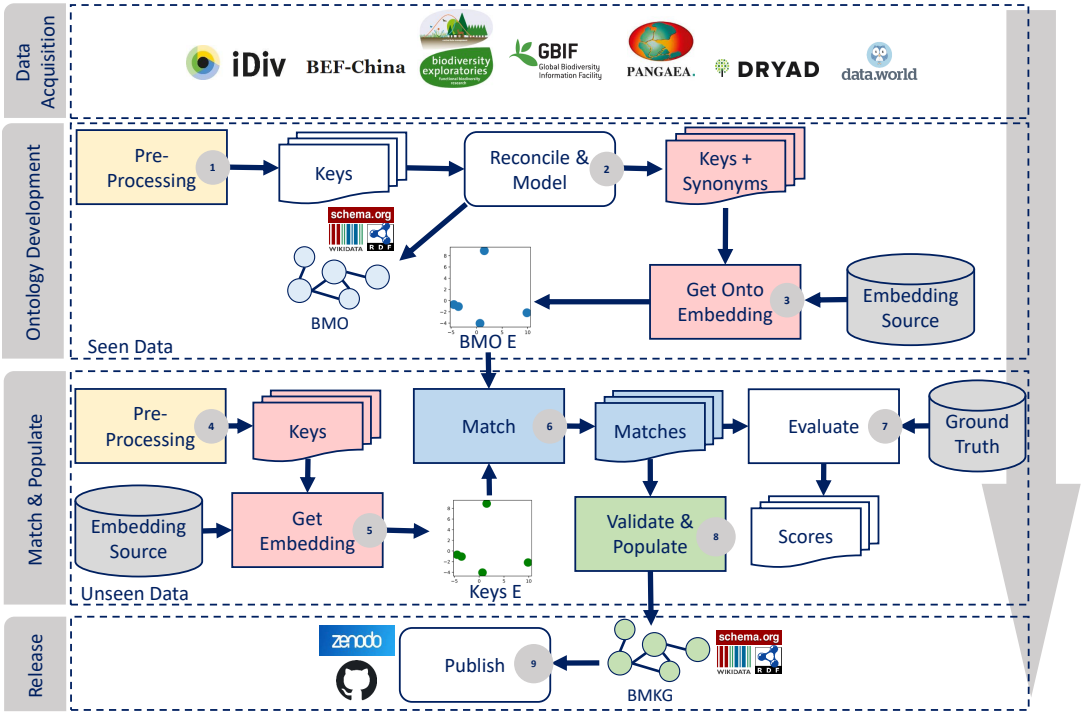

Figure 1: Abstract overview of our workflow to transform raw metadata to KG.

(Biodiversity Metadata Ontology (BMO)). We applied several cleaning steps to the collected data. During this phase, we held several meetings with a biodiversity expert to validate and review our conceptual model. In addition, we developed mean-based techniques to transform BMO to the embedding space (BMOE). 3) **Match & Populate** Our unsupervised learning methods for ontology matching and instance population. For matching, we used cosine similarity in the embedding space between the ontological embeddings, *BMO E*, and metadata embeddings, *Keys E*. We used embeddings to capture the semantic meaning of words. For population, We limit the population to a triple if and only if its value has the expected datatype. For example, we accept the triple, e.g., (author, *phone*, XXX) if "XXX" is a phone. We implemented such kind of validations using regular expressions. 4) **Release** We published our resources and code under the Creative Commons Attribution 4.0 International (CC BY 4.0) and Apache License 2.0, respectively.

Acknowledgments

The authors thank the Carl Zeiss Foundation for the financial support of the project "A Virtual Werkstatt for Digitization in the Sciences (K3, P5)" within the scope of the program line "Breakthroughs: Exploring Intelligent Systems for Digitization" - explore the basics, use applications". In addition, we thank, Cornelia Fürstenau, Sirko Schindler, Muhammad Abbady, and Jan Martin Keil for the fruitful discussions.

References

[1] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutiérrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge Graphs, Synthesis Lectures on Data, Semantics, and Knowledge, Morgan & Claypool Publishers, 2021. doi:10.2200/S01125ED1V01Y202109DSK022.

[2] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016). doi:10.1038/sdata.2016.18.

[3] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguistics 5 (2017) 135–146. URL: https://doi.org/10.1162/tacl_a_00051. doi:10.1162/tacl_a_00051.