Beyond DBpedia and YAGO
The New Kids on the Knowledge Graph Block

Heiko Paulheim[1][0000−0003−4386−8195]

University of Mannheim, Germany
Data and Web Science Group
heiko@informatik.uni-mannheim.de

Abstract. Starting with Cyc in the 1980s [6], the collection of general knowledge in machine interpretable form has been considered a valuable ingredient in intelligent and knowledge intensive applications. Notable contributions in the field include the Wikipedia-based datasets DBpedia [5] and YAGO [10], as well as the collaborative knowledge base Wikidata [11]. Since Google has coined the term in 2012, they are most often referred to as *knowledge graphs* [1, 8]. Besides such open knowledge graphs, many companies have started using corporate knowledge graphs as a means of information representation [7].
In this talk, I will look at two ongoing projects related to the extraction of knowledge graphs from Wikipedia and other Wikis. The first new dataset, *CaLiGraph*[1], aims at the generation of explicit formal definitions from categories [2], and the extraction of new instances from list pages [9]. In its current release, CaLiGraph contains 200k axioms defining classes, and more than 7M typed instances.
In the second part, I will look at the transfer of the DBpedia approach to a multitude of arbitrary Wikis. The first such prototype, *DBkWik*[2], extracts data from Fandom, a Wiki farm hosting more than 400k different Wikis on various topics. Unlike DBpedia, which relies on a larger user base for crowdsourcing an explicit schema and extraction rules, and the "one-page-per-entity" assumption, DBkWik has to address various challenges in the fields of schema learning and data integration [3, 4]. In its current release, DBkWik contains more than 11M entities, and has been found to be highly complementary to DBpedia.

References

1. Ehrlinger, L., Wöß, W.: Towards a definition of knowledge graphs. In: SEMANTiCS (2016)
2. Heist, N., Paulheim, H.: Uncovering the semantics of wikipedia categories. In: International semantic web conference. pp. 219–236. Springer (2019)
3. Hertling, S., Paulheim, H.: Dbkwik: A consolidated knowledge graph from thousands of wikis. In: 2018 IEEE International Conference on Big Knowledge (ICBK). pp. 17–24. IEEE (2018)

[1] http://caligraph.org/
[2] http://dbkwik.org/

4. Hofmann, A., Perchani, S., Portisch, J., Hertling, S., Paulheim, H.: Dbkwik: Towards knowledge graph creation from thousands of wikis. In: International Semantic Web Conference (Posters, Demos & Industry Tracks) (2017)
5. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal **6**(2) (2013)
6. Lenat, D.B.: CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM **38**(11), 33–38 (1995)
7. Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., Taylor, J.: Industry-scale knowledge graphs: Lessons and challenges. Communications of the ACM **62**(8), 36–43 (2019). https://doi.org/10.1145/3331166
8. Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web **8**(3), 489–508 (2017)
9. Paulheim, H., Ponzetto, S.P.: Extending dbpedia with wikipedia list pages. NLP-DBPEDIA@ ISWC **13** (2013)
10. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In: 16th international conference on World Wide Web. pp. 697–706 (2007)
11. Vrandečić, D., Krötzsch, M.: Wikidata: a Free Collaborative Knowledge Base. Communications of the ACM **57**(10), 78–85 (2014)