Enhancing Botanical Knowledge Graphs with Machine Learning

Qianqian Gu‡, Ben Scott‡, Vincent S. Smith‡

‡ Natural History Museum, London, United Kingdom

Corresponding author: Qianqian Gu (qianqian.gu@nhm.ac.uk)

Abstract

Integrating sparse and incomplete biodiversity data into a global, coherent data space and generating machine-readable data infrastructures is a challenge in biodiversity informatics. In recent years, biodiversity data researchers have started proposing Knowledge Graphs (KGs) as one approach to connecting biodiversity data worldwide (Page 2019), representing the connections between the what, when, and where of objects in natural history collections. At the Natural History Museum (NHM) we have constructed a KG of botanical specimens and collectors, encoded into numerical representations, and using a Relational Graph Convolutional Network (RGCN) (Schlichtkrull et al. 2018) to infer gaps in the KG, forging new connections between nodes. The datasets involved in our botanical KG project are NHM Botany Collector database (105,780 entities) and NHM Indian Region Botanical Specimen Dataset (110,043 entities with geographical information).

Our KG with RGCN enables the structured and contextual data to be reasoned across the knowledge content, allowing us to dynamically update its representation according to its closely related neighbours. Our work will explain why and how the KG with RGCN can offer a better way to link digitised botanical data. We use the prototype KG to demonstrate its potential for modelling botanical data and provide a graphical representation for other machine learning applications. For example, the combination of KG with RGCN and Metric Learning (Xing et al. 2002, a form of Machine Learning generally used to automatically construct task-specific distance metrics) supports data completion via entity classification and link prediction for a subset of botanical specimens within a geographic region. These data augmentation models with KGs allow us to identify gaps in specimen provenance, and fill in missing data. After phase one training, our model can achieve 88% accuracy in entity classification and report a reasonable Mean Reciprocal Rank (MRR) in raw ranking link prediction for the Indian Region Botanical Specimen Dataset.

Our research also evaluates the use of the KG and RGCN to improve post-OCR (Optical Character Recognition) correction algorithms as part of automatic specimen digitisation pipelines. This improves the accuracy of entity recognition on specimen label text

identification and transcription, as part of machine learning natural language processing and human-in-the-loop transcription. Human-based transcription can be aided and improved by an interpretation recommendation system predicated on the specimen unit's RGCN-inferred location in the KG. This methodology can also be used to explore the alignment of KGs from different institutions within the global biodiversity network, to identify the relative importance of collectors or determine strengths or gaps in different geographic regions or ecosystems, duplicate items in collections, or objects in collections that have potentially been misidentified.

Keywords

botanical data, linked data, semantic data, knowledge graph, machine learning

Presenting author

Qianqian Gu

Conflicts of interest

References

- Page RM (2019) Ozymandias: A biodiversity knowledge graph. PeerJ 7 (e6739). https://doi.org/10.7717/peerj.6739
- Schlichtkrull M, Kipf T, Bloem P, van den Berg R, Titov I, Welling M (2018) Modeling Relational Data with Graph Convolutional Networks. The Semantic Web 10843: 593-607. https://doi.org/10.1007/978-3-319-93417-4_38
- Xing E, Ng A, Jordan M, Russell S (2002) Distance Metric Learning with Application to Clustering with Side-Information. Advances in Neural Information Processing Systems 15 URL: https://proceedings.neurips.cc/paper/2002/file/c3e4035af2a1cde9f21e1ae1951ac80b-Paper.pdf