# iKNOW: A platform for knowledge graph construction for biodiversity

Samira Babalou[‡,§], David Schellenberger Costa[|], Helge Bruelheide[¶], Jens Kattge[#], Christine Römermann[‡], Christian Wirth[|], Birgitta König-Ries[‡,§]

‡ Friedrich Schiller University Jena, Jena, Germany
§ German Center for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Germany
| University of Leipzig, Leipzig, Germany
¶ Martin Luther University Halle-Wittenberg, Halle, Germany
# Max Planck Institute for Biogeochemistry, Jena, Germany

Corresponding author: Samira Babalou (samira.babalou@uni-jena.de)

## Abstract

Nowadays, more and more biodiversity datasets containing observational and experimental data are collected and produced by different projects. In order to answer the fundamental questions of biodiversity research, these data need to be integrated for joint analyses. However, to date, too often, these data remain isolated in silos.

Both in academia and industry, Knowledge Graphs (KGs) are widely regarded as a promising approach to overcome issues of data silos and lack of common understanding of data (Fensel and Şimşek 2020). KGs are graph-structured knowledge bases that store factual information in the form of structured relationships between entities, like “tree_species has_trait average_SLA” or “nutans is_observed_in SCH_Location" (Hogan et al. 2021). In our context, entities could be, e.g., abstract concepts like a kingdom, a species, or a trait, or a concrete specimen of a species. Example relationships could be "co-occurs" or, "possesses-trait". KGs for biodiversity have been proposed by Page 2019 and have also been the topic at prior TDWG conferences *[1] (Page 2021). However, to date, uptake of this concept in the community has been rather slow (Sachs et al. 2019).

We argue that this is at least partially due to the high effort and expertise required in developing and managing such KGs. Therefore, in our ongoing project, iKNOW (Babalou et al. 2021), we aim to provide a toolbox for reproducible KG creation. While iKNOW is still in an early stage, we aim to make this platform open-source and freely available to the biodiversity community. Thus, it can significantly contribute to making biodiversity data widely available, easily discoverable, and integratable.

For now, we focus on tabular datasets resulting from biodiversity observation or sampling events or experiments. Given such a dataset, iKNOW will support its transformation into (subject, predicate, object) triples in the RDF standard (Resource Description Framework). Every uploaded dataset will be considered as a subgraph of the main KG in iKNOW. If

required, data can be cleaned. After that, the entities and relationships among them should be extracted. For that, a user will be able select one of the existing semi-automatic tools available on our platform (e.g., JenTab (Abdelmageed and Schindler 2020)). The entities in this step can be linked to respective global identifiers in Wikidata, GBIF, the Global Biodiversity Information Facility, or any other user-selected knowledge resource. In the next step, (subject, predicate, object) triples based on the extracted information from the previous steps will be created. After these processes, the generated sub-KG can be used directly. However, one can take further steps such as: Triple Augmentation (generate new triples and extra relations to ease KG completion), Schema Refinement (refine the schema, e.g., via logical reasoning for the KG completion and correctness), Quality Checking (check the quality of the generated sub-KG), and Query Building (create customized SPARQL queries for the generated KG).

iKNOW will include a wide range of functionalities for creating, accessing, querying, visualizing, updating, reproducing, and tracking the provenance of KGs. The reproducibility of such a creation is essential to strengthening the establishment of open science practices in the biodiversity domain. Thus, all information regarding the user-selected tools with parameters and settings, along with the initial dataset and intermediate results, will be saved in every step of our platform. With the help of this, users can redo the previous steps. Moreover, this enables us to track the provenance of the created KG.

The iKNOW project is a joint effort by computer scientists and domain experts from the German Centre for Integrative Biodiversity Research (iDiv). As a showcase, we aim to create a KG of plant-related data sources at iDiv. These include, among others: TRY (the plant trait database) (Kattge and DÍaz 2011), sPlot (the database about global patterns of taxonomic, functional, and phylogenetic diversity) (Bruelheide and Dengler 2019), and PhenObs (the dataset of the global network of botanical gardens monitoring the impacts of climate change on the phenology of herbaceous plant species) (Nordt and Hensen 2021), LCVP, the Leipzig Catalogue of Vascular Plants, (Freiberg and Winter 2020), and many others.

The resulting KG will serve as a discovery tool for biodiversity data and provide a robust infrastructure for managing biodiversity knowledge. From the biodiversity research perspective, iKNOW will contribute to creating a dataset following the Linked Open Data principles by interlinking to cross-domain and specific-domain KGs. From the computer science perspective, iKNOW will contribute to developing tools for dynamic, low-effort creation of reproducible knowledge graphs.

## Keywords

biodiversity informatics, semantic web, knowledge graph platforms

## Presenting author

Samira Babalou

## Presented at

TDWG 2022

## Conflicts of interest

## References

- Abdelmageed N, Schindler S (2020) JenTab: Matching Tabular Data to Knowledge Graphs. In: CEUR (Ed.) In *SemTab@ ISWC*, pp. 40-49. 2020 http://ceur-ws.org/Vol-2775/paper4.pdf.
- Babalou S, Schellenberger Costa D, Kattge J, Römermann C, König-Ries B (2021) Towards a Semantic Toolbox for Reproducible Knowledge Graph Generation in the Biodiversity Domain-How to Make the Most out of Biodiversity Data. INFORMATIK 2021 https://doi.org/10.18420/informatik2021-044
- Bruelheide H, Dengler J, et al. (2019) sPlot – A new tool for global vegetation analyses. Journal of Vegetation Science 30 (2): 161-186. https://doi.org/10.1111/jvs.12710
- Fensel D, Şimşek U, et al. (2020) Introduction: What Is a Knowledge Graph? In: Knowledge Graphs. Springer, Cham. https://doi.org/10.1007/978-3-030-37439-6_1
- Freiberg M, Winter M, et al. (2020) LCVP, The Leipzig catalogue of vascular plants, a new taxonomic reference list for all known vascular plants. Sci Data 7, 416 (2020) https://doi.org/10.1038/s41597-020-00702-z
- Hogan A, Blomqvist E, Cochez M, d'Amato C (2021) Knowledge graphs. Synthesis Lectures on Data, Semantics, and Knowledge, Volume 22 of Synthesis Lectures on Data, Semantics, and Knowledge URL: https://kgbook.org/
- Kattge J, DÍaz S, et al. (2011) TRY - a global database of plant traits. Global Change Biology 17 (9): 2905-2935. https://doi.org/10.1111/j.1365-2486.2011.02451.x
- Nordt B, Hensen I, et al. (2021) The PhenObs initiative: A standardised protocol for monitoring phenological responses to climate change using herbaceous plant species in botanical gardens. Functional Ecology 35 (4): 821-834. https://doi.org/10.1111/1365-2435.13747
- Page RM (2019) Ozymandias: a biodiversity knowledge graph. PeerJ 7:e6739 https://doi.org/10.7717/peerj.6739
- Page RM (2021) Knowledge Graphs. Biodiversity Information Science and Standards 5: e73796 URL: https://biss.pensoft.net/article/73796/
- Sachs J, Page R, Baskauf SJ, Pender J, Lujan-Toro B, Macklin J, Comspon Z (2019) Training and hackathon on building biodiversity knowledge graphs. Research Ideas and Outcomes 5: e36152 https://doi.org/10.3897/rio.5.e36152

## Endnotes

*1 /https://biss.pensoft.net/collection/307/