ORKG

Open Research Knowledge Graph

Sören Auer / Vinodh Ilangovan / Markus Stocker / Sanju Tiwari / Lars Vogt (Eds.)

Cuvillier Verlag
Internationaler wissenschaftlicher Fachverlag

[May 2024, First Edition]

Open Research Knowledge Graph

Open Research Knowledge Graph

Editors

Sören Auer

Vinodh Ilangovan

Markus Stocker

Sanju Tiwari

Lars Vogt

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.
1. Aufl. - Göttingen : Cuvillier, 2024

Cover Design: Nadine Klöver

Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

1. Auflage, 2024

eISBN 978-3-68942-003-9

9. Knowledge synthesis in Invasion Biology: from a prototype to community-designed templates

Maud Bernard-Verdier[1,2], Kamel Fadel[3], Tina Heger[1,4], Jonathan M. Jeschke[1,2], Markus Stocker[3], Lars Vogt[3]

[1] *Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany*
[2] *Freie Universität Berlin, Institute of Biology, Berlin, Germany*
[3] *TIB Leibniz Information Centre for Science and Technology, Hannover, Germany*
[4] *Technical University of Munich, Restoration Ecology, Freising, Germany*

9.1 The prototype with Hi Knowledge data

Motivation

Biological invasions, i.e. the spread of organisms outside their native distributional range as a consequence of human activities, are one of the leading causes of global biodiversity decline. Invasion biology is a subfield of ecological research which has shown an exponential increase in publications in the past 25 years. The Hi Knowledge initiative[18], which was started around 2010 by Jonathan Jeschke and Tina Heger, aims to tackle this by synthesizing and visualizing knowledge in the field of invasion biology and beyond. In a collaborative book by Jeschke & Heger published in 2018, they reviewed the evidence for a set of 12 major hypotheses in invasion biology theory, which predict mechanisms favoring the introduction, spread and impact of species outside their native range. This resulted in a curated dataset assembling information from over 1000 articles testing at least one of these hypotheses.

The collaboration between Hi Knowledge and the ORKG started in Fall 2019. It was quickly clear that the Hi Knowledge dataset could demonstrate the capabilities of ORKG as a service. Ingesting community data into the ORKG, and using ORKG services such as Comparisons to demonstrate what is possible, was an invaluable activity, and with Hi Knowledge the first of this kind.

The SARS-CoV-2 pandemic had postponed more concrete activities towards these aims. However, they were resumed in 2021 in the context of a Master thesis

[18] https://hi-knowledge.org/

by Kamel Fadel (*Fadel, 2021*). In this work, we were able to ingest the Hi Knowledge data into ORKG, build an ORKG Observatory[19] for the community, create ORKG Comparisons[20] for the 10 individual Hi Knowledge hypotheses, and leverage the ORKG integrations with Jupyter to test whether computing environments / dashboards could support the production of tailored visualizations for the community. The Hi Knowledge network of hypotheses was a good objective for our ORKG prototype.

For this prototype with Hi Knowledge data, the research questions were thus of technical nature. Specifically, the work was motivated by the question whether Scientific Knowledge Graphs and ORKG in particular can be exploited in data science and with what technical approaches.

Approach and results

The activity consisted of the following key tasks: (1) Hi Knowledge data ingestion into the ORKG; (2) Create ORKG Comparisons; (3) Data science using the ingested data.

Hi Knowledge data ingestion. The starting point is data that was extracted from articles and published on the Hi Knowledge website[21] in separate files, one file per hypothesis. This data relates to 10 of the 12 hypotheses addressed in the 2018 book, as data on 2 hypotheses were structured in a different way. Both article metadata and extracted essential data as structured content were ingested for these 10 hypotheses, e.g.:

- Article's stance towards the hypothesis: Indicating whether it supports, is undecided, or questions the hypothesis
- The investigated taxa in the article, e.g., plants, birds, mammals, etc.
- Number of investigated taxa in the article
- The continent in which the study was conducted
- Used research method: Experimental or observational/correlational
- If the study was done in the lab, enclosures, or field

This data was first preprocessed to meet the syntax of ORKG CSV file import[22]. We created one CSV file per hypothesis, which thus amounted to a minor transformation of the original Hi Knowledge data to prepare the data for ingestion into ORKG.

19 https://orkg.org/observatory/Invasion_Biology?sort=combined&classesFilter=Paper,Comparison,Visualization

20 https://orkg.org/comparison/R58002/

21 https://hi-knowledge.org

22 https://orkg.org/help-center/article/16/Import_CSV_files_in_ORKG

ORKG Comparisons. Following ingestion, we created ORKG Comparisons, one for each hypothesis[23]. For this, we used the existing ORKG feature and its approach to create comparisons. Figure 9.1 exemplifies the Comparison for the enemy release hypothesis, also available online at https://orkg.org/comparison/R58002/.

Properties	The invertebrate fauna on broom, Cytisus scoparius, in two native and two exotic habitats *Contribution 2 - 2000*	A Comparison of Herbivore Damage on Three Invasive Plants and Their Native Congeners: Implications for the Enemy Release Hypothesis *Contribution 1 - 2000*	Can enemy release explain the invasion success of the diploid Leucanthemum vulgare in North America? *Contribution 1 - 2000*	Incorporation of an invasive plant into a native insect herbivore food web *Contribution 1 - 2000*
Continent	Europe Oceania	North-America	North-America	Europe
Habitat	Terrestrial	Terrestrial	Terrestrial	Terrestrial
has research problem	Testing the enemy release hypothesis in invasion biology	Testing the enemy release hypothesis in invasion biology	Testing the enemy release hypothesis in invasion biology	Testing the enemy release hypothesis in invasion biology
hypothesis	Enemy release	Enemy release	Enemy release	Enemy release
Indicator for enemy release	Infestation	Damage	Damage	Infestation
Investigated species	Plants	Plants	Plants	Plants
Number of species	1	3	1	1
Release of which kind of enemies?	Specialists	no differentiation	no differentiation	no differentiation

Figure 9.1 Comparison for Hi Knowledge data on the enemy release hypothesis.

Data science. An additional aim for this prototype with the Hi Knowledge community was to test if ORKG and its integrations with computing environments such as Jupyter could be used to perform specific analyses of the ingested data, including tailored visualizations that are meaningful for the community. We tested this by performing basic data science tasks with Jupyter Notebooks and web applications that use the ingested data and replicate the Hi Knowledge network of hypotheses. With the ORKG Python library[24], researchers can easily read the data constituting a comparison into a Python data frame and use the powerful scripting environment to implement and execute data science and analysis tasks. With such a setup, we can tackle simple and more advanced data science tasks. For instance, we can easily compute how many contributions support, are undecided, or question a specific hypothesis. Figure 9.2 visualizes the answer to this question for the propagule pressure hypothesis. Thanks to the flexibility of Python data frames, it is possible

[23] https://orkg.org/search/invasion?types=Comparison
[24] https://orkg.readthedocs.io

to slice and dice the data in an arbitrary manner. Figure 9.3 shows the distribution of Hi Knowledge studies across continents. While the approach requires some level of programming, it also shows how the versatility of a computing environment can support much more than predefined visualizations of data on a website. To address the requirement of programming skills, we also created an R Shiny application which, contrary to the Jupyter Notebooks, creates interactive dashboard-style web applications accessible to all users.

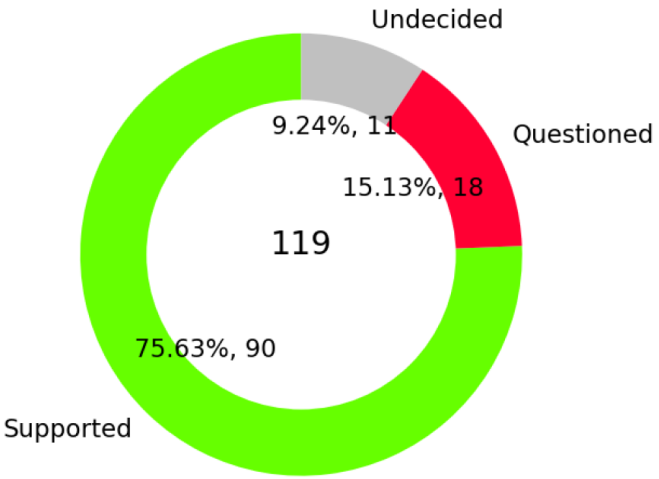

Figure 9.2 Share of contributions that support, question, or are undecided about the propagule pressure hypothesis.

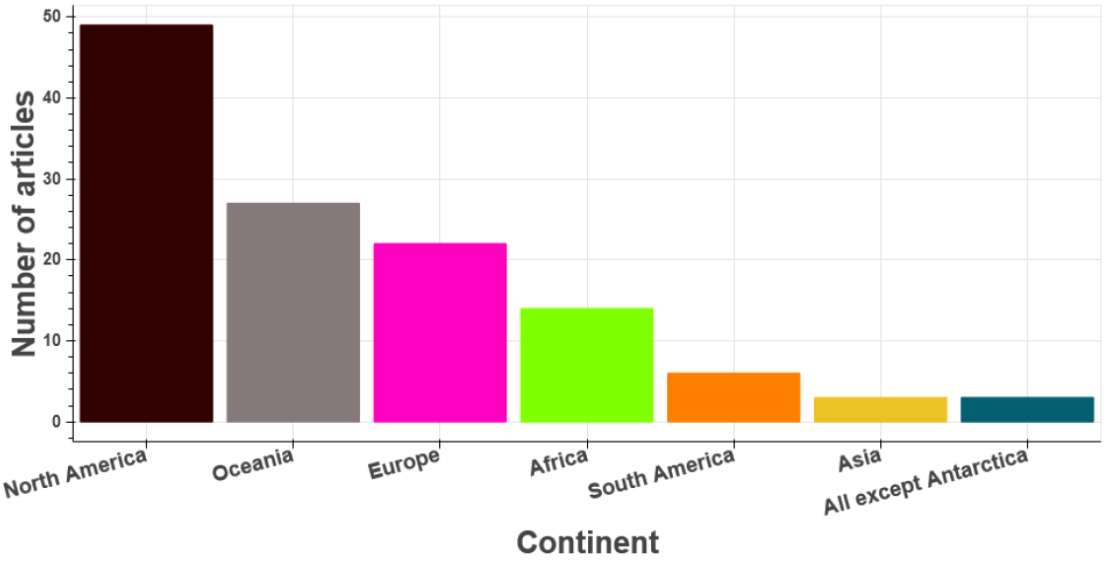

Figure 9.3 Visualization of the number of studies about the propagule pressure hypothesis across continents created with Hi Knowledge data ingested into ORKG using a computing environment.

9.2 The ecologist community gets more involved

Motivation

From 2021 to 2024, the enKORE project (*Jeschke et al., 2021*) within the Hi Knowledge initiative took further steps towards an atlas of knowledge for invasion biology. This project brought together ecologists and data scientists to work on organizing, extracting, synthesizing and visualizing literature in the field of invasion biology. The ORKG was used as a platform in this project to synthesize and visualize current scholarly literature on invasion biology. The effort was led by ecologist Maud Bernard-Verdier, in collaboration with Lars Vogt and Markus Stocker from the ORKG, with the goal first to revisit the existing data on 10 hypotheses in invasion biology.

Method

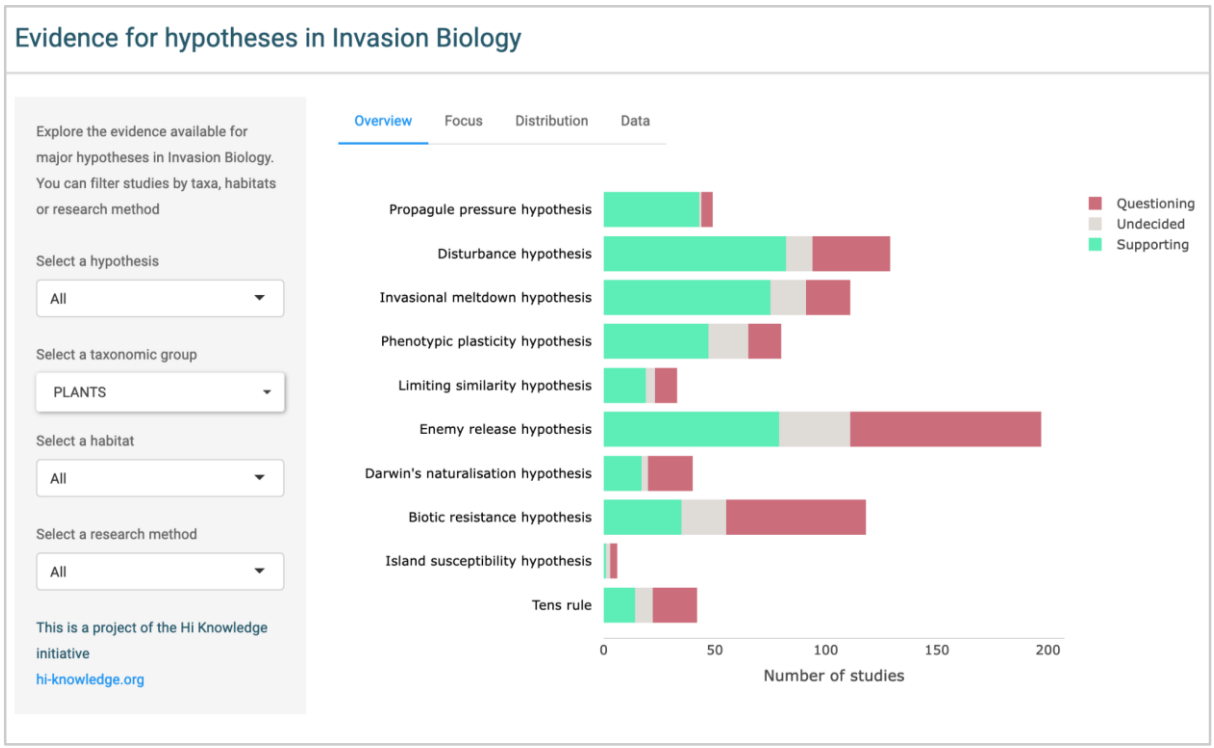

Figure 9.4 Screenshot of an R Shiny app[25] *offering an interactive visualization and summary of evidence for 10 hypotheses in invasion biology, combining 10 ORKG Comparison tables. Studies can be filtered by hypotheses, taxonomic groups, habitats or research methods. The Comparison tables (see Figure 9.1) were obtained by extracting existing published tables for synthetic reviews of hypotheses in invasion biology. The current view presents the distribution of evidence across 10 hypotheses for studies on invasive plants.*

As R is currently the preferred programming language for ecologists (*Lai et al., 2019*), the goal was to develop an R Shiny app for interactive visualization and exploration of the data, building upon the first Jupyter notebooks created by Kamel

[25] Visit the beta app: https://maudbernardverdier.shinyapps.io/Hypothesis-evidence-explorer/; R code accessible on github: https://github.com/maudbv/Hypothesis-evidence-explorer.

Fadel (see above). Using the ORKG package for Python (the R ORKG package was not yet finalized), Maud exported (as .csv) the 10 comparison tables summarizing support for the 10 hypotheses in invasion biology, and used them to create an R Shiny app, aiming first for a proof of concept on static data.

The app (Figure 9.4) presents a small number of curated figures and summary statistics relevant for ecologists to gain an overview of the state of knowledge concerning each hypothesis. Filtering options based on relevant properties annotated in ORKG Comparison tables allow for a customized exploration of the data, as well as data exports.

What we learned

Despite the careful data extraction by Kamel, substantial data cleaning and homogenization were necessary before the app could be created, mainly because the data tables from the original multi-author book (*Jeschke & Heger, 2018*) were themselves not perfectly standardized. For instance, the terms used to designate taxa groupings or habitats were not always comparable across hypothesis tables and had to be manually homogenized. This highlighted early on the need for better quality control (e.g. correcting typographic mistakes) and also standardized vocabulary, in which each term has a unique identifier, if we aim for seamless automatic synthesis. Guiding future ORKG annotations to re-use only pre-determined existing concepts in ORKG, published ontologies, or Wikidata, was identified as a solution to this problem in future steps.

Once data processing was completed, the task of creating visualizations benefited from the specialist perspective of the invasion biology community. While many figures and statistics were possible to compute, the visualizations included in the R Shiny app were selected to address basic questions in ecology concerning the current knowledge gaps and biases existing in the literature, and whether hypotheses are found to be better supported for some species or habitats. The app provides interactive versions of those static figures typically found in published systematic reviews, and one can imagine that systematic reviews could greatly benefit from being accompanied by such additional interactive material.

9.3 Engaging with the broader community of invasion biologists

Motivation

The Hi Knowledge dataset mentioned above is static and had not been updated since the publication of Jeschke & Heger, 2018. Such datasets are the product of an enormous synthesis effort by individual authors, which cannot be realistically reproduced on a regular basis. As mentioned above, the dataset was also not perfectly standardized and reusable, and, importantly, had not been fully semantically

modeled in ORKG (i.e. properties had no link to existing ontologies, Wikidata items or other semantic models).

We decided to use the ORKG as a platform to update the Hi Knowledge dataset, aiming for invasion biologists to contribute data following a comparable structure. The underlying idea is that invasion biologists who published a given study would be motivated to feed information about their study to ORKG, so that it is part of a growing database.

In the first attempts of invasion biologists in the team to add their own papers to ORKG, it quickly became clear that more guidance was needed. Invasion biologists do not typically know about semantic modeling or understand the rules, good practice and constraints associated with semantic annotations as is practiced in ORKG. If we want to motivate invasion biologists to spend time adding their work, and if we want the annotations to be comparable and valuable for automatic synthesis (e.g. in an R Shiny app), a tailored template is needed to guarantee interoperability across their contributions.

Method

Lars and Maud worked together on designing a tailored template for invasion biology that allows the annotation of basic ecological information about a study, as well as information about hypothesis testing following the Hi Knowledge dataset. This collaborative work relied on the input of invasion biologists, providing a list of example statements for Lars to build a first prototype of a semantic model. An online workshop in 2022 with over 70 invasion biologists[26] further identified a list of key concepts relevant to filter literature searches or organize meta-analyses. Building iteratively on this first graph, a first version of the template was implemented by Maud, and further tested and revisited following trial tests during a 2023 in-person workshop in Berlin[27].

We created several templates (Table 9.1): one main template for general scoping of any contribution in ecology and evolution, and five sub-templates, with three specific to invasion biology. It turned out that most of the key information we are interested in in invasion biology is common to the larger field of ecology, and we therefore seized on the opportunity to create a more general template for ecology (**#1**). After several iterations, we decided to simplify the initial template to make it more accessible, and move more complex information, such as descriptions of study design, datasets[28] or study systems, to sub-templates (**#4** and **#5**).

[26] Workshop report: https://zenodo.org/records/8421054

[27] Published workshop report: https://riojournal.com/article/115395/

[28] pre-existing ORKG template: https://orkg.org/template/R178304

Table 9.1: ORKG templates created for the field of invasion biology, and ecology in general.

#	Template name	Purpose	ORKG ID
1	Study in Ecology and Evolution (main template)	General template for any study in the field of ecology (*sensu largo)*	R593657
2	Invasion biology study research question	Annotate theme, research question, hypotheses and invasive taxa, following scheme by Musseau et al.	R593830
3	Hypothesis test in invasion biology	Annotate whether the study supports or not a major hypothesis	R646660
4	Ecological study system description	Describe the properties of a specific ecological study system, which can be shared by multiple studies	R593670
5	Ecological study design description	describe the study design (sample size, treatment, etc.) in an invasion biology study	R593806
6	Hypotheses in invasion biology template	Template for describing major theoretical hypotheses in invasion biology	R602693

Two sub-templates specific to the Hi Knowledge approach to invasion biology were designed. The first (#**2**) is a general description of the main theme, research questions, hypotheses and invasive taxa investigated, following our current conceptual scheme for invasion biology (*Musseau et al., in preparation*). The second (#**3**) describes the testing of major hypotheses in the field (described by template #**6**). It provides information about support or rebuttal of those hypotheses, in the same way as the Hi Knowledge data provided.

To create these templates, not only did new properties have to be modeled in ORKG, reusing as much as possible existing ontologies and Wikidata properties, but also new instance-resources to guide and limit the choices of template users. For instance, we wanted to allow the users to choose from a short list of research approaches, such as observational approaches, experimental approaches or conceptual approaches, and had to model those instances as well as the class to which they belong (class: “research approaches”[29]). We also created classes and instance-resources to describe all items of the conceptual scheme for invasion

[29] https://orkg.org/class/C65001

biology (5 themes, 10 research questions and 64 major hypotheses in invasion biology).

The templates then restricted the possible entries for these fields to only those belonging to the class. Of course, ORKG being fully flexible meant that users could still (and did!) create their own instances of research approach or hypotheses, which in most cases did not fit with what we had intended (e.g. too detailed, redundant with existing instance-resources, etc.). This great freedom in ORKG annotations is here a challenge for better standardization and automated knowledge synthesis.

9.4 Further use of ORKG in the context of invasion biology

ORKG for teaching in ecology

ORKG appeared as a great platform to teach students how to extract information from papers in a systematic way, and provide a published outcome for the class (published ORKG list[30]). In December 2023, we used the ORKG platform to teach (remotely) an introduction to invasion biology to a class of fourth year ecology students at Rhode Island University (USA) with Prof. Laura Meyerson, who had been part of previous workshops of the Hi Knowledge initiative. About 60 students were asked to annotate invasion biology papers using the ORKG templates described above, and with minimal guidance from us.

The pedagogical goals were the following:

1. Learn to extract key ecological information from a scientific paper in a systematic way.
2. Gain an overview of the different themes, research questions and hypotheses in invasion biology.
3. Contribute to community-curated tools for open knowledge synthesis in science.
4. Become familiar with notions of semantic graph modeling.

The students collectively annotated over 100 papers in two 3-hour sessions. The first session provided uneven results, and revealed a steep learning curve for the students to familiarize themselves with ORKG as a tool, as well as with the templates. At the end of the second session, though, most student groups had provided detailed annotations of two to five papers, spending roughly 30-60 mins per paper. This was highly encouraging regarding the usability of the templates, as well as a great learning experience for the students, who reported that they had felt “empowered” as students to actively participate in knowledge extraction rather

[30] https://orkg.org/list/R671240

than passive reading. This highlights the high pedagogical potential of such exercises with ORKG templates, and more ambitious versions of this class could even be designed as small systematic review projects.

Preliminary investigation of the data contributed by students nevertheless revealed a number of pitfalls in the template use, which need to be further analyzed. These might in part be avoided with clearer instructions (with a manual and demonstration) and better modeling. However, the inherent modeling freedom of ORKG means that we should always expect heterogeneities in data quality, and data cleaning strategies will need to be put in place for future data synthesis.

A tool for publishers to collect structured information about submissions

One clear challenge of our approach is to reach out and motivate a large portion of the community of invasion biologists to annotate papers, even their own work. One possibility to tackle this challenge could be to make such annotations part of the normal publication process in scientific journals. It is important, however, to design the process in a way that does not waste the time of authors in the publication submission process. In this perspective, semantic annotations could become a new standard for publishers at the submission level, replacing the current role of article keywords. Such annotations would make all new papers easier to search, group and filter by key ecological criteria. They would also allow dashboard-style automatic syntheses and overviews of the literature, representing the scope and possible research gaps on a given topic (similar to our R Shiny app for Hi Knowledge data), for publishers themselves, as well as any other users if the data is openly published and harvestable with each article.

Whether publishers would want to use ORKG as a platform is uncertain, but we could imagine that the platform could at least be used for preliminary tests and as a proof of concept. Partnerships with publishers willing to invest in open science and technology would be a great boost to the ORKG project. The modeling involved in designing custom templates for a given field should be published in itself as an open resource, and updated by the community around a consensus approach, to allow standardization and interoperability of annotations across journals and publishers and promote FAIR science.

Smart searches

Knowledge graphs allow us in theory to create smart searches with complex scoping and filtering based on statements or class hierarchies. Such smart searches are missing in ORKG, but many invasion biologists and other ecologist users would be interested in it. A good test case for that in ecology would be taxa (species) recognition which, due to the inherently hierarchical organization of taxonomies,

would lend itself particularly well to hierarchical grouping. Users would ideally like to be able to give the Latin name of a species, and it being recognized as a concept with all the known synonyms and taxonomic hierarchy, in such a way that studies could be grouped based on a higher taxonomic level (e.g. plants, insects, birds, etc.). Smart searches would then allow us to search for a certain taxonomic level, no matter the granularity, like “mammals” or “flowering plants”, and filter articles accordingly. While this is not yet possible in ORKG, it is something that would be a real asset to develop in the future.

9.5 Conclusion

Domain-specific templates are necessary for getting community engagement in ORKG, and partnership with scientists from different fields via collaborative projects like enKORE are a good way to build these resources. Outstanding issues are in the difficulty of scaling up engagement of the ecologist community, and data quality control. Data quality and interoperability within a field will depend on the quality of existing domain ontologies and other semantic models for a given field, which in the case of ecology still remain insufficiently developed. Potential solutions to be pursued include guiding “naive” users with better tutorials and explicit templates, engaging in teaching projects to curate certain topics, better workflows to connect with other open knowledge graph projects like Wikidata, and finally getting publishers involved.

References

Jeschke, J.M., & Heger, T. (2018). Invasion biology: hypotheses and evidence. CABI, Wallingford.

Fadel, K. (2021). Data Science with Scholarly Knowledge Graphs. Hannover : Gottfried Wilhelm Leibniz Universität Hannover. https://doi.org/10.15488/11535

Jeschke, J.M., Heger, T., Kraker, P., Schramm, M., Kittel, C., & Mietchen, D. (2021). Towards an open, zoomable atlas for invasion science and beyond. NeoBiota 68:5–18. https://doi.org/10.3897/neobiota.68.66685

Lai J, Lortie CJ, Muenchen RA, Yang J, Ma K (2019) Evaluating the popularity of R in ecology. Ecosphere 10: e02567. https://doi.org/10.1002/ecs2.2567

Jeschke, J.M., & Heger, T. (2018). Invasion biology: hypotheses and evidence. CABI, Wallingford.

Musseau, C., Bernard-Verdier, M., Heger, T., Skopeteas, L., Strasiewsky, D., Mietchen, D., & Jeschke, J. M. A conceptual classification scheme of invasion science. (in preparation)