Bootstrapping a Biodiversity Knowledge Graph

Roderic Page ‡

‡ University of Glasgow, Glasgow, United Kingdom

Corresponding author: Roderic Page (roderic.page@glasgow.ac.uk)

Abstract

The "biodiversity knowledge graph" is a nice metaphor for connecting biodiversity data sources, but can we actually build it? Do we have sufficient linked data available? Given that a knowledge graph is an aggregation of data from multiple sources, how do we give those sources credit for that data, and how do we handle changes to that data? Given that the classic interface to a knowledge graph is an intimidatingly empty SPARQL query box, how do we make the knowledge within a graph more accessible?

This talk discusses an attempt to build a knowledge graph with an eye on how to maintain the graph in the future. It adopts a model similar to Global Biodiversity Information Facility (GBIF) and CheckListBank where individual data providers make datasets available as independently citable units with Digital Object Identifiers (DOIs). Each dataset comprises linked data in the form of N-triples. To create a knowledge graph we simply download one or more such datasets and add them to a triple store. Each data source is assigned to its own named graph, such that we have provenance for each dataset, and we can update any dataset independently. Furthermore, anyone can build their own knowledge graph by mixing and matching the set of data (people, publications, taxa, etc.) most appropriate to their interests.

To bootstrap this approach, exemplar datasets are created based on data harvested from ORCID, Zenodo, and taxonomic name databases. Each demonstration dataset could be replaced in the future by data published directly by those providers. In some cases there are sufficient shared identifiers (such as DOIs and ORCIDs) to form a graph, but taxonomic data typically forms isolated islands. To help the knowledge graph coalesce we need "glue" in the form of datasets that link pairs of different identifiers, such as Life Science Identifiers (LSIDs) for names to DOIs for publications. With the addition of those cross links we can start to generate bibliographies for taxa, discover communities of taxonomic expertise, and more. This model of building a knowledge graph also opens opportunities for smaller, focussed datasets to be added to the graph using the same approach (as set of N-triples archived in an online repository).

In order to be useful, a knowledge graph needs to be easy to query and visualise. Simply providing a SPARQL endpoint is unlikely to be enough. As part of this project, I developed a GraphQL interface to the knowledge graph to provide a set of standard queries that can

support a simple web interface to the graph. This provides a way to explore the graph as it is being developed, which in turn can highlight gaps in connectivity and coverage that need to be addressed.

Keywords

linked data, persistent identifiers, LSIDs

Presenting author

Roderic Page

Presented at

TDWG 2022

Conflicts of interest