Assisting Biologists in Editing Taxonomic Information by Confronting Multiple Data Sources using Linked Data Standards

Franck Michel, Sandrine Tercerie, Antonia Ettorre, Olivier Gargominy, Catherine Faron Zucker

BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37421
doi: 10.3897/biss.3.37421

OPEN ACCESS

Conference Abstract

Assisting Biologists in Editing Taxonomic Information by Confronting Multiple Data Sources using Linked Data Standards

Franck Michel[‡], Catherine Faron-Zucker[‡], Sandrine Tercerie[§], Antonia Ettorre[‡], Gargominy Olivier[§]

‡ Université Côte d'Azur, CNRS, Inria, I3S, Sophia-Antipolis, France
§ Muséum national d'Histoire naturelle, Paris, France

Corresponding author: Franck Michel (franck.michel@cnrs.fr)

Received: 17 Jun 2019 | Published: 26 Jun 2019

Citation: Michel F, Faron-Zucker C, Tercerie S, Ettorre A, Olivier G (2019) Assisting Biologists in Editing Taxonomic Information by Confronting Multiple Data Sources using Linked Data Standards. Biodiversity Information Science and Standards 3: e37421. https://doi.org/10.3897/biss.3.37421

Abstract

During the last decade, Web APIs (Application Programming Interface) have gained significant traction to the extent that they have become a de-facto standard to enable HTTP-based, machine-processable data access. Despite this success, however, they still often fail in making data interoperable, insofar as they commonly rely on proprietary data models and vocabularies that lack formal semantic descriptions essential to ensure reliable data integration. In the biodiversity domain, multiple data aggregators, such as the Global Biodiversity Information Facility (GBIF) and the Encyclopedia of Life (EoL), maintain specialized Web APIs giving access to billions of records about taxonomies, occurrences, or life traits (Triebel et al. 2012). They publish data sets spanning complementary and often overlapping regions, epochs or domains, but may also report or rely on potentially conflicting perspectives, e.g. with respect to the circumscription of taxonomic concepts. It is therefore of utmost importance for biologists and collection curators to be able to confront the knowledge they have about taxa with related data coming from third-party data sources.

To tackle this issue, the French National Museum of Natural History (MNHN) has developed an application to edit TAXREF, the French taxonomic register for fauna, flora

and fungus (Gargominy et al. 2018). TAXREF registers all species recorded in metropolitan France and overseas territories, accounting for 260,000+ biological taxa (200,000+ species) along with 570,000+ scientific names. The TAXREF-Web application compares data available in TAXREF with corresponding data from third-party data sources, points out disagreements and allows biologists to add, remove or amend TAXREF accordingly. This requires that TAXREF-Web developers write a specific piece of code for each considered Web API to align TAXREF representation with the Web API counterpart. This task is time-consuming and makes maintenance of the web application cumbersome.

In this presentation, we report on a new implementation of TAXREF-Web that harnesses the Linked Data standards: Resource Description Framework (RDF), the Semantic Web format to represent knowledge graphs, and SPARQL, the W3C standard to query RDF graphs. In addition, we leverage the *SPARQL Micro-Service* architecture (Michel et al. 2018), a lightweight approach to query Web APIs using SPARQL. A SPARQL micro-service is a SPARQL endpoint that wraps a Web API service; it typically produces a small, resource-centric RDF graph by invoking the Web API and transforming the response into RDF triples.

We developed *SPARQL* micro-services to wrap the Web APIs of GBIF, World Register of Marine Species (WoRMS), FishBase, Index Fungorum, Pan-European Species directories Infrastructure (PESI), ZooBank, International Plant Names Index (IPNI), EoL, Tropicos and Sandre. These micro-services consistently translate Web APIs responses into RDF graphs utilizing mainly two well-adopted vocabularies: Schema.org (Guha et al. 2015) and Darwin Core (Baskauf et al. 2015). This approach brings about two major advantages. First, the large adoption of Schema.org and Darwin Core ensures that the services can be immediately understood and reused by a large audience within the biodiversity community. Second, wrapping all these Web APIs in SPARQL micro-services “suddenly” makes them technically and semantically interoperable, since they all represent resources (taxa, habitats, traits, etc.) in a common manner. Consequently, the integration task is simplified: confronting data from multiple sources essentially consists of writing the appropriate SPARQL queries, thus making easier web application development and maintenance. We present several concrete cases in which we use this approach to detect disagreements between TAXREF and the aforementioned data sources, with respect to taxonomic information (author, synonymy, vernacular names, classification, taxonomic rank), habitats, bibliographic references, species interactions and life traits.

Keywords

Web API, data integration, Linked Data, SPARQL

Presenting author

Franck Michel

Presented at

Biodiversity_Next 2019

References

- Baskauf S, Wieczorek J, Deck J, Webb C, Morris PJ, Schildhauer M (2015) Darwin Core RDF Guide. Biodiversity Information Standards (TDWG). http://rs.tdwg.org/dwc/terms/guides/rdf/
- Gargominy O, Tercerie S, Régnier C, Ramage T, Dupont P, Daszkiewicz P, Poncet L (2018) TAXREF v12, référentiel taxonomique pour la France : méthodologie, mise en oeuvre et diffusion. Muséum national d'Histoire naturelle. Rapport Patrinat 2018-117.
- Guha RV, Brickley D, MacBeth S (2015) Schema.org: Evolution of Structured Data on the Web. ACM Queue - Structured Data 13 (9): 1.
- Michel F, Faron-Zucker C, Gargominy O, Gandon F (2018) Integration of Web APIs and Linked Data Using SPARQL Micro-Services — Application to Biodiversity Use Cases. Information 9 (12): 310. https://doi.org/10.3390/info9120310
- Triebel D, Hagedorn G, Rambold G (2012) An appraisal of megascience platforms for biodiversity information. MycoKeys 5: 45-63. https://doi.org/10.3897/mycokeys.5.4302