Ontological modeling of dynamic biodiversity consensus

Abstract

The digitization of biodiversity in extreme environments, such as the Rub' al Khali (Empty Quarter), relies increasingly on citizen science platforms like iNaturalist. However, the data produced is not static; taxonomic identifications evolve through community consensus, creating a provenance challenge for the Semantic Web. We present a generalized, configurable ETL pipeline and formal OWL 2 DL ontology to model this dynamic ecosystem. Using the Rub' al Khali project as a primary case study, we demonstrate a system that integrates iNaturalist data with the NCBI Taxonomy to detect epistemic conflicts between agents. We separate the TBox (consensus logic) from the ABox (observation data), enabling automated reasoning over conflicting evidence and providing a rigorous structure for integrating shifting classifications into the Linked Open Data cloud.

Keywords

OWL-DL, SIO, Biodiversity, Rub' al Khali, Conflict Detection

1. Introduction

The Rub' al Khali, the world's largest sand desert, represents a significant data void in global biodiversity monitoring. To address this, we established a digitization project on iNaturalist seeded by research expeditions. While effective for data mobilization, the platform's consensus mechanism, where an observation's identity "flips" based on user voting, presents a semantic challenge. Existing Darwin Core mappings [3] capture only the snapshot of the current state, losing the history of disagreement essential for scientific rigor. We propose a solution using OWL 2 DL aligned with the Semanticscience Integrated Ontology (SIO) [1].

By strictly separating the ontological schema (TBox) from the instance data (ABox), we enable automated reasoning to detect logical inconsistencies in taxonomic assertions. While developed for the Rub' al Khali, the architecture is generalized to support any iNaturalist project.

2. Ontology Design

To model scientific disagreement while respecting biological hierarchy, we treat taxonomy as a system of class subsumptions and disjointness. We separate epistemic assertions (the act of claiming an identification) from the underlying reality (the biological individual).

SWAT4HCLS 2026: Semantic Web Applications and Tools for Health Care and Life Sciences, February 16–19, 2026, Leiden, The Netherlands

robert.hoehndorf@kaust.edu.sa (R. Hoehndorf)
https://bio-ontology.org (R. Hoehndorf)

CEUR Workshop Proceedings (CEUR-WS.org)

2.1. Generalized Architecture

We implemented a configurable Groovy-based pipeline that fetches data from the iNaturalist API and transforms it into an OWL ontology. The system utilizes a generic namespace (`http://example.org/biodiversity/`) to support multiple projects. The resulting knowledge graph is hosted in a Virtuoso SPARQL endpoint and visualized via a lightweight web dashboard.

2.2. Hybrid Taxonomy and Deep Disjointness

A key challenge is the lack of formal logical definitions in citizen science data. We address this by constructing a hybrid taxonomic tree:

1. **iNaturalist Backbone:** We import the complete taxonomic hierarchy from the iNaturalist observations, representing taxa as OWL classes.
2. **NCBI Integration:** Where possible, iNaturalist taxa are mapped to the NCBI Taxonomy OBO ontology. This grounds the layman terms in a curated scientific standard.
3. **Deep Disjointness:** To enable conflict detection across the entire tree, we algorithmically generate `isIncompatibleWith` property assertions for all pairs of taxa that diverge in the hierarchy. This ensures that a disagreement between a Genus and a distinct Family is flagged just as robustly as a sibling Species conflict.

2.3. The Identification Process

We model an identification as a reified `sio:process`. An `IdentificationAct`:

1. Is performed by an Agent (`sio:has-agent`).
2. Occurs at a specific time (`sio:has-value`).
3. Targets a specific Observation (`sio:has-target`).
4. Outputs a Taxon determination (`sio:has-output`).

$$\begin{aligned}\texttt{IdentificationAct} \sqsubseteq\ & \texttt{sio:process} \\ & \sqcap \exists \texttt{sio:has-agent.sio:Agent} \\ & \sqcap \exists \texttt{sio:has-target.Observation} \\ & \sqcap \exists \texttt{sio:has-output.Taxon}\end{aligned} \tag{1}$$

2.4. Epistemic Conflict Detection

We define conflict at the assertion level using SWRL rules. If an observation is the target of two *active* identification acts (i.e., not superseded by a later revision) that output incompatible taxa, the observation is classified as a `ConflictingObservation`.

$$\begin{aligned}\texttt{ActiveID}(?a1) \wedge \texttt{ActiveID}(?a2) \wedge \texttt{target}(?a1, ?o) \wedge \texttt{target}(?a2, ?o) \\ \wedge \texttt{output}(?a1, ?t1) \wedge \texttt{output}(?a2, ?t2) \\ \wedge \texttt{isIncompatibleWith}(?t1, ?t2) \\ \rightarrow \texttt{ConflictingObservation}(?o)\end{aligned} \tag{2}$$

3. Results: Instantiation (ABox)

The system was instantiated with data from the Rub' al Khali project (https://www.inaturalist.org/projects/rub-al-khali).

3.1. Data Provenance

As of February 2026, the dataset comprises 227 observations and 546 identification acts. The hybrid taxonomy contains 269 unique taxa, of which 80 were successfully mapped to NCBI IDs, while 14 remained specific to iNaturalist. The remaining taxa serve as intermediate nodes in the hierarchy.

3.2. Conflict Analysis

The HermiT reasoner successfully classified the ontology, detecting **10 conflicting observations**. These conflicts represent genuine scientific disagreement or uncertainty that persists despite community voting. For example, Observation `obs_321457657` was flagged due to active disagreement between two experts on the specific species classification.

The "Deep Disjointness" algorithm generated over 33,000 incompatibility assertions, ensuring that the reasoner could detect conflicts at any level of the taxonomic tree, not just between direct siblings.

3.3. Deployment

The final knowledge graph is deployed in a Dockerized environment using Virtuoso. A custom web interface allows researchers to view live statistics, visualize conflicting observations (with photos fetched dynamically from iNaturalist), and execute arbitrary SPARQL queries against the consensus model.

4. Discussion and Conclusion

By distinguishing between the TBox (the logic of consensus) and the ABox (the expedition data), and by employing a hybrid taxonomy, we gain several advantages:

1. **Monotonicity:** New identifications are added without deleting prior assertions, preserving the history of scientific opinion.
2. **Robustness:** The hybrid approach leverages the best of both worlds: the coverage of iNaturalist and the rigorous semantics of NCBI.

3. **Querying Disagreement:** The system allows targeted retrieval of contentious observations, prioritizing them for DNA barcoding or expert review.

This approach ensures that the digital representation of the Rub' al Khali's flora is not just a static archive, but a living knowledge graph that accurately reflects the scientific process of identification.

References

[1] Dumontier, M., et al.: The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. Journal of Biomedical Semantics 5, 14 (2014).

[2] iNaturalist: A Community for Naturalists. https://www.inaturalist.org.

[3] Wieczorek, J., et al.: Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1), e29715 (2012).

[4] Hitzler, P., et al.: OWL 2 Web Ontology Language Primer. W3C Recommendation (2009).