Conference Abstract

A High-throughput Data Ingest Pipeline for Semantic Data-stores

John Deck[‡], Brian Stucky[§], Ramona Walls[|], Rodney Ewing[¶], Melissa Genazzio[#], Henry W Loescher[#], Robert Guralnick[¤]

‡ University of California at Berkeley, Berkeley, United States of America
§ Florida Museum of Natural History, University of Florida, Gainesville, United States of America
| CyVerse, Tucson, United States of America
¶ Biocode, LLC, Junction City, United States of America
# National Ecological Observatory Network, Boulder, United States of America
¤ Vertnet, Florida, United States of America

Abstract

Ontologies offer multiple benefits for biodiversity data processing and analysis, including precisely defined vocabularies, robust pathways for data integration, and support for automated machine reasoning. However, ontologies have yet to be widely deployed for biodiversity data processing and analysis. Reasons for this include: specialized skills and coordination are needed for mapping terms to source data, data processing and machine reasoning are computationally expensive, and there is a scarcity of tools for working with ontologies and RDF triples. In this presentation we will discuss a data processing pipeline (available at https://github.com/biocodellc/ppo-data-pipeline) which simplifies complex implementation tasks, offers tools for data ingest, triplifying, and reasoning, and makes datasets available for indexing.

Keywords

Ontology, Pipeline, Workflow, Data Integration

Presenting author

John Deck

Presented at

TDWG 2017