PROCEEDINGS OF TDWG

Biodiversity Information Science and Standards 1: e20067
doi: 10.3897/tdwgproceedings.1.20067

OPEN ACCESS

Conference Abstract

Argo as a platform for integrating distinct biodiversity analytics tools into workflows for building graph databases

Riza Batista-Navarro[‡], Nhung T. H. Nguyen[‡], Axel J. Soto[‡], William Ulate[§], Sophia Ananiadou[‡]

‡ University of Manchester, Manchester, United Kingdom
§ Missouri Botanical Garden, St. Louis, MO, United States of America

Corresponding author: Riza Batista-Navarro (riza.batista@manchester.ac.uk)

Received: 06 Aug 2017 | Published: 07 Aug 2017

Citation: Batista-Navarro R, Nguyen N, Soto A, Ulate W, Ananiadou S (2017) Argo as a platform for integrating distinct biodiversity analytics tools into workflows for building graph databases. Proceedings of TDWG 1: e20067. https://doi.org/10.3897/tdwgproceedings.1.20067

Abstract

Together with the increasingly growing amount of available data on biodiversity comes the proliferation of various informatics tools aimed at the collection, management and analysis of biodiversity-relevant knowledge. Consequently, we have seen how several data formats and programming languages or environments have come into use, giving rise to a problem in interoperability should anyone wish to combine the outputs of distinct tools, or to integrate them into one solution.

Argo (Rak et al. 2012), an online text mining workbench based on the Unstructured Information Management Architecture (UIMA) interoperability standard, offers a means for seamlessly unifying various tools and resources into customisable text processing workflows. Among many other features, Argo provides: (1) a library of diverse tools, i.e., UIMA components, each of which is dedicated to a specific task such as loading datasets or gazetteers of interest (e.g., the Biodiversity Term Inventory), recognition of species names and their semantically related terms (Nguyen et al. 2017); (2) a graphical interface for designing workflows using components as building blocks; (3) an environment for executing and monitoring the progress of workflows; and (4) a user-interactive annotation editor for manually revising or validating results of automated processing.

Recently, Argo has been extended to provide support for incorporating into workflows external web services conforming with the Representational State Transfer (REST) protocol. Taking advantage of these features, we demonstrate how we combine in-house tools and resources for named entity recognition (Batista-Navarro et al. 2017) with externally developed ones, e.g., EXTRACT (Pafilis et al. 2016), in order to build text mining workflows for populating neo4j graph databases with biodiversity-relevant knowledge. To provide a few exemplars, we focus on use cases that seek to leverage various sources of literature to capture fine-grained information on the habitat and reproductive conditions of: (1) a subset of plants catalogued in World Flora Online (Jackson and Miller 2015), and (2) tropical trees belonging to the *Dipterocarpaceae* family.

Keywords

text mining, information extraction, graph databases, workflows, knowledge curation

Presenting author

Riza Batista-Navarro

Funding program

Newton Fund Institutional Links

Grant title

Conserving Philippine Biodiversity by Understanding Big Data (COPIOUS): Integration and analysis of heterogeneous information on Philippine biodiversity

References

- Batista-Navarro R, Zerva C, Nguyen NH, Ananiadou S (2017) A text mining-based framework for constructing an RDF-compliant biodiversity knowledge repository. Communications in Computer and Information Science. 656. Springer, Cham, 30-42 pp. https://doi.org/10.1007/978-3-319-55209-5_3
- Jackson PW, Miller J (2015) Developing a World Flora Online - a 2020 challenge to the world's botanists from the international community. Rodriguésia 66 (4): 939-946. https://doi.org/10.1590/2175-7860201566402

- Nguyen NH, Soto A, Kontonatsios G, Batista-Navarro R, Ananiadou S (2017) Constructing a biodiversity terminological inventory. PLOS ONE 12 (4): e0175277. https://doi.org/10.1371/journal.pone.0175277
- Pafilis E, Buttigieg PL, Ferrell B, Pereira E, Schnetzer J, Arvanitidis C, Jensen LJ (2016) EXTRACT: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation. Database 2016: baw005. https://doi.org/10.1093/database/baw005
- Rak R, Rowley A, Black W, Ananiadou S (2012) Argo: an integrative, interactive, text mining-based workbench supporting curation. Database 2012: bas010. https://doi.org/10.1093/database/bas010