Author-Driven Computable Data and Ontology Production for Taxonomists

Hong Cui[‡], Bruce Ford[§], Julian Starr[|], James Macklin[¶], Anton Reznicek[#], Noah W. Giebink[‡], Dylan Longert[|], Étienne Léveillé-Bourret[|], Limin Zhang[‡]

‡ University of Arizona, Tucson, United States of America
§ University of Manitoba, Winnipeg, Canada
| University of Ottawa, Ottawa, Canada
¶ Agriculture and Agri-Food Canada, Ottawa, Canada
# University of Michigan, Ann Arbor, United States of America

Corresponding author: Hong Cui (hongcui@email.arizona.edu)

Abstract

It takes great effort to manually or semi-automatically convert free-text phenotype narratives (e.g., morphological descriptions in taxonomic works) to a computable format before they can be used in large-scale analyses. We argue that neither a manual curation approach nor an information extraction approach based on machine learning is a sustainable solution to produce computable phenotypic data that are FAIR (Findable, Accessible, Interoperable, Reusable) (Wilkinson et al. 2016). This is because these approaches do not scale to all biodiversity, and they do not stop the publication of free-text phenotypes that would need post-publication curation. In addition, both manual and machine learning approaches face great challenges: the problem of inter-curator variation (curators interpret/convert a phenotype differently from each other) in manual curation, and keywords to ontology concept translation in automated information extraction, make it difficult for either approach to produce data that are truly FAIR. Our empirical studies show that inter-curator variation in translating phenotype characters to Entity-Quality statements (Mabee et al. 2007) is as high as 40% even within a single project. With this level of variation, curated data integrated from multiple curation projects may still not be FAIR.

The key causes of this variation have been identified as semantic vagueness in original phenotype descriptions and difficulties in using standardized vocabularies (ontologies). We argue that the authors describing characters are the key to the solution. Given the right tools and appropriate attribution, the authors should be in charge of developing a project's semantics and ontology. This will speed up ontology development and improve the semantic clarity of the descriptions from the moment of publication. In this presentation, we will introduce the Platform for Author-Driven Computable Data and Ontology Production for Taxonomists, which consists of three components:

1. a web-based, ontology-aware software application called 'Character Recorder,' which features a spreadsheet as the data entry platform and provides authors with

the flexibility of using their preferred terminology in recording characters for a set of specimens (this application also facilitates semantic clarity and consistency across species descriptions);

2. a set of services that produce RDF graph data, collects terms added by authors, detects potential conflicts between terms, dispatches conflicts to the third component and updates the ontology with resolutions; and
3. an Android mobile application, 'Conflict Resolver,' which displays ontological conflicts and accepts solutions proposed by multiple experts.

Fig. 1 shows the system diagram of the platform.

The presentation will consist of:

1. a report on the findings from a recent survey of 90+ participants on the need for a tool like Character Recorder;
2. a methods section that describes how we provide semantics to an existing vocabulary of quantitative characters through a set of properties that explain where and how a measurement (e.g., length of perigynium beak) is taken. We also report on how a custom color palette of RGB values obtained from real specimens or high-quality specimen images, can be used to help authors choose standardized color descriptions for plant specimens; and
3. a software demonstration, where we show how Character Recorder and Conflict Resolver can work together to construct both human-readable descriptions and RDF graphs using morphological data derived from species in the plant genus *Carex* (sedges). The key difference of this system from other ontology-aware systems is that authors can directly add needed terms to the ontology as they wish and can update their data according to ontology updates.

The software modules currently incorporated in Character Recorder and Conflict Resolver have undergone formal usability studies. We are actively recruiting *Carex* experts to participate in a 3-day usability study of the entire system of the Platform for Author-Driven Computable Data and Ontology Production for Taxonomists. Participants will use the platform to record 100 characters about one *Carex* species. In addition to usability data, we will collect the terms that participants submit to the underlying ontology and the data related to conflict resolution. Such data allow us to examine the types and the quantities of logical conflicts that may result from the terms added by the users and to use Discrete Event Simulation models to understand if and how term additions and conflict resolutions converge.

We look forward to a discussion on how the tools (Character Recorder is online at http://shark.sbs.arizona.edu/chrecorder/public) described in our presentation can contribute to producing and publishing FAIR data in taxonomic studies.

Keywords

community ontology building, FAIR, taxonomic descriptions, survey, software, ontology-aware data editor, conflicts in ontology, mobile application, Character Recorder, Conflict Resolver

Presenting author

Hong Cui

Presented at

TDWG 2021

Funding program

Advances in Biological Informatics of U.S. National Science Foundation (NSF)

Conflicts of interest

References

- Mabee P, Ashburner M, Cronk Q, Gkoutos G, Haendel M, Segerdell E, Mungall C, Westerfield M (2007) Phenotype ontologies: the bridge between genomics and evolution. Trends in Ecology & Evolution 22 (7): 345-350. https://doi.org/10.1016/j.tree.2007.03.013
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, ’t Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18

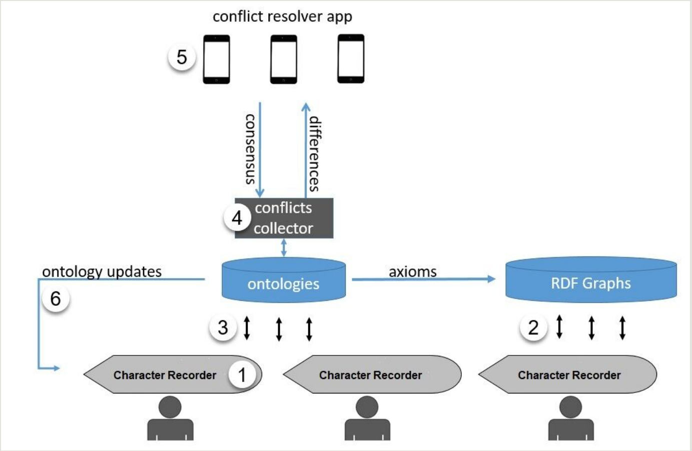

Figure 1.

System diagram for the author-driven computable data and ontology development platform.