Technical Considerations for a Transactional Model to Realize the Digital Extended Specimen

Nelson Rios[‡], Sharif Islam[§,|], James Macklin[¶], Andrew Bentley[#]

‡ Yale University, New Haven, United States of America
§ Naturalis Biodiversity Center, Leiden, Netherlands
| DiSSCo, Leiden, Netherlands
¶ Agriculture and Agri-Food Canada, Ottawa, Canada
# University of Kansas, Lawrence, KS, United States of America

Corresponding author: Nelson Rios (nelson.rios@yale.edu)

Abstract

Technological innovations over the past two decades have given rise to the online availability of more than 150 million specimen and species-lot records from biological collections around the world through large-scale biodiversity data-aggregator networks. In the present landscape of biodiversity informatics, collections data are captured and managed locally in a wide variety of databases and collection management systems and then shared online as point-in-time Darwin Core archive snapshots. Data providers may publish periodic revisions to these data files, which are retrieved, processed and re-indexed by data aggregators. This workflow has resulted in data latencies and lags of months to years for some data providers. The Darwin Core Standard Wieczorek et al. (2012) provides guidelines for representing biodiversity information digitally, yet varying institutional practices and lack of interoperability between Collection Management Systems continue to limit semantic uniformity, particularly with regard to the actual content of data within each field. Although some initiatives have begun to link data elements, our ability to comprehensively link all of the extended data associated with a specimen, or related specimens, is still limited due to the low uptake and usage of persistent identifiers. The concept now under consideration is to create a Digital Extended Specimen (DES) that adheres to the tenets of Findable, Accessible, Interoperable and Reusable (FAIR) data management of stewardship principles and is the cumulative digital representation of all data, derivatives and products associated with a physical specimen, which are individually distinguished and linked by persistent identifiers on the Internet to create a web of knowledge.

Biodiversity data aggregators that mobilize data across multiple institutions routinely perform data transformations in an attempt to provide a clean and consistent interpretation of the data. These aggregators are typically unable to interact directly with institutional data repositories, thereby limiting potentially fruitful opportunities for annotation, versioning, and repatriation. The ability to track such data transactions and satisfy the accompanying legal

implications (e.g. Nagoya Protocol) is becoming a necessary component of data publication which existing standards do not adequately address. Furthermore, no mechanisms exist to assess the “trustworthiness” of data, critical to scientific integrity, reproducibility or to provide attribution metrics for collections to advocate for their contribution or effectiveness in supporting such research.

Since the introduction of Darwin Core Archives Wieczorek et al. (2012) little has changed in the underlying mechanisms for publishing natural science collections data and we are now at a point where new innovations are required to meet current demand for continued digitization, access, research and management. One solution may involve changing the biodiversity data publication paradigm to one based on the atomized transactions relevant to each individual data record. These transactions, when summed over time, allows us us to realize the most recently accepted revision as well as historical and alternative perspectives. In order to realize the Digital Extended Specimen ideals and the linking of data elements, this transactional model combined with open and FAIR data protocols, application programming interfaces (APIs), repositories, and workflow engines can provide the building blocks for the next generation of natural science collections and biodiversity data infrastructures and services. These and other related topics have been the focus of phase 2 of the global consultation on converging Digital Specimens and Extended Specimens. Based on these discussions, this presentation will explore a conceptual solution leveraging elements from distributed version control, cryptographic ledgers and shared redundant storage to overcome many of the shortcomings of contemporary approaches.

Keywords

transactional publishing, provenance, Digital Extended Specimen

Presenting author

Nelson Rios

Presented at

TDWG 2021

Conflicts of interest

References

- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1). https://doi.org/10.1371/journal.pone.0029715