I Know Something You Don’t Know: The annotation saga continues…

James A Macklin[‡], David Peter Shorthouse[‡], Falko Glöckler[§]

‡ Agriculture and Agri-Food Canada, Ottawa, Canada
§ Museum für Naturkunde Berlin, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany

Corresponding author: James A Macklin (james.macklin@canada.ca)

Abstract

Over the past 20 years, the biodiversity informatics community has pursued components of the digital annotation landscape with varying degrees of success. We will provide an historical overview of the theory, the advancements made through a few key projects, and will identify some of the ongoing challenges and opportunities. The fundamental principles remain unchanged since annotations were first proposed. Someone (or something): (1) has an enhancement to make elsewhere from the source where original data or information are generated or transcribed; (2) wishes to broadcast these statements to the originator and to others who may benefit; and (3) expects persistence, discoverability, and attribution for their contributions alongside the source.

The Filtered Push project (Morris et al. 2013) considered several use cases and pioneered development of services based on the technology of the day. The exchange of data between parties in a universally consistent way necessitated the development of a novel draft standard for data annotations via an extension of the World Wide Web Consortium’s Web Annotation Working Group standard (Sanderson et al. 2013) to be sufficiently informative for a data curator to confidently make a decision. Figure 2 from Morris et al. (2013), reproduced here as Fig. 1, outlines the composition of an annotation data package for a taxonomic identification. The package contains the data object(s) associated with an occurrence, an expression of the motivation(s) for updating, some evidence for an assertion, and a stated expectation for how the receiving entity should take action. The Filtered Push and Annosys (Tschöpe et al. 2013) projects also considered implementation strategies involving collection management systems (e.g., Symbiota) and portals (e.g., European Distributed Institute of Taxonomy, EDIT). However, there remain technological barriers for these systems to operate at scale, the least of which is the absence of globally unique, persistent, resolvable identifiers for shared objects and concepts.

Major aggregation infrastructures like the Global Biodiversity Information Facility (GBIF) and the Distributed System of Scientific Collections (DiSSCo) rely on data enhancement to improve the quality of their resources and have annotation services in their work plans.

More recently, the Digital Extended Specimen (DES) concept (Hardisty et al. 2022) will rely on annotation services as key components of the proposed infrastructure. Recent work on annotation services more generally has considered various new forms of packaging and delivery such as Frictionless Data (Fowler et al. 2018), Journal Article Tag Suite XML (Agosti et al. 2022), or nanopublications (Kuhn et al. 2018). There is risk in fragmentation of this landscape and disenfranchisement of both biological collections and the wider research community if we fail to align the purpose, content, and structure of these packages or if these fail to remain aligned with FAIR principles.

Institutional collection management systems currently represent the canonical data store that provides data to researchers and data aggregators. It is critical that information and/or feedback about the data they release be round-tripped back to them for consideration. However, the sheer volume of annotations that could be generated by both human and machine curation processes will overwhelm local data curators and the systems supporting them. One solution to this is to create a central annotation store with write and discovery services that best support the needs of all stewards of data. This will require an international consortium of parties with a governance and technical model to assure its sustainability.

Keywords

collections, biodiversity, round-tripping

Presenting author

James Macklin

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Agosti D, Benichou L, Addink W, Arvanitidis C, Catapano T, Cochrane G, Dillen M, Döring M, Georgiev T, Gérard I, Groom Q, Kishor P, Kroh A, Kvaček J, Mergen P, Mietchen D, Pauperio J, Sautter G, Penev L (2022) Recommendations for use of annotations and persistent identifiers in taxonomy and biodiversity publishing. Research Ideas and Outcomes 8 https://doi.org/10.3897/rio.8.e97374

- Fowler D, Barratt J, Walsh P (2018) Frictionless Data: Making Research Data Quality Visible. International Journal of Digital Curation 12 (2): 274-285. https://doi.org/10.2218/ijdc.v12i2.577
- Hardisty AR, Ellwood ER, Nelson G, Zimkus B, Buschbom J, Addink W, Rabeler RK, Bates J, Bentley A, Fortes JAB, Hansen S, Macklin JA, Mast AR, Miller JT, Monfils AK, Paul DL, Wallis E, Webster M (2022) Digital Extended Specimens: Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet. BioScience 72 (10): 978-987. https://doi.org/10.1093/biosci/biac060
- Kuhn T, Banda J, Willighagen E, Ehrhart F, Evelo C, Malas T, Dumontier M, Merono-Penuela A, Malic A, Poelen J, Hurlbert A, Centeno Ortiz E, Furlong L, Queralt-Rosinach N, Chichester C (2018) Nanopublications: A Growing Resource of Provenance-Centric Scientific Linked Data. 2018 IEEE 14th International Conference on e-Science (e-Science) https://doi.org/10.1109/escience.2018.00024
- Morris R, Dou L, Hanken J, Kelly M, Lowery D, Ludäscher B, Macklin J, Morris P (2013) Semantic Annotation of Mutable Data. PLoS One 8 (11). https://doi.org/10.1371/journal.pone.0076093
- Sanderson R, Ciccarese P, Van de Sompel H (2013) Designing the W3C open annotation data model. Proceedings of the 5th Annual ACM Web Science Conference https://doi.org/10.1145/2464464.2464474
- Tschöpe O, Macklin J, Morris R, Suhrbier L, Berendsohn W (2013) Annotating biodiversity data via the Internet. TAXON 62 (6): 1248-1258. https://doi.org/10.12705/626.4

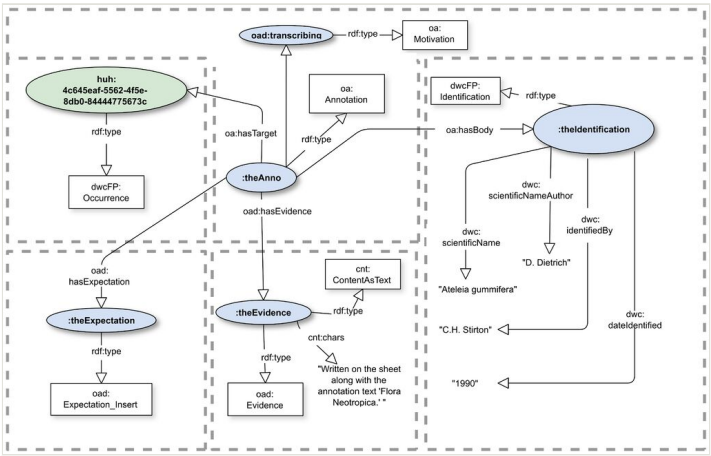

Figure 1.

Figure 2. Annotation providing a taxonomic identification.

Figure illustrates an abbreviated annotation providing a taxonomic identification for an occurrence record. The record is selected by reference to a lengthy identifier in the namespace of the Harvard University Herbaria (prefix “huh:”). [RDF S1] is a complete RDF representation in N3 syntax. The prefixes “oa:”, “oad:” and “dwcFP:” indicate terms respectively from the Open Annotation Ontology [61], the extension ontology we propose [Ontology S1], and a purpose built OWL ontology [Ontology S2] representation of the Darwin Core vocabulary [29]. https://doi.org/10.1371/journal.pone.0076093.g002