A Specimen-based View of the World

Using the Biological Collections Ontology to Model Biodiversity Collections

Ramona Walls
iPlant Collaborative
University of Arizona
Tucson, AZ, USA
rwalls@iplantcollaborative.org

John Deck
Berkeley Natural History Museums
University of California
Berkeley, California, USA
deck@berkeley.edline

Robert Guralnick
Ecology and Evolutionary Biology
University of Colorado at Boulder
Boulder, Colorado, USA
robert.guralnick@colorado.edu

Andréa Matsunaga
Dept. of Electrical and Computer Engineering
University of Florida
Gainsville, Florida, USA
ammatsun@ufl.edu

Abstract—**Application ontologies for biodiversity and biomedical specimen data are being developed within the OBO Foundry framework. Both the Biological Collections Ontology (BCO) and the Ontologized Minimum Information About BIobank data Sharing (OMIABIS) are ontologies rooted in specimens and the need to track, share and query data about specimens. In this paper, we briefly describe the structure of the BCO and the way that it is used to annotate and reason over biodiversity data. We conclude with a discussion of the relationship of the BCO to bio-bank ontologies and areas of potential collaboration through the Ontology for Biomedical Investigations (OBI).**

Keywords—ontology; biodiversity; specimen; material sample; Darwin Core; MIxS

I. Introduction

Museum specimens and the data associated with them are a critical foundation of biodiversity knowledge. They provide evidence of an organism's occurrence at a particular place and time and are source material for genetic, genomic, and metagenomic sequence data as well as for morphological, physiological, and biochemical trait measurements. Environmental data and field notes taken at the time of specimen collection or observation provide needed context that can form the basis of ecological studies into species' distributions and interactions (e.g. [1–3]). As our ability to measure and record scientific data has grown through technologies such as sequencing and digital data capture, so too has the need to store, track, access, and understand new types of specimens and their associated data.

The Biological Collections Ontology (BCO) [4] is a semantic model that describes and links both traditional and novel types of biodiversity data. While observations play a key role in biodiversity research and the BCO, the need to track and describe relationships between specimens, their origins, and their derivatives continues to be the BCO's primary driving use case. Although the BCO models basic biodiversity domain knowledge, it is primarily an application ontology that relies on imports from and coordination with other ontologies such as the Ontology for Biomedical Investigations (OBI) [5], the Environment Ontology (ENVO) [6], and the Population and Community Ontology (PCO) [4]. Coordination with ENVO is crucial for describing the environments in which specimens are collected, and coordination with PCO allows the BCO to describe multi-organism specimens, such as metagenomic samples.

The BCO grew out of a series of workshops [7,8] aimed at harmonizing traditional museum collection data, typically described using Darwin Core (DwC) [9], and genomic-based biodiversity data, typically described using using Minimum Information for any (x) Sequence (MIxS) [10,11]. Although MIxS is a standard for sequence data, there is overlap with specimen-based standards given that many of the MIxS terms describe the specimen that was sequenced or the conditions under which it was collected. The first term developed for the BCO was *material sample*[1], which was defined as a *material entity* (from the Basic Formal Ontology or BFO) [12,13] that **realizes** a **material sample role** by being the output of some *material sampling process*. As the BCO matured, it became apparent that BCO classes for *material sample* and *material sampling process* were very similar to *specimen* and *specimen collection* in OBI. Because these and many of the other concepts needed to describe biodiversity data were already present in OBI, a decision was made to coordinate BCO development with OBI. Classes that can be applied to biological investigations generally, such as *specimen*, should be housed in OBI, while only those specific to biodiversity studies, such as *museum specimen*, should be maintained in the BCO.

Funding was provided by EAGER: An Interoperable Information Infrastructure for Biodiversity Research (NSF-IIS-1255035), by RCN4GSC: A Research Coordination Network for the Genomic Standards Consortium (NSF-DBI-0840989), and by Collaborative Research: BiSciCol Tracker: Towards a tagging and tracking infrastructure for biodiversity science collections (NSF-DBI: 0956371, 0956350, 0956426).

[1] Ontology class names are shown in italic and relations in bold.

II. Ontology Design

The BCO is being developed according to OBO Foundry principles [14]. It is organized around two of the key processes that generate biodiversity data: *specimen collection* and *observing process* (Fig. 1). Both have a *material entity* as an input, but the key difference is that *specimen collection* generates a *material entity* (a *specimen*) as output while *observing process* generates an *information content entity* (from the Information Artifact Ontology or IAO) [15]. The BCO interprets *specimen collection* in a broad sense to include collection of museum or herbarium specimens, subsampling processes such as tissue sampling or DNA extraction, and collection of environmental (e.g., metagenomic) samples. Sequence generation and its output data are also crucial to biodiversity studies, but they have been modeled in OBI and the Sequence Ontology [16], so we import classes as needed for those concepts.

An essential functionality of the BCO is the ability to trace data through a series of processes. For example, one may have organismal sequence data stored in GenBank [17] and metagenomic data stored in another database and want to determine if those sequences came from the same museum sample (Fig. 2). To make queries like this, we needed a transitive property chain that links inputs and outputs of planned processes, which was not available in the Relation Ontology (RO) [18] or BFO. We created two relations using property chains, defined as follows:

is_specified_input_of o 'has output' subPropertyOf ‘derives from by planned process’
(http://purl.obolibrary.org/obo/BCO_0000067)

is_specified_output_of o 'has input' subPropertyOf ‘is derived into by planned process’
(http://purl.obolibrary.org/obo/BCO_0000068)

The **is_specified_input_of** and **is_specified_output_of** relations are from OBI and **has output** and **has input** are from the RO We are working with curators of RO to develop more broadly applicable definitions and names for these relations.

III. Using the BCO

One of the main uses of the BCO is to query over data sets that have metadata associated with both *specimens* and *specimen collection*. We have held several workshops in which we mapped column headings to ontology terms and specified relations among columns in order to convert datasets from the typical spread-sheet format to RDF [19]. Work is ongoing to develop tools that can automate mapping of data in common formats (such as Darwin Core archives or MIxS spread sheets) to RDF using BCO and other ontologies (see theBiSciCol Triplifier [20] and Biocode FIMS tools [21]). Three major challenges in this endeavor are that few researchers distinguish between specimens and specimen collection processes when they are recording data, the information content of many spreadsheets is ambiguous, and the lack of a clear standard for instance identifier assignment for biodiversity data.

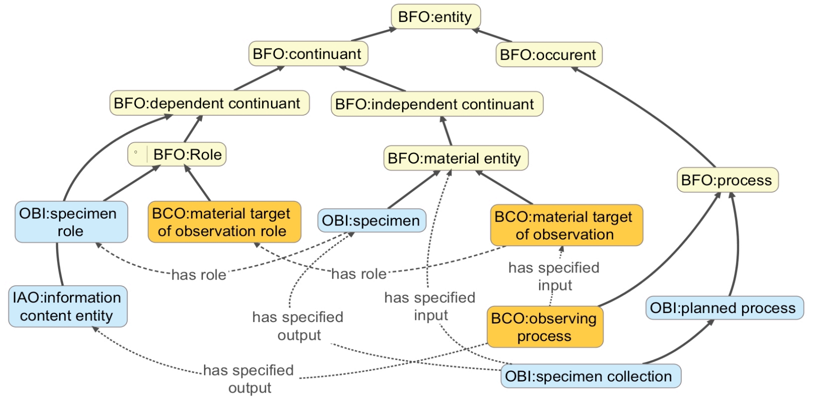

Fig. 1. Key classes of the BCO (orange), in a hierarchy based on BFO (yellow), OBI and IAO (blue). Solid arrows indicate subclassOf relations.

IV. The BCO and Biobank Ontologies

Similar to biodiversity specimen repositories, biomedical specimen repositories (biobanks) need to track and share data on specimens, their sources and their derived products or data. Here we compare the BCO to the Ontologized MIABIS (OMIABIS), named after the Minimum Information About BIobank data Sharing (MIABIS) [22]. OMIABIS is not the only existing biobank ontology, but is the only published, freely available one of which we are aware. We also considered the ontology described by [23] but do not have access to a current version for direct comparison. BCO and OMIABIS not only share a similar focus on specimens but also reuse many of the same terms from OBI and IAO. Some aspects that appear to differ between the two ontologies are in fact simply domain-specific differences in terminology. For example, descriptions of the environment in which a sample was collected in the BCO are, from a knowledge modeling perspective, very similar to a description of patient disease status and history in biobank ontologies, as both describe the conditions under which a specimen was collected. We recommend that developers of both ontologies work to develop a shared set of design patterns that can be used to model the environmental context of specimens, other aspects of specimen collection processes, and relations such as **derives from by planned process**, described earlier.

Both the biodiversity and biobank domains have standards for describing data (MIABIS for biobanks and DwC and MIxS for biodiversity specimens) and infrastructure for aggregating relevant data – the European Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) [24] for biobanks and the Global Biodiversity Information Facility (GBIF) [25] for biodiversity data. Nonetheless, the diversity of specimen and data types, the distributed nature of collections, and the novelty of informatic approaches to many researchers in both fields lead to uneven application of standards and additional challenges for semantic reasoning, particularly across legacy data sets. These challenges call for tools that make the ontologies easier to work with, on top of ontology development. We see an opportunity for BCO and OMIABIS developers to collaborate on tool development in areas such as universally unique specimen identifiers, data itegration across legacy data sources, and reasonig over large data sets.

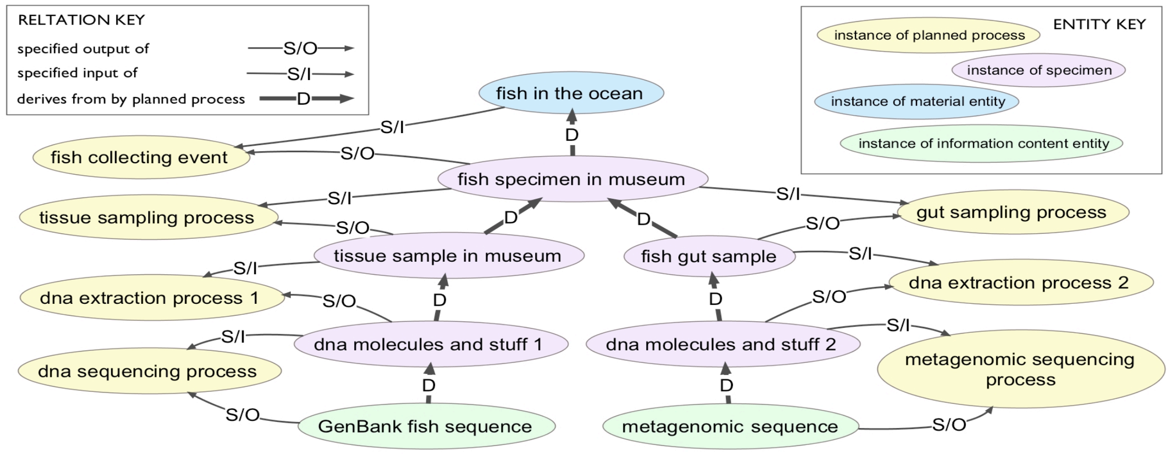

Fig. 2. Transitive derivation in the BCO. By declaring the **derives from by planned process** relation transitive and specifying a property chain for the relation based on inputs and outputs, we can infer that both the GenBank fish sequence and metagenomic sequence stored elsewhere are derived from the same fish, without having to assert the **derives from by planned process** relation.

It is clear that there are many areas of overlap and potential collaboration in modeling biodiversity specimen collections and biobanks. We are interested in discussing re-use of OBI terms without importing the entire OBI logic chain. Much of OBI's logic is not necessary for BCO's use cases and is likely to put off potential users, and we would like to learn if OMIABIS faces a similar situation. The only major conflict we found during this comparison was a difference in the version of BFO used by the two ontologies, and this is a conflict we think can be easily resolved. Concepts from OMIABIS, such as the **owns** and **administers** relations are highly useful and important to the BCO to capture the administrative data and relationships among biodiversity collections, may be better housed in the more general OBI. We recommend that curators from both domains work with the OBI to develop common terminology wherever possible.

References

[1] C.H. Graham, S. Ferrier, F. Huettman, C. Moritz, and A.T. Peterson, "New developments in museum-based informatics and applications in biodiversity analysis," *Trends Ecol Evol*, vol. 19, pp. 497-503, 2004.

[2] C. Moritz, J.L. Patton, C.J. Conroy, J.L. Parra, G.C. White, and S.R. Beissinger, "Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA," *Science* vol. 322, pp. 261-264, 2008.

[3] G.H. Pyke, P.R. Ehrlich, "Biological collections and ecological/environmental research: a review, some observations and a look to the future," *Biol Rev Camb Philos Soc*, vol. 85, pp. 247-266, 2010.

[4] R.L. Walls, J. Deck , R. Guralnick, S. Baskauf, R. Beaman, S. Blum, P.L. Buttigieg, et al., "Semantics in support of biodiversity knowledge discovery: an introduction to the Biological Collections Ontology and related ontologies," *PLoS ONE*, vol. 9, p. e89606, 2014.

[5] R. Brinkman, M. Courtot, D. Derom, J. Fostel, Y. He, P. Lord, et al., "Modeling biomedical experimental processes with OBI," *J Biomed Semant*, vol. 1(Suppl 1), p. S7, 2010.

[6] P.L. Buttigieg, N. Morrison, B. Smith, C.J. Mungall, S.E. Lewis, "The environment ontology: contextualising biological and biomedical entities," *J Biomed Semant*, vol. 4, p. 43, 2013.

[7] EÓ.. Tuama, J. Deck, G. Dröge, M. Döring, D. Field, R. Kottmann, et al., "Meeting Report: Hackathon-Workshop on Darwin Core and MIxS Standards Alignment (February 2012).," *Stand Genomic Sci*, vol. 7, pp.166-170, 2012.

[8] J. Deck, K. Barker, R. Beaman, P.L. Buttigieg, G. Dröge, R. Guralnick, et al., "Clarifying concepts and terms in biodiversity informatics," *Stand Genomic Sci*, vol. 8, pp. 352-359, 2013.

[9] J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, et al., "Darwin Core: an evolving community-developed biodiversity data standard," *PloS One*, vol. 7, p. e29715, 2012.

[10] P. Yilmaz, J.A. Gilbert, R. Knight, L. Amaral-Zettler, I. Karsch-Mizrachi, G. Cochrane, et al., "The Genomic Standards Consortium: bringing standards to life for microbial ecology," *ISME J*, vol. 5, pp. 1565-1567, 2011.

[11] P. Yilmaz, R. Kottmann, D. Field, J.R. Cole, L. Amaral-Zettler, J.A. Gilbert, et al., "Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications," *Nat Biotechnol*, vol. 29, pp. 415-420, 2011.

[12] P. Grenon, B. Smith, "SNAP and SPAN: towards dynamic spatial ontology," *Spat Cogn Comput*, vol. 4, pp. 69-103, 2004.

[13] B. Smith, "On classifying material entities in Basic Formal Ontology," in: *Interdisciplinary Ontology. Proceedings of the Third Interdisciplinary Ontology Meeting*. Keio University Press, Tokyo, 2012, pp. 1-13.

[14] http://www.obofoundry.org/wiki/index.php/Category:Accepted

[15] https://code.google.com/p/information-artifact-ontology/

[16] M. Bada, K. Eilbeck, "Efforts toward a more consistent and interoperable Sequence Ontology," in: *Proceedings of the 3rd International Conference on Biomedical Ontology (ICBO 2012)*. KR-MED Series. Graz, Austria; 2012.

[17] http://www.ncbi.nlm.nih.gov/genbank

[18] https://code.google.com/p/obo-relations/

[19] R.L. Walls, R. Guralnick, J. Deck, A. Buntzman, P.L. Buttigieg, N. Davies, et al., "Meeting report: Advancing practical applications of biodiversity ontologies," *Stand Genomic Sci*, unpublished.

[20] https://code.google.com/p/triplifier/

[21] https://code.google.com/p/biocode-fims/

[22] M. Brochhausen, M.N. Fransson, N.V. Kanaskar, M. Eriksson, R. Merino-Martinez, R.A. Hall, et al., "Developing a semantically rich ontology for the biobank-administration domain," *J Biomed Semant*, vol. 4, p. 23, 2013.

[23] A. Andrade, M. Kreutzhaler, J. Hastings, M .Krestyaninova, S. Schulz, "Requirements for semantic biobanks," *Stud Health Technol*, vol. 80, pp. 569-573, 2012.

[24] http://bbmri.eu/

[25] http://www.gbif.org