The Biological Collections Ontology links traditional and contemporary biodiversity data

Ramona L. Walls
CyVerse
University of Arizona
Tucson, AZ
rwalls@cyverse.org

Robert P. Guralnick
Florida Museum of Natural History
University of Florida
Gainsville, FL, USA

***Abstract*—Biodiversity data comes from many sources, ranging from museum specimens to field surveys to genomic sequences. Domain specific standards provide vocabularies for many types of these data, but they do not fully support integrating data across methods, scales, and domains. The Biological Collections Ontology (BCO) was designed to bridge the terminology gap between traditional museum-based specimen collections and more contemporary environmental sampling methods, such as metagenomic sequencing, by providing a logically defined set of terms for biodiversity that map to standards such as the Darwin Core and Minimum Information for any Sequence. The BCO is expanding to encompass observational biodiversity data such as field surveys and taxonomic inventories. A key design principle of the BCO is to clearly distinguish the different types of processes involved in biodiversity data collection along with the inputs and outputs of those processes. The BCO has applications to plant biodiversity studies for linking herbarium specimens to sequence data, connecting trait data to specimens, and describing survey data.**

Keywords—biodiversity; Darwin Core; museum specimen; observation; inventory; MIxS

I. Sources of Biodiversity data and corresponding standards

For hundreds of years, information about the existence and location of organisms has been preserved in museum specimen collections. Written observations by naturalists in field notebooks and on specimen labels are another related source of data, not only for occurrences but for traits and habitat data. Such observations are complemented by formal taxonimic surveys such as plot surveys, transects, and even floras for entire regions. Darwin Core (DwC) is a standardized vocabulary for sharing these types of biodiversity data [1, 2]. DwC is supported by the Biodiversity Information Standards organization (also known as the Taxonomic Databases Working Group or TDWG) [3]. DwC is intentionally simple in its structure and is not intended to be an ontology. The majority terms in DwC are metadata properties, recently formalized in RDF [4]. DwC properties cover areas such as taxon, location, identification, and geological context (important for fossil specimens). In addition, DwC has a small set of classes such as dwc:Occurrence and dwc:Location.

With the advent of molecular biology, a new type of biodiversity data emerged in the form of genetic, genomic, and metagenomic sequences. Although sequence data are comprised of a standardized series of symbols and stored in cannoical repositories such as NCBI [5], the associated data about sequencing methods and the specimen from which a sequence was derived, is often collected in an *ad hoc* manner and not included in published reports. The Genomics Standards Consortium is an open-membership working body formed in September 2005 with the goal of promoting mechanisms that standardize the description of genomes and the exchange and integration of genomic data, including metagenomes. GSC published the Minimum Information for any (x) Sequence (MixS) [6], which, like DwC, consists of a set of shared properties and is now available as RDF [7]. Core MixS terms cover areas such as sample collection, library preparation, and assembly, with environmental packages that contain terms specific to different types of samples (e.g., soil or water).

A third major source of biodiversity data are surveys such as vegetation plot surveys or regional taxonomic inventories. Unlike musuem specimens that generally record the presence of a single organism at a place and time, surveys can cover multiple taxa and occur over extended periods of time. Surveys also can provide information on the abundance or absence of taxa. Most surveys follow a sampling protocal, so information on the protocol needs to be captured along with the primary distribution and abundance data. Finally, any of the above data sources may generate data on the traits of an individual or taxa. Although DwC has been used to exchange survey and trait data, and trait data can be recorded using MixS, their relatively flat structure limits their ability to capture complex biodiversity data without loss of information.

II. The BCO

A. Overview

The Biological Collections Ontology (BCO) grew out of a series of workshops aimed at integrating traditional biodiversity data about museum specimens with more contemporary genetic and genomic biodiversity data [8-11]. A major outcome of these workshops was the differentiation between processes that yields specimens and processes that yield data (observing processes). After an initial release of the BCO, it was determined that the Ontology Biomedical Investigations (OBI) [12] already contained an appropriate design pattern and set of terms for specimen collection (http://purl.obolibrary.org/obo/OBI_0000659), and those terms were imported into BCO. OBI has a term for assay, which is the parent term for BCO observing process. BCO relies on the Information Artifact Ontology (IAO), which is closely aligned with OBI, for terms related to information and data.

B. *Integrating Darwin Core and MIxS with BCO*

Both DwC and MixS intentionally lack logical structure. Whereas this makes them broadly applicable, it severely limits the ability to make inferences or classificaitons from data associated with these vocabularies. BCO aims to provide a logical structure that is consistent with DwC and MixS, thereby providing an optional semantic layer. A seemlingly small but crucial step to the integration of DwC and MixS annotated datasets was the adoption of the OBI term specimen (http://purl.obolibrary.org/obo/OBI_0100051) by both TDWG and GSC, originally spearheaded by BCO curators. When used outside the context of BCO, this shared identifier simply provides consistent vocabulary for compatability across sources. If annotated datasets are mapped to BCO, it allows additional information related to specimens to be connected via logical definitions for inference [10] (see next section).

BCO imports all DwC terms and maps them to parent or equivalent BCO classes. For example, in BCO, dwc:Event is a subclass of OBI:planned process and dwc:Identification is equivalent to BCO:taxonomic identification. DwC properties are classified as data properties in BCO. To add reasoning power to these, work is underway that will link these data properties to object properties and classes in BCO.

As community standards adopted by multiple international organizaitons, both DwC and MixS must be stable and undergo a lengthy community review process before a term can be added or modified. BCO offers a more flexible proving ground for testing the efficacy and definitions of new terms such as specific subclasses of specimen or observing process. These new terms can then be considered for inclusion as part of a controlled vocabulary in MixS or DwC.

III. Applicability of BCO to plant biodiversity

A. *Linking herbarium specimens to their sequence data*

The primary use cases for the BCO is to link museum specimens to sequence data derived form a specimen. This is illustrated in detail in figure 3 in [10]. In brief, BCO can be used to specify a set of planned processes, each with specified inputs and outputs, which can then be used to infer the original source of a material or data. If MixS terms are used to annotate the sequence, they can be connected to data about the original specimen in a DsC archive by sharing the OBI:specimen identifier. Herbaria that use DwC to share specimen data now have the option of using a material sample core [13] (as opposed to an occurrence core) that explicitly specifies that the basis of the record of the occurrence is a specimen.

B. *Connecting trait data to specimens*

Another important use case for BCO is connecting trait data to the entity that bears the trait. This is relevant not only for biodiversity studies for population-level studues such as plant breeding. This case has several additional complications. First, traits can be measured at the organism, population, or species level. Second, trait data is often compiled from multiple sources, such us from an organism in situ, specimens, a photograph of a specimen or organism in situ, or published species descriptions. BCO models the relations among traits, entities that bear traits, evidence for a trait, and the taxonomic identification of the entity bearing the trait. BCO can also be used in with the Environment Ontlogy (ENVO) [14] to specify data about the environment, by connecting the environmental parameters to a location in which a specimen collecting or trait measurement event takes place.

C. *BCO for describing survey data*

Arguably the most complex type of biodiversity data is data about biodiversity inventories or surveys. Surveys generally include multiple locations and times as well as complex sampling protocols (e.g., Fig. 1 in [10]), and capturing computable information about surveys in a lossless way with flat formats such as a Darwin Core Archive is difficult or impossible. Work is underway in BCO to model this type of data, including the addition of terms for taxonomic inventory types.

IV. Conclusions

The BCO is used for many types of biodiversity data, including data about museum specimens, sequences, traits, and surveys. The BCO is under development and community input is welcome on the issue tracker [14] or mailing list [15].

RW is supported by CyVerse (NSF-DBI-0735191 and DBI-1265383),

References

[1] 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: e29715.

[2] http://rs.tdwg.org/dwc/terms/

[3] http://www.tdwg.org/

[4] http://rs.tdwg.org/dwc/terms/guides/rdf/

[5] http://www.ncbi.nlm.nih.gov/

[6] Yilmaz P, Gilbert JA, Knight R, Amaral-Zettler L, Karsch-Mizrachi I, Cochrane G, Nakamura Y, Sansone S-A, Glöckner FO, Field D. 2011. The Genomic Standards Consortium: bringing standards to life for microbial ecology. *The ISME Journal* 5: 1565–1567.

[7] https://github.com/pyilmaz/mixs

[8] Deck J, Barker K, Beaman R, Buttigieg PL, Dröge G, Guralnick R, *et al.* 2013. Clarifying concepts and terms in biodiversity informatics. *Standards in Genomic Sciences* 8: 352–359.

[9] Deck J, Guralnick R, Walls R, Blum S, Haendel M, Matsunaga A, Wieczorek J. 2015. Meeting report: Identifying practical applications of ontologies for biodiversity informatics. *Standards in Genomic Sciences* 10: 25.

[10] Walls RL, Deck J, Guralnick R, Baskauf S, Beaman R, Blum S, *et al.* 2014a. Semantics in support of biodiversity knowledge discovery: an introduction to the Biological Collections Ontology and related ontologies. *PLoS ONE* 9: e89606.

[11] Walls RL, Guralnick R, Deck J, Buntzman A, Buttigieg PL, Davies N, *et al.* 2014b. Meeting report: Advancing practical applications of biodiversity ontologies. *Standards in Genomic Sciences* 9: 17.

[12] Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, *et al.* 2016. The Ontology for Biomedical Investigations. *PloS One* 11: e0154556.

[13] http://tools.gbif.org/dwca-validator/extension.do?id=dwc:MaterialSample

[14] Buttigieg PL, Morrison N, Smith B, Mungall CJ, Lewis SE. 2013. The environment ontology: contextualising biological and biomedical entities. *Journal of Biomedical Semantics* 4: 43.