Biodiversity Ontologies

Ramona Walls1\*, Steve Baskauf2, Reed Beaman3\*, Pier Luigi Buttigieg4\*, John Deck5\*, Dag Endresen6, Maria Alejandra Gandolfo7, Robert Guralnick8\*, Robert Hanner9, Alyssa Janning10, Andréa Matsunaga11, Peter Midford12, Norman Morrison13\*, Barry Smith14, Andrea Thomer15, John Wieczorek16, Jamie Whitacre17

1The iPlant Collaborative, University of Arizona, Thomas J. Keating Bioresearch Building, 1657 East Helen St., Tucson, AZ 85721

2Department of Biological Sciences, Vanderbilt University, PMB 351634, 2301 Vanderbilt Place, Nashville, TN 37235

3University of Florida, Florida Museum of Natural History, 1659 Museum Rd., Gainesville, FL 32611

4Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven

5University of California at Berkeley, 3001 Valley Life Science Building, Berkeley, CA 94720

6GBIF Norway, Natural History Museum, University in Oslo, Sars gate 1, N-0562 Oslo

7LH Bailey Hortorium, Dept. of Plant Biology, 410 Mann Library Building, Cornell University, Ithaca, NY 14853.

8Department of Ecology and Evolutionary Biology and the CU Museum of Natural History, University of Colorado at Boulder, Boulder CO 80309-0265.

9Biodiversity Institute of Ontario, University of Guelph, 50 Stone Road East, Guelph, ON, Canada, N1G 2W1

10School of Information Resources and Library Science, University of Arizona, 1515 East First Street, Tucson, AZ 85719

11University of Florida, Gainesville, FL, 32611

12Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS, 660451

13The BioVeL Project, School of Computer Science, The University of Manchester, Oxford Road, Manchester, UK. M13 9PL

14Department of Philosophy, University at Buffalo, Buffalo, NY 14260.

15Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

163101 VLSB, Museum of Vertebrate Zoology, University of California, Berkeley, 94720

17National Museum of Natural History, Smithsonian Institution, Washington, D.C.

\*These authors contributed equally to the writing of the original manuscript draft.

Corresponding author: [rwalls@iplantcollaborative.org](mailto:rwalls@iplantcollaborative.org)

# Keywords

Standards in biodiversity, ontology, metagenomics, environmental sampling, ecology

# Abstract

# The study of biodiversity spans many disciplines and involves several types of data, including those pertaining to species occurrences and abundances, genetic sequence data, trait measurements, and environmental parameters, as well as information on collection and measurement protocols. An evaluation of the current landscape of metadata standards and ontologies in biodiversity science suggests that existing standards such as the [Darwin Core terminology](http://rs.tdwg.org/dwc/terms/index.htm) are inadequate for describing biodiversity data in a semantically meaningful way. On the other hand, existing ontologies, such as those in the [Open Biological and Biomedical Ontologies (OBO) Foundry Library](http://www.obofoundry.org/), lack many of the necessary terms to describe biodiversity data. In this paper, we describe the motivation for and ongoing development of the Biological Collections Ontology, the Environment Ontology, and the Population and Community Ontology. These ontologies share the aim of improving data aggregation and integration across the biodiversity domain and can be used to describe samples and sampling processes that encompass: 1) museum specimens or other samples that underpin taxonomic knowledge, 2) environmental samples, including metagenomic samples, and 3) survey-based ecological samples and observations. We provide a detailed discussion of how these ontologies can be applied to a real-life use case and argue that, if adopted as a standard and enriched by the biodiversity community, they would significantly reduce barriers to data discovery, integration, and exchange among biodiversity resources.

# Introduction

The loss of biodiversity is a major societal issue of our time [1, 2, 3]. Maintaining the diversity of genetic resources, species, and ecosystems is a critical investment for the future integrity of the biosphere. Trustworthy data about biodiversity will act as the basis of resource management and of policy decisions, and the development of infrastructures to support communication of this data is an essential investment in human and ecosystem sustainability.

The study of biodiversity now spans almost all biological disciplines, from field ecology to genomics, paleontology to neontology, systematics to phylogenetics, and depends on the collection of data on taxon occurrences and abundances, gene sequences, traits, and environmental parameters, as well as information on collecting and measurement protocols. The Convention on Biological Diversity (CBD) promotes a biodiversity management strategy that is based on surveys of existing biological resources as well as an understanding their value and importance [4] – a strategy highly dependent on long-term access to data for which the structure and content are well described and which can be used in a wide variety of meaningful analyses.

Biodiversity research is a branch of “big data” science that is still in its infancy [5]. While it has been suggested that the integration of taxonomic names will be the key to "the new big biology" [6], this is only one part of the data integration puzzle. Biodiversity science requires data about organisms, their morphology and genetics, their habitats, their geographical ranges, and the effects of habitat fragmentation on genetic diversity. However, all of these kinds of data are currently stored in a fragmented network of resource silos, where data are complex, highly heterogeneous, and incomplete. Additionally, current use of the term “biodiversity” (including that of the CBD) relates to organisms, populations, species, and whole ecosystems. As we move toward a global, scientifically-based understanding and survey of biodiversity [7, 8], traditional methods of description and classification are increasingly proving insufficient and are being complemented by scalable, high-throughput molecular biology methods. This implies an ever more urgent need for the means to better describe the many ways that biodiversity scientists capture and assemble data at different scales, along with resilient vocabularies, standards, and ontologies that will help scientists make use of these heterogeneous data in a reliable, harmonized manner – one that relies wherever possible on automatic reasoning rather than on *ad hoc* manual comparison and assemblage of data. Providing these tools allows us to build a dynamic picture of population and community assemblages and ultimately to test hypotheses of how these communities function and interact within a given niche, ecosystem, or region.

The notions of sampling, specimen collection, and observation are central to all areas of biodiversity research, yet despite their importance, the terminology describing samples and sampling processes is often vague and inconsistent across datasets, creating serious limitations for data aggregation, integration, re-use, and large-scale analyses. In this paper we describe the current landscape of metadata standards and ontologies in biodiversity science, especially as they relate to sampling. Existing standards, such as the [Darwin Core terminology](http://rs.tdwg.org/dwc/terms/index.htm), are not sufficient for describing the full gamut of biodiversity data in a semantically meaningful way, while existing ontology resources, such as the [Open Biological and Biomedical Ontologies (OBO) Foundry library](http://www.obofoundry.org/), lack many of the necessary terms to describe biodiversity data and metadata.

The lack of a clearly defined, semantically enriched terminology in the biodiversity domain is the motivation for development of ontologies that address this need. Herein, our main focus is the Biological Collections Ontology (BCO), but we also describe the Environment Ontology ([EnvO](http://www.environmentontology.org)) and the Population and Community Ontology ([PCO](http://code.google.com/p/popcomm-ontology/)). These ontologies share the goal of aligning to the same upper-level ontology – the Basic Formal Ontology (BFO) – that is also used by the other ontologies in the OBO Foundry library. Throughout this paper, ontology terms (classes and relations/predicates) are printed in italics. When it is not obvious from the context which ontology a term comes from, it is prefixed by the corresponding acronym (e.g., BCO: *material sample*). We discuss how this set of ontologies can be used to describe samples and sampling processes that encompass 1) museum specimens or other samples that underpin taxonomic knowledge, 2) environmental samples, including metagenomic samples, and 3) survey-based ecological samples and observations. We first describe the approaches used for collections across these related biodiversity domains, noting the commonalities and differences in processes and methods of sampling, pointing out where and how these affect the need for further integration. We then examine the landscape of previous work from which the Biological Collections Ontology was developed, before presenting it and describing its relationship to other upper-level and biodiversity-related ontologies. We close with a more detailed examination of broad-scale use cases of the BCO and its sibling ontologies in overcoming key impediments to doing global-scale, integrative biodiversity science.

**Approaches to Collection of Biodiversity Data**

*Museum collections based on individual organisms:* Museum and herbarium specimens are physical evidence that document biodiversity via individual organisms (or their parts) collected and preserved in perpetuity. These specimens are the subject of the morphological observations, descriptions, and citations that have underpinned the formal application of scientific names and taxonomy for over 250 years [9]. Specimens also provide the basis for measurements of phenotypic traits and trait variance, as well as a fundamental source of data pertaining to species distribution and abundance, and, often, data on the environmental characteristics of the collection location. The institutions that house scientific collections do so with a commitment to the permanent physical conservation of the collections, and typically also manage data about individual specimens, for example through specimen labels, collection catalogues, and field notebooks. Increasingly, data and metadata are being made available through institutional, national [e.g., [Atlas of Living Australia (ALA)](http://www.ala.org.au/), [Centro de Referência em Informação Ambiental (CRIA)](http://www.cria.org.br/), [iDigBio](https://www.idigbio.org/)), and international resources [e.g., the Global Biodiversity Information Facility, [(GBIF)](http://www.gbif.org/), [Biological Collection Access Services (BioCASE)](http://www.biocase.org/)], in a form adherent to standards such as the Darwin Core (DwC) [10]. A basic use case for biological collections hinges on the idea that specimens are themselves samples taken from the environment through specific collection events. They thereby inherit data pertaining to such events (e.g., as identified under DwC:Event), which can serve as sources for further sub-sampling. Specimens also accrue data after collecting events, for example through events of identification (DwC:Identification) that link the specimen to a particular taxon name (DwC*:*Taxon).[[1]](#footnote-2)

New data products – both digital and physical – can be derived from specimens. In some cases this involves preparations that are destructive in the sense that they involve removing a physical subsample from the original material, such as harvesting of tissues for DNA extraction and follow-on genetic and genomic analyses. A subsampling event may occur at any stage in the curatorial workflow, and in many cases the subsamples find their way into other types of collections, for example, in cryo-facilities, which are independent from that of the source specimen. Other preparations include the acquisition of digital images and other media objects. Tracking objects and data across multiple resources, and communicating relationships derived from specimen subsampling and distribution to multiple repositories (physical and digital), is not easily accommodated by current data infrastructure.

*Environmental samples, including metagenomic samples:* The field of metagenomics addresses the function and composition of microbial communities [11]. The main distinguishing feature of metagenomics is that it is concerned with the analysis of DNA from whole communities, compared to genomics, which is concerned with the analysis of DNA from a single cell or organism. A goal of metagenomics is to study ecosystem functioning by connecting taxonomic composition to function and environmental characteristics at the both local and global level. In contrast to what is the case in museum collections, approaches to biodiversity based on environmental sampling and metagenomics often yield samples whose species composition is not known. Samples themselves thus become the key targets of data integration for biodiversity assessment. These samples are contextualized by properties of their environment of origin such as location and associated physio-chemical properties. The sample must be physically extracted from some environmental material (e.g., a liter of ocean water, a cubic centimeter of soil, the contents of an animal’s gut) and then assayed to determine something about what is contained within that specific sample. The sequences derived from metagenomic analysis of an environmental sample inherit the data pertaining to location and environmental conditions linked to the sample itself. Other analyses of the sample, for example biogeochemical analyses to measure particular compounds, can also provide information on the environment affecting the sequenced community. As sampling procedures may be performed on virtually any material entity, environmental samples may also be extracted from the interiors of other organisms.

*Survey-based, ecological approaches:* Ecological survey methods are heterogeneous, but they are often based on a defined time spent counting the number of individuals or species within a particular spatial range, such as a linear transect or a gridded area. The target of the survey is often a measure of abundance and distribution, rather than to a point occurrence, since multiple units may be contained within the relevant sampling area. The targets may also be hierarchical with one nested within the other, for example, a leaf within a plant within a subplot within a plot (Figure 1). As a result of this nesting, environmental variables associated with a plot can also be associated with a leaf collected within that plot and with the DNA or other moletcules extracted from that leaf. In contrast to metagenomic studies, many ecological studies are based on observations or measurements taken from samples that are neither conserved nor curated. Databasing efforts (e.g., [TRY](http://www.try-db.org), [BIEN](http://www.bien.org.au/), [NEON](http://www.neoninc.org)), while sufficient for capturing the complexity of ecological sampling, do not effectively link data or metadata collected at one level of sampling layer with another layer of that study or with data from other studies. Linking ecological data derived from the application of survey methods or from sampling efforts to terms from ontologies such as the BCO, EnvO, and PCO can provide the semantic framework needed for automated access to, and reasoning over, what is potentially a huge source of networked data.

Below we examine further how the process of sampling in the museum collections community coincides with that in environmental and ecological communities through the BCO.

## The biodiversity standards landscape

The Biodiversity Information Standards organization (TDWG) is a community dedicated to the development of standards for biodiversity information. TDWG has ratified and maintains the Darwin Core (DwC) [10] and the Access to Biological Collections Data (ABCD) [12] standards for documentation of specimens in biological collections. TDWG initiated an ontology effort at their 2006 annual conference to build a semantic framework for the TDWG standards tied to Life Science Identifiers (LSIDs). This initiative was named the TDWG LSID Ontology (<http://rs.tdwg.org/ontology/>) and was developed using the Web Ontology Language (OWL; [http://www.w3.org/TR/owl2-overview/](http://www.w3.org/TR/owl-features/)) [13]. The foundation of the TDWG LSID Ontology was defining a base ontology that would play a similar role to that played by BFO within the OBO Foundry framework [14].

For a variety of socio-technical reasons, the TDWG LSID ontology, which was built as a prototype for the TDWG Technical Architecture Group, was effectively discontinued. At the 2011 TDWG conference, a new ontology with a more limited focus on the Darwin Core terminology was presented. This product was dubbed Darwin Core Semantic Web ([D-SW](https://code.google.com/p/darwin-sw/)) [15]. D-SW provides pairs of inverse object properties that can be used to relate instances of DwC-defined classes. It thereby codifies a particular outlook on the relationships among the DwC classes that includes a semantic separation of museum specimens and species observations (DwC:Occurrence) [16]. A task group established in 2010 by the Global Biodiversity Information Facility (GBIF) to explore options for the implementation of Knowledge Organization Systems for biodiversity information standards proposed to initiate a closer integration between the TDWG standards and the OBO framework [17, 18]. The work presented here for the BCO is primarily based on ensuring semantic compatibility between the OBO Foundry framework and the Darwin Core standard.

*Review of Recent Activities*

In 2011, the NSF-funded Research Coordination Network project for the Genomic Standards Consortium (RCN4GSC, hosted at UCSD, with John Wooley as PI) [19, 20] began a series of meetings to reconcile discrepancies among terms between the DwC and Minimum Information about any (x) Sequence (MIxS) standards [21]. The vocabulary alignment meetings [22] recognized conceptual differences and conflicts between the use of the terms sample, specimen, and occurrence. In response, the RCN4GSC organized a Semantics of Biodiversity (SoB) workshop in Lawrence, Kansas in May of 2012. SoB brought together a range of domain experts to comment on a proposal for aligning terms within a larger framework, using the Basic Formal Ontology (BFO) as a guide. BFO was used to both provide a conceptual framework for problematic terms and help clarify terms in the DwC vocabulary and illustrate mechanisms for managing vocabularies.

Building on SoB, another workshop was held at [GSC14](http://gensc.org/gc_wiki/index.php/GSC_14), in Oxford [23] to build the framework for a bio-collections ontology. The purpose of this workshop was to undertake a formal definition of samples and sampling processes and formalize the concepts outlined at the SoB workshop as an ontology. This workshop composed the first formal definitions for the BCO, including terms for samples and sampling, as described below under the section for *The Biological Collections Ontology*.

### *Developing within BFO*

The BCO, PCO, and EnvO each model distinct portions of reality, or domains, pertinent to biodiversity research. In aid of interoperability and semantic homogeneity, these ontologies either are aligned or are currently being aligned to the BFO [24, 25]. BFO is an upper-level ontology and is concerned with providing formal specifications of basic metaphysical entities, such as *spatial region, temporal region*, *object*, *quality*, and *process*,from which to derive the semantics of entities in a given domain ontology. Domain ontologies sharing this formal basis are, thus, more readily interoperable with all other BFO-derived ontologies (e.g., the Gene Ontology or GO [26, 27] and the Sequence Ontology or SO [28]) allowing broader scope for cross-domain knowledge representation and reasoning. Additionally, BFO is associated with several formal theories that may be re-used by every domain ontology aligned to it. For example, domain ontologies may avail themselves of formal mereotopological notions – i.e. notions of wholes, parts, parts of parts, and boundaries – associated with BFO without having to define them independently. Users of and contributors to the emerging biodiversity and environment ontologies BCO, PCO, and EnvO may thus interface with a growing and cross-domain semantic resource to enhance their research.

## Results – Ontology descriptions

### *The Biological Collections Ontology*

The BCO is being developed in keeping with the [OBO Foundry principles](http://obofoundry.org/wiki/index.php/Category:Accepted), using in the Web Ontology Language and Protégé software (<http://protege.stanford.edu>). A draft version of the BCO is available to view or download in OWL at<http://biocode-commons.googlecode.com/svn/trunk/ontologies/biocollections/biocode.owl>, or the draft can be browsed via BioPortal at<http://bioportal.bioontology.org/ontologies/3201>.

Development in the BCO to date has focused on the terms *material sample* and *material sampling process*, plus related terms (Figure 2). A *material sample* (Figure 2A) is defined as a BFO:*entity* that is the product of a *material sampling process* and which has a *material sample role*. Examples include a jar of ocean water, a herbarium specimen, a preserved animal in a museum collection, or a fossil specimen. If a sampling process is carried out on a *material sample*, the resulting *material sample* will be conceptually similar to the procedural notion of “subsample”. This is the consequence of *material samples* being sequentially extracted from a *material entity* with nested parts. For example, in an experimental process where DNA was extracted from a sample of a microbial community, itself extracted by filtration from a jar of marine water, *material samples* downstream of the jar of marine water can be called subsamples. We are considering ways to logically connect samples and subsamples in the BCO, for example, through addition of a relation such as *is\_subsample\_of* that links two instances or through a new term for *subsampling process*. Alternatively, an instance-level representation of the targets and products of a *sampling process* could be used to identify procedural subsamples, without the creation of an explicit *subsample* class.

*Material sampling process* (Figure 2B) is a subclass of *planned process*, from the Ontology for Biomedical Investigations ([OBI](http://obi-ontology.org/page/Main_Page)) [29]. Three other types of processes are also needed: a *selecting process* (a planned process by which a person or machine decides that a particular material entity as worthy of collection), a *physical extraction process* (a planned process that involves removing a material sample from one site to another), and a *submitting process* (a planned process whereby a person submits a material sample to an organization). A *material sampling process* is distinguished from an *observing process* in that an *observing process* has as output an *information content entity* (from the Information Artifact Ontology or [IAO](http://code.google.com/p/information-artifact-ontology/)) about some BFO:*material entity*, rather than a *material sample*, although both processes have a *selecting process* as a part. Other processes related to biodiversity studies, such as photographing organisms or specimens, will also be covered by the BCO.

The BCO imports a number of terms from other ontologies, including the entirety of the BFO and the IAO. The BFO serves as the upper-level framework for the BCO, while the IAO provides IAO:*information content entity* and its subclasses. These include IAO:*data item*, which represents information generated as a result of an assay, such as a DNA sequence as found in Genbank, or IAO*:plan specification*, which represents the information in an experimental protocol, for example. CARO:*organism* andCARO*:organism or virus or viroid* are imported from the Common Anatomy Reference Ontology [30].

### OBI is an ontology that aims to enable the description of all biological and clinical investigations with a shared vocabulary [29]. The notions of material sample, subsample, observation, and measurement are all covered in OBI, but were developed independently of the BCO effort, Coordination between OBI and BCO is just beginning. At first blush, the shared terms align well between the BCO and OBI, and several OBI terms are included directly in the BCO. However, given the much broader scope of OBI, many of OBI’s term names are not familiar to biodiversity scientists, even if the definitions are familiar. We are considering ways to align the BCO with OBI, such as importing OBI terms into BCO while providing them with alternative labels readily understandable to biodiversity scientists. The future use of OBI terms in BCO will ensure that it is interoperable with other descriptions of investigations in the OBO foundry.

### *The Environment Ontology*

EnvO is a community-developed ontology for the standardized description of environments. Any instances of a BCO:*material sample* or BFO:*process* may be annotated with and contextualized by EnvO terms. Pertinent examples include instances of a BCO: *material sample, sampling process,* or *observing process* as well as those of a PCO:*species, population,* or *community*. Importantly, EnvO’s terms make no reference to specific locations; however, EnvO annotations are naturally linked to geospatial information, which may be expressed with resources such as [GAZ](http://gensc.org/gc_wiki/index.php/GAZ_Project), a first step towards an open source gazetteer constructed on ontological principles.

EnvO is currently being aligned with the OBO Foundry principles and BFO. The latest version, in OBO format, may be found on EnvO’s website ([www.environmentontology.org](http://www.environmentontology.org)) or through its Google code repository (<http://envo.googlecode.com>). Recently, EnvO has been approved as a project under the Genomic Standards Consortium (GSC) framework and is used in the MIxS checklist [21]. Several other users are listed on the EnvO website.

As of 2013-Mar-01, EnvO comprised 1556 classes[[2]](#footnote-3) partitioned into three hierarchies: *biome,* *environmental feature*, and *environmental material*. When annotating entities with EnvO, terms from each of these hierarchies should be combined to describe environments across a range of granularity. EnvO’s *biome* terms aim to describe the broad, ecosystemic context in which a given entity (henceforth: an occupant) is immersed. Rather than being scale-dependent, the boundaries of a given *biome* are defined by the spatial range of a PCO:*community*, which shows adaptations to a set of structuring environmental conditions (such as water or light availability) conserved over that *biome*. Next, children of the *environmental feature* class are intended to describe some physical discontinuity (feature) that strongly influencesthe occupant’s local environment. This hierarchy is broad and may include geographic features alongside anatomical features, therefore, terms from other domain ontologies will often be imported when augmenting thishierarchy. Finally, children of the *environmental material* class describe the ecological medium or media immediately surrounding an occupant. An example of a valid, minimal annotation of a pelagic shark observed feeding near a shallow coral reef would be: *neritic epipelagic zone biome*, *coral reef*, *coastal water*.

Work is underway to define the autecological concepts of *habitat* and *niche* with reference to the relevant concepts in PCO. These classes, together with *biome, environmental feature,* and *environmental material,* will lay the foundation for more refined and standardized semantic handling of these key ecological concepts.

### *The Population and Community Ontology*

The PCO is rooted in the BFO and contains terms related to collections of organisms such as populations, communities, and species. Upper-level terms in the PCO include *collection of organisms* (a subclass of BFO:*material entity* that includes PCO:*population*, PCO:*community*, and PCO:*species*), *population process* and *community process* (subclasses of BFO:*process*), and *populations quality* and *community quality* (subclasses of BFO:*quality*). *Population*, *community*, and *species* are defined broadly, in recognition of the many ways that these terms have been used and defined in the scientific literature. The PCO also includes terms for the roles of interacting organisms, such as *symbiont role* or *predator role*. Relations that could be used to describe inter-organismal interactions, such as *preys on* or *pollinates*, are under consideration. Terms related to individual organisms and their parts are already covered in other ontologies such as CARO, GO, and clade-specific anatomy ontologies, but may be imported into the PCO as needed.

Many of the processes needed to describe species interactions, such as *competition with other organism* or *symbiosis encompassing mutualism through parasitism*, are covered by the *biological process* branch of the GO [26, 27] and are subclasses of GO:*multi-organism process*. The bulk of these terms were developed as part of the PAMGO project [31], and focus on interactions among micro-organisms or between pathogens and hosts. As a result, many processes that are important for other use cases, such as *predation* or *herbivory*, are missing from the GO. A goal of the PCO project is to develop these terms in collaboration with the appropriate domain experts, and then continue discussions with GO curators about if and how they should be integrated into the GO.

Biological processes that occur at a population level but pertain to infectious diseases (such as IDO:*disease epidemic*) are covered by the Infectious Disease Ontology (IDO) [32]. Other processes found particularly in animal populations, such as migration or animal culture, should be represented in PCO at the population level. Although *migration* is defined in the IDO-Malaria ontology [33], its definition and placement is specific to mosquitoes as malaria vectors. Terms for such processes may be related to terms for the underlying behavior processes (e.g., *migratory behavior*, *social learning*) in the NeuroBehavior Ontology (NBO) [34].

PCO:*population quality* includes subclasses such as *carrying capacity* and *population growth rate*, and PCO:*community quality* includes subclasses such as *community species richness*. The development of these terms is taking place in consideration of the Phenotypic Quality Ontology (PATO) [35]. PATO already has a term *population quality*, but it has only *mixed sex* and *morbidity* as subclasses, and no term for community quality exists in PATO. As with GO processes, PCO curators are working to develop population and community quality terms with domain experts, then will work with PATO curators to determine which ontology is appropriate to host those terms.

The PCO aims to serve the bioinformatics needs of ecology and evolutionary biology as well as other population-based studies (e.g., biomedical research or community healthcare). Within the context of biodiversity studies, PCO terminology is important for describing multi-organism (e.g., metagenomic or ecological) samples and sampling, as well as for the construction of logical definitions of terms such as niche or habitat (see section on EnvO, above). The PCO is in the early stages of development, and anyone who is interested in using or contributing terms may contact the curators through the PCO mailing list or [issue tracker](http://code.google.com/p/popcomm-ontology/issues/list).

## Discussion

Ontologies can be used to link information on samples and their derivatives as materials and data move through various processes and institutions. We show the complexity of the problem domain and the utility of the ontologies described above by focusing on a use case represented diagrammatically in Figure 3. The use case is one based on sampling protocols for the [Moorea Biocode project](http://mooreabiocode.org/). This project aims to create the first comprehensive inventory of all non-microbial life in a complex tropical ecosystem by creating a library of genetic markers and physical identifiers for every species of plant, animal, and fungus on the island. As part of this project, museum specimens, such as insects, are collected (Essig Museum specimen, Figure 3A). From a specimen, tissue may be sampled for preservation (Smithsonian tissue sample, Figure 3A) and for sequence analysis (Genbank sequence, Figure 3A). In addition, the gut contents of the insect may be collected for metagenomic sequencing and submission to the [CAMERA portal](http://camera.calit2.net/) (CAMERA gut sample and metagenomic sequence, Figure 3A). Each step in this process follows protocols, has inputs and outputs, and is accompanied by metadata collection. Starting at any step in the chain, researchers need to access data/metadata associated with any other step.

Figure 3B shows selected terms from ontologies that can be used to annotate biological sampling data, plus their relationships to BFO terms (for a more complete BFO hierarchy, see Figure 2). Sequences (metagenomic or single species) can be annotated with the SO term *region*, but because the SO is currently in process of being formally aligned with the BFO, we do not show its relationship here (see [36] for more details)

Figure 3C is a graphical representation of the annotation of the workflow shown in Figure 3A, using ontology terms from Figure 3B. Instances are shown as rectangles, ontology classes as ovals, and relations between instances or classes as arrows. Color coding is used to link instances in Figure 3C to classes in Figure 3B (e.g., white boxes with blue outlines in Figure 3C represent instances of BCO:*material sample*, which is shown as the white oval with a blue outline in Figure 3B).

One outcome of the annotation process is to enable a linked data approach [37] by representing relationships between instances and between instances and ontology term identifiers, using uniform resource identifiers (URIs) as globally unique identifiers. The BiSciCol project (<http://biscicol.blogspot.com/>) is implementing such an approach, storing relationships harvested from community-accessible datasets in a database, enabling queries using relevant ontologies. Some examples of the types of queries that could be performed in the context of the Moorea Biocode use case include:

1. Do taxon identifications determined through keying (e.g., TaxonID A in Figure 3C) and those determined through sequencing (e.g., TaxonID Y in Figure 3C) provide consistent information regarding the identity of the originally collected specimen?
2. What are the EnvO:*biome* and other environmental parameters associated with metagenomic sequence X (as recorded during the Moorea Biocode sampling process)?
3. What are all the taxa that have been collected as part of the Moorea Biocode project and where can I find the specimens and DNA samples associated with those taxa?

The linked open data framework provides the means to answer key questions not only within a particular large-scale e-infrastructure project, but also across such projects. Thus, Internet-scale discoverability, integration, and machine-inferencing provided by linked data support beautifully the identified need for global scale, integrated, and fine-grained information on the state of the planetary biosphere, and especially on biodiversity and changes therein.

For the sake of clarity, Figure 3C does not show every relationship that could or should be annotated in this workflow. Moreover, this use case, although usefully comprehensive in showing many possible inputs and outputs and chains of provenance that need tracking, only scratches the surface of the use cases that can be modeled using carefully constructed ontologies in biodiversity and ecology domains. Many use cases relevant to biodiversity studies will not directly involve the collection of specimens at all. Implicit in the gene-based view of biodiversity is that the genes accumulated from a metagenomics analysis belong to specimens that are present but effectively destructively sampled during analyses. The gene(s) sampled might only provide the roughest guide to accurately represent the actual specimen and its link into taxonomy. Individual analyses, in which phylogenetic trees or operational taxonomic taxonomic units can be used, are not as much a challenge in the gene-based world as how to re-use sequence information, especially given the large number of sequences that have no reference to taxonomic names at all. How can trees or operational units be easily combined across analyses? How can the trees and operational units be reconciled with names and specimens? Solving the problem of how to reconcile genes, specimens, and samples remains a core task that will become a key part of how BCO evolves in the future.

Annotation of sequence submissions has been developed as a Genomic Standard Consortium (GSC) standard using the MIxS family of checklists, providing an avenue for contextualizing sequences at the time of submission to repositories. The MIxS family consists of checklists for genome/metagenome sequences (MIGS/MIMS) and marker sequences (MIMARKS) with shared descriptors across all three checklists, checklist specific descriptors, and a suite of environmental packages adopted from EnvO [21]. In Figure 3c, sequence annotation is performed between the sequencing and sequence alignment activities (BLAST), describing events and material leading up to the sequence. A challenge, then, is integrating BCO outputs with both the MIxS and DwC standards, providing consistency of terms and conceptual grounding across the biodiversity (DwC) and molecular (MIxS) namespaces.

In addition to the types of samples described in Figure 3, many biodiversity datasets reference neither specimens nor genes, but instead simply provide a list of taxa observed in an area, or even of taxa confirmed not to have been detected. The BCO will need to grow to encompass such ecological survey and inventorying data, which would allow a way to more effectively integrate different sources of knowledge via shared, linked and well-understood terms (e.g., DwC:taxon identification or BCO:*identification process*).

Figures 1 and 3 nicely illustrate one further complexity worthy of discussion: biodiversity knowledge is structured spatially and hierarchically. For example, a community of microbes from the gut of a collected insect in Figure 3 may relate not only to the topological relationships of the gut described in the Arthropod Anatomy Ontology ([AAO](http://code.google.com/p/arthropod-anatomy-ontology/)), but also to the local environmental conditions peculiar to the gut, and possibly even to the external environmental conditions measured at the time of collection. BCO, in conjunction with DwC, EnvO, PCO, and potentially other ontologies such as AAO, provides a powerful semantic language to create such spatially nested and hierarchical sets of comparisons.

**Acknowledgements**  
We gratefully acknowledge the support from the US National Science Foundation (NSF) through the following grants: DBI-0840989 (RCN4GSC), EAGER, 0956371, 0956350, 0956426 (BiSciCol), DBI-0735191(The iPlant Collaborative). We also thank the contributions of participants of the Semantics of Biodiversity Workshop in Lawrence, Kansas (May, 2012) and the Biocode Ontology Hackathon in Oxford, England (October, 2012) who are not co-authors of this manuscript: Vijay Barve, Jim Beach, Matthiew Bietz, Stan Blum, Shawn Bowers, Gabi Droege, Michelle Koo, Kris Krishtalka, John Kunze, Chuck Miller, Gil Nelson, Cynthia Parr, Sujeevan Ratnasingham, Jai Rideout, Robert Robbins, Joel Sachs, Inigo San Gil, Herbert Schentz, Mark Schildhauer, Peter Sterk, Steve Stones-Havas, Brian Stucky, Mellisa Tulig, Dave Vieglais, Brian Wee, Trish Whetzel, and Greg Whitbread.

**Literature Cited (preparing for SIGS formatting)**

1. Pereira HM, Leadley PW, Proença V, Alkemade R, Scharlemann JPW, Fernandez-Manjarrés JF, Araújo MB, et al. Scenarios for Global Biodiversity in the 21st Century. Science 2010;330(6010):1496–501. doi:10.1126/science.1196624
2. Pereira HM, Ferrier S, Walters M, Geller GN, Jongman RHG, Scholes RJ, Bruford MW, et al. Ecology. Essential Biodiversity Variables. Science 2013; 339(6117):277–8. doi:10.1126/science.1229931
3. Cardinale BJ, Duffy JE, Gonzalez A, Hooper DU, Perrings C, Venail P, Narwani A, Mace GM, Tilman D, Wardle DA, Kinzig AP, Daily GC, Loreau M, Grace JB, Larigauderie A, Srivastava DS, Naeem S. Biodiversity loss and its impact on humanity. Nature 2012;486:59–67. doi:10.1038/nature11148
4. United Nations. 1993;Convention on Biological Diversity. Opened for signature at the Earth Summit 5 June 1992, Rio de Janeiro, Brazil. Available at <http://treaties.un.org/doc/Treaties/1992/06/19920605%2008-44%20PM/Ch_XXVII_08p.pdf>, verified 2 April 2013
5. Kelling S, Hochachka HM, Fink D, Riedewald M, Caruana R, Ballard G, Hooker G. Data-Intensive Science: A New Paradigm for Biodiversity Studies. BioScience 2009;59(7):613–620. doi:10.1525/bio.2009.59.7.12
6. Patterson DJ, Cooper J, Kirk PM, Pyle RL, Remsen DP. Names are key to the big new biology. Trends in Ecology & Evolution 2010;25(12):686-691. [PMID:20961649](http://www.ncbi.nlm.nih.gov/pubmed/20961649) [doi:10.1016/j.tree.2010.09.004](http://dx.doi.org/10.1016/j.tree.2010.09.004)
7. Jetz W, McPherson JM, Guralnick RP. Integrating biodiversity distribution knowledge: toward a global map of life. Trends in Ecology & Evolution 2012;27(3):151-159. [PMID:22019413](http://www.ncbi.nlm.nih.gov/pubmed/22019413) [doi:10.1016/j.tree.2011.09.007](http://dx.doi.org/10.1016/j.tree.2011.09.007)
8. Davies, ZG, Fuller RA, Loram A, Irvine KN, Sims V, Gaston K. A National Scale Inventory of Resource Provision for Biodiversity Within Domestic Gardens. Biological Conservation 2009;142 (4):761–771. doi:10.1016/j.biocon.2008.12.016
9. Beach J, Blum S, Donoghue M, Ford L, Guralnick R, Mares M, Thiers B, Westneat M, Wheeler Q, Wiegmann B and others. A Strategic Plan for Establishing a Network Integrated Biocollections Alliance. 2010. <http://digbiocol.wordpress.com/brochure/>
10. Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 2012;7(1):e29715. [PMID:22238640](http://www.ncbi.nlm.nih.gov/pubmed/22238640) [doi:10.1371/journal.pone.0029715](http://dx.doi.org/10.1371/journal.pone.0029715)
11. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology 1998;5(10):R245-R249. [PMID:9818143](http://www.ncbi.nlm.nih.gov/pubmed/9818143) [doi:10.1016/S1074-5521(98)90108-9](http://dx.doi.org/10.1016/S1074-5521(98)90108-9)
12. Holetschek J, Dröge G, Güntsch A, Berendsohn WG. The ABCD of primary biodiversity data access. Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology 2012;146(4):771-779. [doi:10.1080/11263504.2012.740085](http://dx.doi.org/10.1080/11263504.2012.740085)
13. Horridge M, Drummond N, Goodwin J, Rector A, Stevens R, Wang HH. The Manchester OWL Syntax. In: Proc. of the 2006 OWL Experiences and Directions Workshop (OWL-ED2006), Athens, Georgia (USA) November 10-11, 2006, Cuenca Grua B and Hitzler P, eds. Available online at <http://ceur-ws.org/Vol-216/submission_9.pdf>
14. Kennedy J, Gales R, Hyam R, Kukla R, Wieczorek J, Hagedorn G, Döring M, Vieglais D. Developing a core ontology for taxonomic data. In: Proceedings of Taxonomy Data Working Group (TDWG) 2006, 2006, Missouri Botanical Gardens, St. Louis, Missouri, USA; 15-22 October 2006. <http://www.tdwg.org/proceedings/article/view/13>
15. Webb C, Baskauf S. Darwin-SW: Darwin Core data for the Semantic Web. In: Proceedings of Taxonomy Data Working Group (TDWG) 2011, 2011, New Orleans, Louisiana, USA. <https://mbgserv18.mobot.org/ocs/index.php/tdwg/2011/paper/view/152>
16. Baskauf SJ. Organization of occurrence-related biodiversity resources based on the process of their creation and the role of individual organisms as resource relationship. Biodiversity Informatics 2010;7(1):17-44.
17. Lapp H, Morris RA, Catapano T, Hobern D, Morrison N. Organizing our knowledge of biodiversity. Bulletin of the American Society for Information Science and Technology 2011;37(4):38-42. [doi:10.1002/bult.2011.1720370411](http://dx.doi.org/10.1002/bult.2011.1720370411)
18. Catapano T, Hobern D, Lapp H, Morris RA, Morrison N, Noy N, Schildhauer M, Thau D. Recommendations for the use of knowledge organization systems by GBIF version 1.0. Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark. 2011. <http://links.gbif.org/gbif_kos_whitepaper_v1.pdf>
19. Wooley J, Field D, Glöckner FO. Extending Standards for Genomics and Metagenomics Data: A Research Coordination Network for the Genomic Standards Consortium (RCN4GSC). Standards in Genomic Sciences 2009;1(1):159. [PMID:21304642](http://www.ncbi.nlm.nih.gov/pubmed/21304642/) [doi:10.4056/sigs.26218](http://dx.doi.org/10.4056/sigs.26218)
20. Robbins RJ, Amaral-Zettler L, Bik H, Blum S, Edwards J, Field D, Garrity G, Gilbert JA, Kottmann R, Krishtalka L and others. RCN4GSC Workshop Report: Managing Data at the Interface of Biodiversity and (Meta)Genomics, March 2011. Standards in Genomic Sciences 2012;7(1). [PMID:23451294](http://www.ncbi.nlm.nih.gov/pubmed/23451294) [doi:10.4056/sigs.3156511](http://dx.doi.org/10.4056/sigs.3156511)
21. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G and others. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotech 2011;29(5):415-420. [PMID:21552244](http://www.ncbi.nlm.nih.gov/pubmed/21552244) [doi:10.1038/nbt.1823](http://dx.doi.org/10.1038/nbt.1823)
22. Tuama ÉÓ, Deck J, Dröge G, Döring M, Field D, Kottmann R, Ma J, Mori H, Morrison N, Sterk P and others. Meeting Report: Hackathon-Workshop on Darwin Core and MIxS Standards Alignment (February 2012). Standards in Genomic Sciences 2012;7(1). [PMID:23451295](http://www.ncbi.nlm.nih.gov/pubmed/23451295/) [doi:10.4056/sigs.3166513](http://dx.doi.org/10.4056/sigs.3166513)
23. Deck, J, et al. Clarifying Concepts and Terms in Biodiversity Informatics, Standards in Genomic Sciences, submitted March 2013
24. Grenon P, Smith B. SNAP and SPAN: Towards Dynamic Spatial Ontology. Spatial Cognition and Computation 2004;4(1):69-103. [doi:10.1207/s15427633scc0401\_5](http://dx.doi.org/10.1207/s15427633scc0401_5)
25. Arp R, Smith B. Function, Role, and Disposition in Basic Formal Ontology. In: The 11th Annual Bio-Ontologies Meeting, 2008, Toronto, Canada. p 1-4. [doi:10101/npre.2008.1941.1](http://dx.doi.org/10101/npre.2008.1941.1)
26. The Gene Ontology Consortium. The Gene Ontology: enhancements for 2011. Nucleic Acids Research 2012;40(D1):D559-D564. [PMID:22102568](http://www.ncbi.nlm.nih.gov/pubmed/22102568) [doi:10.1093/nar/gkr1028](http://dx.doi.org/10.1093/nar/gkr1028)
27. The Gene Ontology Consortium. Creating the Gene Ontology Resource: Design and Implementation. Genome Research 2001;11(8):1425-1433. [PMID:11483584](http://www.ncbi.nlm.nih.gov/pubmed/11483584/) [doi:10.1101/gr.180801](http://dx.doi.org/10.1101/gr.180801)
28. Eilbeck K, Lewis SE. Sequence Ontology Annotation Guide. Comparative and Functional Genomics 2004;5(8):642-647. [PMID:18629179](http://www.ncbi.nlm.nih.gov/pubmed/18629179/) [doi:10.1002/cfg.446](http://dx.doi.org/10.1002/cfg.446)
29. Brinkman R, Courtot M, Derom D, Fostel J, He Y, Lord P, Malone J, Parkinson H, Peters B, Rocca-Serra P and others. Modeling biomedical experimental processes with OBI. Journal of Biomedical Semantics 2010;1(Suppl 1):S7. [PMID:20626927](http://www.ncbi.nlm.nih.gov/pubmed/20626927) [doi:10.1186/2041-1480-1-S1-S7](http://dx.doi.org/10.1186/2041-1480-1-S1-S7)
30. Haendel MA, Neuhaus F, Osumi-Sutherland D, Mabee PM, Mejino JLV, Jr., Mungall CJ, Smith B. CARO - The Common Anatomy Reference Ontology. In: Burger A, Davidson D, Baldock R, editors. Anatomy Ontologies for Bioinformatics. Volume 6, Computational Biology: Springer London; 2008. p 327-349. [doi:10.1007/978-1-84628-885-2\_16](http://dx.doi.org/10.1007/978-1-84628-885-2_16)
31. Giglio MG, Collmer CW, Lomax J, Ireland A. Applying the Gene Ontology in microbial annotation. Trends in microbiology 2009;17(7):262-268. [PMID:19577473](http://www.ncbi.nlm.nih.gov/pubmed/19577473) [doi:10.1016/j.tim.2009.04.003](http://dx.doi.org/10.1016/j.tim.2009.04.003)
32. Cowell LG, Smith B. Infectious Disease Ontology. In: Sintchenko V, editor. Infectious Disease Informatics: Springer-Verlag New York; 2010. p 373-395. [doi:10.1007/978-1-4419-1327-2\_19](http://dx.doi.org/10.1007/978-1-4419-1327-2_19)
33. Topalis P, Mitraka E, Bujila I, Deligianni E, Dialynas E, Siden-Kiamos I, Troye-Blomberg M, Louis C. IDOMAL: an ontology for malaria. Malaria Journal 2010;9(1):230. [PMID:20698959](http://www.ncbi.nlm.nih.gov/pubmed/20698959) [doi:10.1186/1475-2875-9-230](http://dx.doi.org/10.1186/1475-2875-9-230)
34. Gkoutos GV, Schofield PN, Hoehndorf R. Chapter Four - The Neurobehavior Ontology: An Ontology for Annotation and Integration of Behavior and Behavioral Phenotypes. In: Chesler EJ, Haendel MA, editors. International Review of Neurobiology. Volume 103: Elsevier; 2012. p 69-87. [PMID:23195121](http://www.ncbi.nlm.nih.gov/pubmed/23195121) [doi:10.1016/B978-0-12-388408-4.00004-6](http://dx.doi.org/10.1016/B978-0-12-388408-4.00004-6)
35. Mungall CJ, Gkoutos GV, Smith CL, Haendel MA, Lewis SE, Ashburner M. Integrating phenotype ontologies across multiple species. Genome Biology 2010;11(1):R2. [PMID:20064205](http://www.ncbi.nlm.nih.gov/pubmed/20064205) [doi:10.1186/gb-2010-11-1-r2](http://dx.doi.org/10.1186/gb-2010-11-1-r2)
36. Eilbeck K, Bada M, Efforts toward a more consistent and interoperable Sequence Ontology. Proceedings of the 3rd International Conference on Biomedical Ontology (ICBO 2012), KR-MED Series, Graz, Austria, July 21-25, 2012 R. Cornett and R. Stevens, Editors. Available online at <http://ceur-ws.org/Vol-897/session3-paper13.pdf>
37. Bizer C, Heath T, Berners-Lee T, Linked data—The story so far. International Journal on Semantic Web and Information Systems 2009;5(3):1–22. doi:10.4018/jswis.2009081901
38. Guralnick RP, Wieczorek J, Beaman R, Hijmans RJ, The BioGeomancer Working Group. BioGeomancer: Automated Georeferencing to Map the World's Biodiversity Data. PLoS Biol 2006;4(11):e381. [PMID:17105347](http://www.ncbi.nlm.nih.gov/pubmed/17105347) [doi:10.1371/journal.pbio.0040381](http://dx.doi.org/10.1371/journal.pbio.0040381)
39. Duckworth WD, Genoways HH, Rose CL. Preserving natural science collections: Chronicle of our environmental heritage. National Institute for the Conservation of Cultural Property (Washington, DC); 1993.
40. Ariño AH. Approaches to estimating the universe of natural history collections data. Biodiversity Informatics 2010;7(2):81–92.

**Figure Legends**

**Figure 1:** Sampling schemes in ecological studies are often nested and may include plot; subplot or transect within plot; individual within plot, subplot, or transect; organ (e.g., leaf) within individual; tissue within organ; and DNA or mineral (e.g., C or N) within tissue. DNA extracted from a leaf of a tree that is present in a sub-plot may be characterized by environmental features of the plot.

**Figure 2:** Core terms of the Biological Collections Ontology (BCO) and their relationships to the Basic Formal Ontology (BFO) and Ontology for Biomedical Investigations (OBI). BCO terms that are subclasses of BFO:*continuant* are shown in **(A)**. For example, BCO:*material sample* is a subclass of BFO:*material entity* and hasRole BFO:*material sample role* (which is a BFO:*role*). BCO terms that are subclasses of BFO:*occurrent* are shown in **(B)**. For example, BFO:*material sampling process* is a subclass of OBI:*planned process*, which is a subclass of BFO:*process*. The ontology graphs in A and B are connected through the BFO root term *entity*, but are shown separately due to space constraints.

**Figure 3:** Linking samples and derivatives from the Moorea Biocode project: **A.** Biodiversity data is collected at many different levels that may be connect to one another in biologically meaningful ways, such as an Essig Museum specimen collected as part of a Biocode bioinventory event, a Smithsonian tissue or CAMERA metagenomic gut sample collected from that specimen, or DNA extracted from the tissue or metagenomic sample. **B.** Ontology classes (ovals) from multiple, coordinated ontologies can be used to describe the samples and processes represented in A. These terms are rooted in the same upper-level ontology (the BFO) and connected by ontological relations such as *subClass* of (solid black arrows), *has specified output* (blue arrows), and *has specified input* (green arrows). The relationship between the sequence ontology (SO) term *region* and the BFO is not yet explicit in the SO file, and so is shown as a dashed line. **C.** Annotation of the workflow shown in A with ontology terms from B is represented graphically in C. Ontology classes are shown as ovals (as in B), instances are shown as rectangles, and relations between instances or classes as arrows (as in B). Color-coding links instances in C to classes in B. This figure shows how, for example, TaxonID Y resulting from the BLAST identification process on Genbank sequence Y can be linked back to the original Moorea Biocode sampling event. The relation between the island of Moorea and the EnvO *biome* term *oceanic island* is shown as a dashed gray arrow, because an ontological relation between a geographic location or a material entity and EnvO classes has not yet been specified.

1. The application of taxon names is more complex than it first seems, and requires its own separate treatment to do it justice. However, given the critical importance of names associated with biological collection objects, we briefly summarize key issues here. In short, the processes used to generate taxon names are diverse and often contradictory, resulting in a well-known (yet still difficult) need to reference names to a common standard. Despite continued work and products (e.g., TaxMeOn (<http://schema.onki.fi/taxmeon/>) and the Euler project (<http://www.nsf.gov/awardsearch/showAward?AWD_ID=1118088>) there is no single community solution to this crucial issue. [↑](#footnote-ref-2)
2. Note that EnvO grows through a phase of “greedy term capture” followed by a curation phase. In the former, contributors work with either EnvO’s core development team or the wider consortium to specify their knowledge of environments with which they are familiar. Contributed terms, their positioning in EnvO’s hierarchies, and their working definitions are then curated. In this phase, and wherever possible, logical definitions are created, redundancies removed, and cross-references added to enhance the semantic usefulness of new content. Users should thus ensure that the terms they employ for annotation and analysis are appropriate to their needs. [↑](#footnote-ref-3)