9-1-2023

The Ontology of Biological Attributes (OBA)-computational traits for the life sciences.

Ray Stefancsik

James P Balhoff

Meghan A Balk

Robyn L Ball

Susan M. Bello

See next page for additional authors

Authors

Ray Stefancsik, James P Balhoff, Meghan A Balk, Robyn L Ball, Susan M. Bello, Anita R Caron, Elissa J Chesler, Vinicius de Souza, Sarah Gehrke, Melissa Haendel, Laura W Harris, Nomi L Harris, Arwa Ibrahim, Sebastian Koehler, Nicolas Matentzoglu, Julie A McMurry, Christopher J Mungall, Monica C Munoz-Torres, Tim Putman, Peter N Robinson, Damian Smedley, Elliot Sollis, Anne E Thessen, Nicole Vasilevsky, David O Walton, and David Osumi-Sutherland

Mammalian Genome (2023) 34:364–378
https://doi.org/10.1007/s00335-023-09992-1

The Ontology of Biological Attributes (OBA)—computational traits for the life sciences

Ray Stefancsik[1] · James P. Balhoff[2] · Meghan A. Balk[3] · Robyn L. Ball[4] · Susan M. Bello[4] · Anita R. Caron[1] · Elissa J. Chesler[4] · Vinicius de Souza[1] · Sarah Gehrke[5] · Melissa Haendel[5] · Laura W. Harris[1] · Nomi L. Harris[6] · Arwa Ibrahim[1] · Sebastian Koehler[7] · Nicolas Matentzoglu[8] · Julie A. McMurry[5] · Christopher J. Mungall[6] · Monica C. Munoz-Torres[5] · Tim Putman[5] · Peter Robinson[4] · Damian Smedley[9] · Elliot Sollis[1] · Anne E. Thessen[5] · Nicole Vasilevsky[10] · David O. Walton[4] · David Osumi-Sutherland[1]

Received: 27 January 2023 / Accepted: 6 April 2023 / Published online: 19 April 2023

Abstract
Existing phenotype ontologies were originally developed to represent phenotypes that manifest as a character state in relation to a wild-type or other reference. However, these do not include the phenotypic trait or attribute categories required for the annotation of genome-wide association studies (GWAS), Quantitative Trait Loci (QTL) mappings or any population-focussed measurable trait data. The integration of trait and biological attribute information with an ever increasing body of chemical, environmental and biological data greatly facilitates computational analyses and it is also highly relevant to biomedical and clinical applications. The Ontology of Biological Attributes (OBA) is a formalised, species-independent collection of interoperable phenotypic trait categories that is intended to fulfil a data integration role. OBA is a standardised representational framework for observable attributes that are characteristics of biological entities, organisms, or parts of organisms. OBA has a modular design which provides several benefits for users and data integrators, including an automated and meaningful classification of trait terms computed on the basis of logical inferences drawn from domain-specific ontologies for cells, anatomical and other relevant entities. The logical axioms in OBA also provide a previously missing bridge that can computationally link Mendelian phenotypes with GWAS and quantitative traits. The term components in OBA provide semantic links and enable knowledge and data integration across specialised research community boundaries, thereby breaking silos.

Introduction

Animal models have greatly contributed to the progress of genomics research. In addition to mutant strains identified by traditional phenotypic selection and breeding methods, genome engineering in model organisms allow the generation of transgenic lines and targeted mutants using homologous recombination or CRISPR-Cas9 technology (Bello et al. 2021; Hsu et al. 2014; Clark et al. 2020). Collectively, these technologies allow researchers worldwide to generate a large body of genetic data using mouse and other model organisms, and the resulting data is made available in several biomedical databases curated by experts (Alliance of Genome Resources Consortium 2022; Blake et al. 2021;

✉ Ray Stefancsik
ray@ebi.ac.uk

1 European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridgeshire CB10 1SD, UK

2 Renaissance Computing Institute, University of North Carolina, Chapel Hill, NC 27517, USA

3 Natural History Museum, University of Oslo, Oslo, Norway

4 The Jackson Laboratory, Bar Harbor, ME 04609, USA

5 Anschutz Medical Campus, University of Colorado, Aurora, CO 80045, USA

6 Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

7 Ada Health GmbH, Berlin, Germany

8 Semanticly, Athens, Greece

9 William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London EC1M 6BQ, UK

10 Data Collaboration Center, Critical Path Institute, Tucson, AZ 85718, USA

Kaldunski et al. 2022; Groza et al. 2023). Currently there are more than 800 biological databases collecting genotype, phenotype and variation data from a wide range of organisms (Ma et al. 2022). The valuable knowledge in these databases on variants, phenotypes and gene function is highly relevant to human and veterinary medicine, agriculture, evolutionary biology, ecology and comparative genomics in general.

Technological advances in next-generation DNA sequencing also yield an ever increasing number of new genomic variants with unknown functional significance across the tree of life (Stephens et al. 2015; Cantelli et al. 2022). The identification of the phenotypically and clinically relevant subset of the new DNA variants in both human and veterinary medicine, as well the characterisation of the mechanisms of how these variations exert their phenotypic effects, pose serious challenges that cannot be met successfully without advancements in data integration and computational tools (Thessen et al. 2020). A standardised and computationally amenable representation of traits is critical for many biomedical and agricultural use cases involving DNA variants, from genome-wide association studies (GWAS) to Quantitative Trait Loci (QTL) mappings (Rehm 2021; Sollis et al. 2023; Bogue et al. 2023; Pathak and Kim 2022; Moses et al. 2018). Currently, the lack of consistent computational modelling and annotation of traits from various data sources restricts their interoperability and hinders not only genetic mechanisms of discovery for medicine, but also agriculture, biodiversity and evolutionary biology.

Ontologies provide standardised sets of concepts (terms) that are understandable by human users and also allow for logical inference, computational reasoning and sophisticated data queries. There are several phenotype ontologies that differ in their scope of specialisation or focus on certain taxonomic groups. For example, the Mammalian Phenotype Ontology (MP) (Smith and Eppig 2009) and the Human Phenotype Ontology (HPO) (Köhler et al. 2021) have different taxonomic focusses to categorise phenotypes of primarily Mendelian-type inheritance. Each of these ontologies is used to annotate genotypes, where the annotations represent phenotypic states that deviate from a reference, which is usually the wild-type or typical phenotype for the species and population of focus. The phenotypic deviation or abnormality is always represented in the logical axioms in these phenotype ontologies. This is in contrast to trait ontologies, where the logical axioms define generic attributes without reference to any specific phenotypic alterations or states. For example, a "blood glucose amount" can manifest in a "Hyperglycemia" phenotype, where the former manifests in an "increased amount" phenotypic state with an "abnormal" quality component in the logical equivalence axioms. This is a fundamental difference between modelling traits and phenotypes ontologically.

OBA is a standardised, representational framework for observable attributes that are characteristics of organisms, or parts of organisms. For example, the attribute "trochanter size" (OBA:0002360) is a characteristic of the anatomical entity "trochanter" (UBERON:0000980); and "blood glucose amount" (OBA:VT0000188) is a characteristic of glucose (CHEBI:17234) in the blood (UBERON:0000178). This way of defining attributes, called the Entity-Quality (EQ) pattern, is used by many biomedical ontologies, including the Plant Trait Ontology (TO) (Cooper et al. 2018) for defining attributes of plants such as "petal length" (TO:0002605), the Environment Ontology (ENVO) (Buttigieg et al. 2016) for defining attributes of environmental materials, such as "soil pH" (ENVO:09200010) and the Human Phenotype Ontology (HPO) for defining phenotypic abnormalities such as "Abnormal telomere morphology" (HP:0031412). The same EQ pattern has also been employed in data annotation using a post-compositional approach—combining an entity term and a quality term within an annotation, rather than creating a separately defined trait term—to describe both phenotypic abnormalities (e.g. in zebrafish) (Bradford et al. 2011) as well as natural evolutionary variation in the Phenoscape Knowledgebase (Mabee et al. 2012; Dahdul et al. 2010). The initial design of OBA was significantly inspired by work from the creators of the Plant Trait Ontology.

The majority of attributes in OBA are logically defined using the Web Ontology Language (OWL). These logical definitions use terms from relevant reference ontologies, such as Uberon (Mungall et al. 2012) or ChEBI (Hastings et al. 2016). With the exception of a small number of high-level concepts, most of the classification in OBA is automatically computed on the basis of the classifications of the various reference ontologies, using an automated reasoner. The advantage of this approach is twofold: firstly, we do not have to manually classify any concepts, which drops the cost of curating the classification significantly whilst increasing its completeness. Secondly, the numerous links to reference ontologies can be exploited for a wide variety of applications, including querying (e.g. select all data where the morphology of a part of the renal system is affected), knowledge graph integration (e.g. automatic linking to phenotypic abnormalities from widely used ontologies such as HPO or MP) and knowledge inference (e.g. inferring missing data from logical implications) (Dececchi et al. 2015). A rich logical axiomatisation based on design patterns is necessary to ensure interoperability with existing phenotype ontologies and other data types, such as anatomical, chemical and biological pathway data. Existing ontologies such as the Vertebrate Trait ontology (VT) (Park et al. 2013) and the Experimental Factor Ontology (EFO) (Malone et al. 2010) are widely used to annotate traits, but do not contain such axiomatisation.

In this paper, we introduce OBA, an ontology and logical framework for representing biological attributes. We show how we use the Entity-Quality modelling framework to automatically classify attributes reliably using reference ontologies. Additionally, we demonstrate how OBA can be used to automatically integrate data from other widely used phenotype ontologies, thereby breaking silos.

Methods

Logical framework

As ontologies grow in size, they become increasingly hard to maintain. Phenotype (and trait) ontologies are inherently polyhierarchical, as they combine a variety of interwoven external classifications, such as attributes and biological entities. This makes it hard to ensure that the classification is complete by manual curation (no subclass axioms are missing), and that the existing classification is consistent with other ontologies (for example, "head size" should not be a parent of "eye size"). Instead of relying on manual classification of biological attributes in OBA, we use logical definitions and automated reasoning to compute the hierarchical classification. OBA is represented in the Web Ontology Language (OWL), a knowledge representation formalism based on Description Logics, a fragment of First Order Logic. It is fully aligned with the Core Ontology for Biology and Biomedicine (COB) (COB: An experimental ontology) because all concepts in OBA are, implicitly, children of "characteristic" (PATO:0000001), which itself is part of COB. However, we currently do not import COB directly (though this is planned as future work).

The Entity-Quality (EQ) pattern (Mungall et al. 2010; Washington et al. 2009) is widely used for representing traits and phenotypes in ontologies such as the Human (Köhler et al. 2021), Mammalian (Smith and Eppig 2009) and Xenopus (Fisher et al. 2022) phenotype ontologies. There are a number of variants of this pattern, but at its core, a phenotypic quality (Q, which is currently more frequently referred to as a "characteristic" rather than "quality") such as "height", "mass" or "amount", usually from the Phenotype And Trait Ontology (PATO) (Gkoutos et al. 2005), is combined with an entity (E), such as an anatomical or chemical entity, to form the concept of a "biological attribute", sometimes referred to as a "trait" (see Fig. 1). For example, "lysine in blood amount" (OBA:2020005) is composed of the PATO class of "amount" (PATO:0000070), lysine (CHEBI:25094) and blood (UBERON:0000178). PATO defines basic categories of phenotypic qualities (attributes or characteristics) and it can be used for quantitative trait or Mendelian phenotype annotation (Gkoutos et al. 2018).

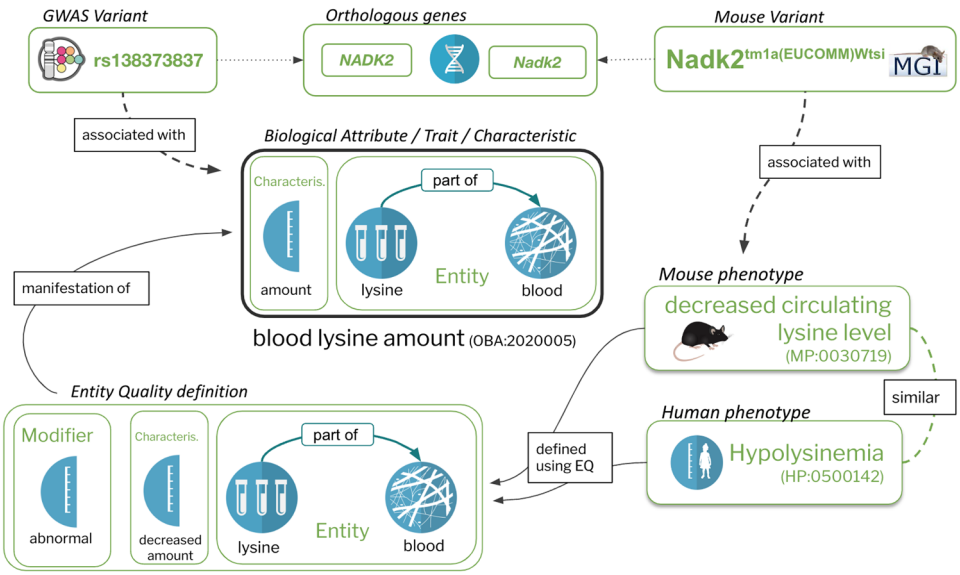


Fig. 1 The Entity-Quality model enables composing biological attributes in a way that is compatible with the logical definitions of widely used ontologies such as the MP and HPO which are used to document phenotypes associated with diseases or genes. On the right is a specific example of a human phenotype term, "Hypolysinemia" (HP:0500142), which means a lower than normal amount of lysine in the blood. The EQ (phenotypic effect) on the left is not only used to logically define Hypolysinemia, but also the mouse phenotype "decreased circulating lysine level" (MP:0030719). This ensures that an automated reasoner can compute the appropriate relationship between the two (in this case equivalence), as well as to the specific biological attribute they are concrete manifestations of ("blood lysine amount"). Representing phenotype and phenotypic attributes this way enables the grouping of quantitative variant data (e.g. GWAS) and qualitative variant data (e.g. MGI)

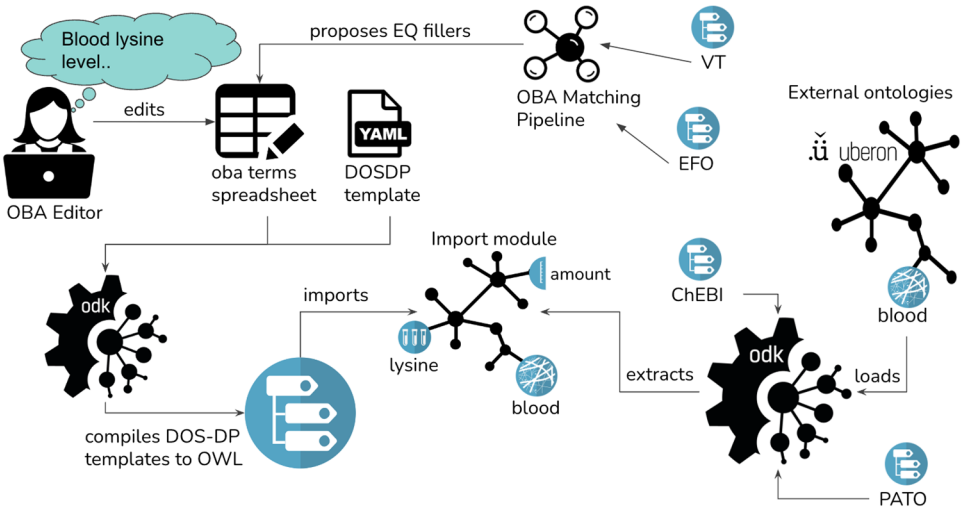
PATO is species-neutral in its scope but does not provide relationships to the biological entities whose phenotypic qualities it is meant to describe (Washington et al. 2009). Using EQ logical definitions in OWL enables us to use automated reasoners to automatically classify our traits: if, for example, lysine is an "amino acid" according to ChEBI, there is no need to remember to manually classify "lysine in blood amount" under "amino acid in blood amount"—the reasoner will do this for us based on the classification in ChEBI. A second feature of such axiomatisation is that it can be used for powerful logical querying using OWL DL Queries (Grau et al. 2008) and SPARQL (Detwiler et al. 2008). This enables us to group data in ways that are not easily available in traditional databases. For example, it allows us to query for data related to morphology of a tissue that is considered part of the cardiovascular system—even if no such term exists in OBA. A query capturing this can be found in the supplemental materials (S2, Supplementary materials).

Ontologies of phenotypic abnormalities such as HPO, MP, XPO (Fisher et al. 2022) and ZP make extensive use of the EQ pattern, but are primarily used to capture phenotypic effects compared to some reference (usually "wild type") rather than unqualified biological attributes as in OBA. For example, "decreased circulating lysine level" (MP:0030718) in the Mammalian Phenotype Ontology is defined as "abnormal(ly)" (PATO:0000460) "decreased amount" (PATO:0000470) of "lysine" in the "blood" (Fig. 1). Since both the biological attributes in OBA and the phenotypic effects in MP are represented using the Web Ontology Language (OWL), we can use an automated reasoner, such as ELK (Kazakov et al. 2014), to automatically compute links between the two. Other examples of links between OBA attributes and phenotypic effects: head circumference (OBA:VT0000047) has "Decreased head circumference" (HP:0040195), "Microcephaly" (HP:0000252) and "Progressive microcephaly" (HP:0000253) as manifestations; "brain ventricle size" (OBA:0002294) has manifestation "Ventriculomegaly" (HP:0002119).

Template-based ontology curation with DOS-DP

Ontologies, especially those with logically rich axiomatisation, enable powerful services such as automated reasoning, classification and logical querying, but logical modelling is difficult (Slater et al. 2020) and appropriate expertise is scarce. A popular approach to deal with this problem is to use design patterns and templating systems for logical axioms (Osumi-Sutherland et al. 2017). This allows for decoupling the curation of reference terms used for logical definitions from their exact axiomatic pattern. The central idea is to employ a small number of axiom templates (which implement design patterns such as the EQ model described above) that can be created and maintained by logic experts, and have content curators focus on the selection of appropriate filler terms (e.g. terms from Uberon to define anatomical attributes). There are a number of available approaches, but many Open Biological and Biomedical Ontologies (OBO) Foundry (Jackson 2021) ontologies use Dead Simple Ontology Design Patterns (DOS-DP) (Osumi-Sutherland et al. 2017), a system that allows capturing a logical model (design pattern) in a specific YAML file (YAML is a structured, yet human-readable file format; the OBO Foundry is a community-driven organisation which promotes the standardisation of metadata and logical patterns in the Biological and Biomedical Ontology community). This YAML file is maintained separately from the actual biological attributes, which are maintained in tables (TSV files). For example, the "entity attribute" template (see Fig. 3), the most basic of all OBA design patterns, has two filler terms, the entity (e.g. a chemical, or an anatomical entity) and the attribute (a characteristic from PATO, such as "amount") and defines how a new biological attribute in OBA following that pattern should be converted to OWL (amongst other aspects, it describes how the logical equivalent class definition should be instantiated). Curators simply add a row to a spreadsheet with the OBA identifier and the two fillers (see Fig. 3). The identifier scheme used for new OBA terms corresponds to the standard recommendation (OBO foundry). All identifiers are represented as a globally unique, persistent and resolvable identifier (GUPRI) (Le Franc et al. 2020), otherwise known simply as "persistent URL" (PURL), starting with the http://purl.obolibrary.org/obo/OBA_ URI prefix, followed by an numeric identifier. GUPRIs are essential to the vision of global data integration, and provide a stable way to refer to domain concepts such as the biological attributes discussed here. Some identifiers are prefixed with the literal "VT" to indicate that they are sourced from VT. A specialised toolkit (DOS-DP tools) (dosdp-tools: Utility for working with DOSDP design patterns and OWL ontologies) then translates the spreadsheet into OWL axioms using the template file.

OBA currently uses ten DOS-DP term templates for different trait patterns; see Table S1 (supplemental materials). These were selected because they cover the majority of anatomical, chemical level and cellular attributes which are central for the integration of genomics data. By far the majority of biological (especially anatomical) attribute terms in OBA can be represented using a basic entity-attribute pattern (e.g. "head size"). All templates can also be found online (src, patterns, dosdp-patterns at master · obophenotype, bio-attribute-ontology). In addition to ensuring a consistent axiomatisation of the ontology across thousands of terms (a general advantage of template systems, not just DOS-DP), one major advantage of using DOS-DPs as a framework for managing OWL ontologies is their generative capabilities.




```yaml
def:
  text: "The %s of a %s."
  vars:
    - attribute
    - entity
  annotations:
    - annotationProperty: xref
      text: "AUTO:patterns/patterns/entity_attribute"

exact_synonym:
  value: exact_synonyms

xref:
  value: xrefs

equivalentTo:
  text: "%s and 'characteristic_of' some %s"
  vars:
    - attribute
    - entity
```

entity-attribute.yaml

entity-attribute.tsv

defined_class	defined_class_name	entity	entity_name	attribute	attribute_name
OBA:0000166	tracheal tube diameter	UBERON:0000117	respiratory tube	PATO:0001334	diameter
OBA:0000167	tracheal tube length	UBERON:0000117	respiratory tube	PATO:0000122	length
OBA:0000168	tracheal tube size	UBERON:0000117	respiratory tube	PATO:0000117	size
OBA:1000031	neck circumference	UBERON:0000974	neck	PATO:0001648	circumference
OBA:1000045	iris color	UBERON:0001769	iris	PATO:0000014	color
OBA:1000072	coat color pattern	UBERON:0010166	coat of hair	PATO:0000019	color pattern
OBA:1000073	coat spatial pattern	UBERON:0010166	coat of hair	PATO:0000060	spatial pattern
OBA:VT0002641	erythrocyte shape trait	CL:0000232	erythrocyte	PATO:0000052	shape
OBA:1000006	cilium length	GO:0005929	cilium	PATO:0000122	length
OBA:1000018	nucleus shape	GO:0005634	nucleus	PATO:0000052	shape
OBA:VT0000222	neutrophil quantity	CL:0000775	neutrophil	PATO:0000070	amount
OBA:VT0001586	erythrocyte quantity	CL:0000232	erythrocyte	PATO:0000070	amount
OBA:0000036	body fluid levels	UBERON:0006314	bodily fluid	PATO:0000918	volume
OBA:0000006	Malpighian tubule diameter	UBERON:0001054	Malpighian tubule	PATO:0001334	diameter
OBA:0000007	Malpighian tubule size	UBERON:0001054	Malpighian tubule	PATO:0000117	size
OBA:0000011	actin filament length	GO:0005884	actin filament	PATO:0000122	length

Fig. 2 Overview of the OBA Workflow. The OBA matching pipeline searches existing trait ontologies for new terms and proposes suitable EQ fillers. The OBA editors curate EQ fillers (new ones and the ones proposed by the matching pipeline). The ODK then compiles the curated terms into OWL and imports all the referenced terms (EQ fillers) from their respective external ontologies, e.g. Uberon, into a special import module

Fig. 3 DOS-DP template example. The fillers declared in the template above (attribute, entity) are mapped to the respective column names in the TSV file below. A specialised tool reads both files and generates the axioms specified by the template file

Not only can we dynamically generate labels, definitions and synonyms based on the filler terms provided, but we can also add contextual axioms which can be exploited for automated reasoning. In OBA, for example, we generate General Concept Inclusion (GCI) axioms which define how attributes are related mereologically (e.g. "ulna size" part of "forelimb skeleton size"). These axioms are defined as part of the DOS-DP design patterns.

Automated mapping pipeline for external sources

One of the central use cases for OBA is to provide additional structure to other generally weakly axiomatised ontologies, mainly VT and the Experimental Factor Ontology (EFO). To synchronise these vocabularies with OBA, we execute the following workflow:

1. Match: link external terms to OBA terms if they exist
2. Sync: identify external terms that do not exist in OBA, decompose them into their logical components (reference entities) and curate them as instances in our DOS-DP template pipeline
3. Compile: compile the new terms into OWL and integrate them into the ontology

We have built a custom pipeline (oba_alignment.ipynb at master · obophenotype, bio-attribute-ontology) that supports our curators in steps 1 and 2 (see Fig. 2). To that end, we implemented a matching process that works on the ontology labels and exact synonyms. After applying a series of normalisation steps (including the removal of stop words like "measurement" or "trait"), if a direct match between an external term and an OBA term can be identified, we present it as a candidate match to a curator. The curator just has to review and approve or reject the match. For step 2, we sequentially match all our reference ontologies (ChEBI, Uberon, PRO, GO, PATO) to the external term. For example, if an external term "lysine measurement" contains the term "lysine", we record that as a potential match for the "entity" column in the "entity-attribute" DOS-DP pattern (Fig. 3). Thus our curators are presented with a set of potential EQ-decompositions, which they proceed to either accept or reject. Mappings to external ontologies, as generated in steps 1 and 2, are documented using the Simple Standard for Sharing Ontological Mappings (SSSOM) (Matentzoglu 2022a) and shared as part of the OBA GitHub repository (https://github.com/obophenotype/bio-attribute-ontology). Note that in contrast to other synchronisation workflows such as those used by Uberon or the Mondo disease ontology (Vasilevsky et al. 2020), we do not import any curated information from external ontologies (synonyms, definitions, etc.) but rely entirely on automated templated processes.

OBA life cycle management

OBA has been a member of the OBO Foundry (Jackson et al. 2021) for more than seven years and has a team of 6 regular contributors. It is managed by members of the European Bioinformatics Institute (EBI) and the Monarch Initiative (Shefchek et al. 2020) using modern ontology workflows and curation practices. To manage our releases, quality control and external dependencies we use the Ontology Development Kit (ODK (Matentzoglu 2022b), version 1.3.2). The ODK provides mechanisms to version and publish OBA releases in a variety of serialisations (JSON, RDF/XML, OBO) and release file variants according to OBO Foundry practices, relying largely on the ROBOT tool (Jackson et al. 2019). It fully supports DOS-DP workflows which ensures a seamless integration of mostly TSV based curation into the general ontology life cycle. For example, terms that are used as fillers during the decomposition of biological attributes are automatically imported from their respective external ontologies. The ODK is also used for continuous integration testing. Whenever one of our curators makes a pull request on GitHub with changes to OBA, we automatically execute the DOS-DP pipeline, followed by a number of strict quality control checks. For example, these checks ensure that all terms added fall under the "biological attribute" root term, are unique (no other equivalent attribute exists) and are logically consistent. Lastly, the ODK imports relevant terms and axioms from our reference ontologies (e.g. Uberon, ChEBI, PATO), which ensures that OBA is fully consistent with their axiomatisation (see Fig. 2). To ensure consistency, we use the ELK reasoner, which is suitable for OWL 2 EL ontologies (see Results). OBA publishes a new version every 2–3 months, using the GitHub releases mechanism for versioning and dissemination.

Results

The Ontology of Biological Attributes (OBA) is published under the CC0-1.0 licence (public domain) and is in its 17th release (21 December 2022) (bio-attribute-ontology) at the time of writing this paper. The ontology is expressed using the OWL 2 EL profile of the Web Ontology Language (OWL) (Motik et al. 2009). Note that some imports use higher expressivity axioms (beyond OWL 2 EL), which means that there are corner cases where using an OWL 2 EL reasoner such as ELK (Kazakov et al. 2014) may be incomplete. (Note: all elements required by the Minimum Information for Reporting of an Ontology (MIRO) guidelines (Matentzoglu et al. 2018b) are reported here.)

OBA defines 7807 biological attributes, most of which have logical equivalence axioms (full logical definition).

OBA attributes have, on average, 1.54 parents (indicating a high degree of polyhierarchy) and one or more associated synonyms. Most attributes (69.6%) are anatomical, 12.4% are attributes of biological processes and 9.8% are cellular attributes (see Fig. 4). Anatomical attributes are defined with terms from the Uberon anatomy ontology. In total there are 302 CHEBI, 346 CL, 708 GO, 40 MONDO, 109 NBO, 139 PATO, 69 PRO, 4 SO and 2469 UBERON terms referred to by OBA ids. The latest version of OBA is available under the persistent URL http://purl.obolibrary.org/obo/oba.owl. The version referred to as part of this paper can be accessed by the versioned persistent URL http://purl.obolibrary.org/obo/oba/releases/2022-12-21/oba.owl

If an entity is obsoleted, the label is changed to have "obsolete" at the beginning, the metadata field "deprecated" is true, and all logical axioms are removed. When the entity is completely redefined, the metadata field "term replaced by" is added to indicate the substitute term.

The four main relationships used in OBA are "part of" (BFO:0000050) primarily to denote mereological links between anatomical entities, "characteristic of" (RO:0000052) to link a characteristic (e.g. "morphology") to a biological entity (e.g. "heart"), "characteristic of part of" (RO:0002314) to link a characteristic to a biological entity and all of its parts (e.g. we can use "characteristic of part of" to define a trait that applies to all parts of the cardiovascular system), and "subclass of" (rdfs:subClassOf) to classify biological attributes. In addition, the relationship has_role (RO:0000087) is used in cases where the definition of a biological attribute requires a reference to a chemical role, such as "serum metabolite" in "serum metabolite amount" (OBA:2050092).

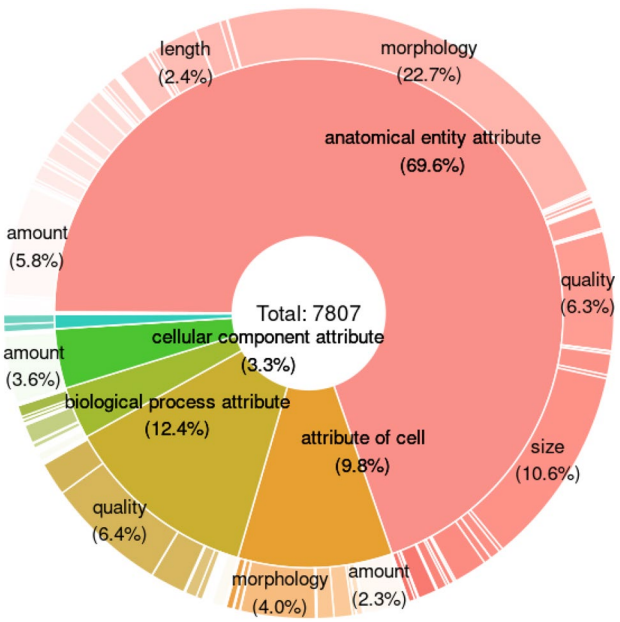


Fig. 4 Distribution of OBA attributes across categories and qualities

OBA coverage of biological attributes relevant for cross-species data integration

The template-based curation workflow ("Template-based ontology curation with DOS-DP" section) makes the process of adding new attributes highly scalable, as we do not need to worry about logical modelling. In the following we show (as an example) how to rapidly curate relevant biological attributes in OBA to cover the needs of the International Mouse Phenotyping Consortium (IMPC) database which captures information that includes the effect of gene knockouts on phenotype (Groza et al. 2023). IMPC uses 1102 phenotypic abnormality terms from the Mammalian Phenotype Ontology (MP). To ensure that we capture the relevant attribute terms for these, we first extract the fillers for the EQ logical definitions from MP using DOS-DP tools (dosdp-tools: Utility for working with DOSDP design patterns and OWL ontologies), and then transform the fillers into the respective OBA pattern. Since this mapping approach relies on the presence of EQ logical definitions (and only about half of MP terms have one), only 532 (~50%) of the IMPC phenotypes could be matched this way. A curator then manually assigns appropriate PATO characteristics (e.g. "amount" in cases where the quality of the phenotypic abnormality was "increased amount"). This process resulted in a total of 179 new trait terms added to OBA. Note that the remaining 50% of IMPC phenotypes need more extensive effort, in some cases manual curation. However, the mapping approach described in the automated mapping pipeline section can be employed to streamline the effort.

The Mouse Phenome Database (MPD) (Bogue et al. 2023) enables the integration of genomic and phenomic data by providing access to primary experimental data, well-documented data collection protocols and analysis tools. OBA terms currently cover 80.1% (5066 of 6325) of trait measures annotated in MPD via mappings to VT. We estimate that we will cover most of the remaining 20% by the end of 2023.

Identifying mouse traits reflective of human disease is critical to prioritise preclinical models of disease and aspects of complex disease. Prior to the development of OBA, researchers hoping to retrieve mouse trait measures reflective of human disease characteristics had to know specifically which mouse traits were associated with each disease, searching trait by trait to find a complete set. A workaround approach involves retrieval of disease terms (DOID) to vertebrate traits (VT) using a gene-centric mapping performed by retrieving Alliance of Genome Resources (AGR) (Alliance of Genome Resources Consortium 2022) annotated genes associated with each DOID and identifying Mammalian Phenotype (MP) terms to which their phenotypic alleles

were annotated in Mouse Genome Database (MGD) (Blake et al. 2021). These MP terms were used to retrieve mouse trait measures from the Mouse Phenome Database (MPD) and the VT terms to which the traits were also annotated. In an effort to validate the utility of OBA to retrieve relevant mouse data, we compared this gene-centric approach to the OBA's semantic mappings of disease terms from the Human Disease Ontology (Schriml et al. 2022) to vertebrate traits (VT). In total, OBA mapped 3033 of 3455 (87.8%) disease terms to VT terms with mouse trait measures in MPD. From the two approaches combined, we identified 4910 disease terms that had associations to VT terms. For 1348 disease terms, at least one VT term was mapped to the disease using each approach and 558 (41.4%) were associated with at least one shared VT term. In these cases, the mean overlap of VT terms per disease was 26.6% for OBA and 16.7% for the gene-centric approach.

Using OBA to group phenotypic abnormalities

To determine how well existing ontologies of phenotypic abnormality aggregate under OBA, we count the total number terms falling under any OBA class, and the total number of links between a phenotype ontology and OBA. For an accurate edge count, we follow the Ubergraph approach, which essentially converts an ontology to a knowledge graph with nodes and edges instead of axioms. We (1) merge the phenotype ontologies with OBA, then (2) materialise the relationships necessary for connecting OBA biological attributes with phenotypic abnormalities using a regular OWL 2 reasoner (ELK). Next, we (3) convert the resulting ontology to a knowledge graph using "relation graph" (Balhoff et al. 2022). Lastly, (4) we extract the OBA mappings from the knowledge graph using the SSSOM toolkit (Matentzoglu 2022a). The results can be found in Table 1. It is important to understand that no particular effort was made to cover all phenotype classes—as described in the section above, coverage can be rapidly increased by reusing the logical definitions. This experiment only illustrates the breadth of integration, not its depth: it includes classes that only link to very general attributes like "morphology anatomical entity".

Alignment with other trait ontologies

In contrast to the alignment with ontologies of phenotypic abnormalities as described in the previous section, alignment with most other trait vocabularies has to be performed using a semi-automated approach based on automated matching (see "Automated mapping" section above) and manual curation. To date, we have curated 2314 mappings to VT (version 12.5) and 150 mappings to EFO (version 3.14.0). 2,332 terms in OBA (which can be recognised by their ID, i.e. OBA:VT123 instead of OBA:123) have been derived from the VT ontology, i.e. OBA terms decompose and generalise them using the EQ pattern.

Table 1 Current degree of integration between OBA and existing phenotype ontologies

Ontology	# Links to OBA	# Classes under OBA
HPO	217,474	16,544
MP	187,405	13,620
ZP	117,023	39,373
XPO	38,567	20,340

A handful of other ontologies of attributes use the same EQ system to define attribute classes. For example, the Plant Trait Ontology (TO) has 1144 classes that classify under OBA attributes, and the Plant Phenology Ontology has 60 such classes.

How to access OBA and how to contribute to it

The EMBL-EBI Ontology Lookup Service (OLS) (Jupp et al. 2015) and Ontobee (Ong et al. 2017) are platforms from which one can find or browse OBA terms manually. There is also an OBA GitHub repository for those who wish to contribute to OBA, view documentation or download public releases and source files.

Users can explore OBA by entering free text into the search box on OLS or by using unique, permanent OBA identifiers. It is also possible to browse terms in the ontology hierarchy tree view or the interactive graph layout which displays colour-coded term relations. Search results return OBA terms, textual and logical definitions in addition to terms dynamically imported from other ontologies. Users can also query for OBA terms using the linked ontology server, Ontobee (Table 2). A complete list of terms can be downloaded in ".xlxs" or ".tsv" formats from Ontobee's home page. The OBA ".obo" or ".owl" ontology files can be viewed in an ontology editor such as Protégé, where users can browse terms and construct DL queries (Musen and Protégé Team 2015).

OBA welcomes contributions or suggestions for improvements from the research community. Contributions, suggestions or bug reports can be initiated via the OBA issue tracker on GitHub (Table 2).

Programmatic access to OBA

OBA is distributed in RDF/OWL, OBO Format and OBO Graph JSON format, so any programming library that is capable of reading these formats can be used to explore OBA. For data science use cases we recommend the use

Table 2 Different ways to access OBA

Link	Name	Note
https://www.ebi.ac.uk/ols/ontologies/oba	EMBL-EBI Ontology Lookup Service (OLS)	Find or browse OBA terms manually
https://github.com/obophenotype/bio-attribute-ontology	OBA GitHub repository	To contribute to OBA, read documentation, or download OBA public releases and source files
https://www.ebi.ac.uk/ols/docs/api	OLS API	Query OBA terms programmatically
https://ubergraph.apps.renci.org/sparql	ubergraph SPARQL endpoint	Query OBA terms programmatically
https://api.triplydb.com/s/DwuipbH9o	Example query using ubergraph SPARQL endpoint	Query OBA terms programmatically
https://ontobee.org/ontology/OBA	Ontobee	Query OBA terms manually or programmatically
https://github.com/obophenotype/bio-attribute-ontology/issues	OBA issue tracker	Contribute bug reports or suggestions to OBA
https://github.com/INCATools/ontology-access-kit/blob/main/notebooks/Monarch/OBA-Tutorial.ipynb	Accessing OBA using the Ontology Access Kit (OAK)	We provide a Jupyter notebook showing examples of querying OBA using OAK

of the Ontology Access Kit (OAK) (ontology-access-kit: Ontology Access Kit: A python library and command line application for working with ontologies.(Github), which provides both Python bindings and a command line interface. Additionally, there are several ways to query OBA terms using public APIs: the OLS API, Ubergraph (Balhoff et al. 2022) and Ontobee SPARQL endpoints (Table 2).

Use cases

OBA is used across a wide range of biological domains and processes, including genomics and drug discovery. In this section, we list some examples of its use.

The Gene Ontology (Gene Ontology Consortium 2021) is using OBA for axiomatising their "regulation of characteristic" branch (~ 100 terms), which describes biological processes that qualitatively or quantitatively modulate a biological attribute. For example, biological processes "regulation of lysosomal lumen pH" (GO:0035751) and "lysosomal lumen pH elevation" both regulate the biological attribute (trait) "lysosomal lumen pH" (OBA:0000091). OBA trait terms imported into EFO can facilitate computational drug target identification via the Open Targets Platform (Ochoa et al. 2023). For example, OBA, in tandem with other ontologies, has proved useful for computational drug target identification in a study of drug-induced adverse events in animal models (Giblin et al. 2021). In a study that identified associations between drug-induced preclinical and clinical adverse events in animal models and humans, a number of ontologies including the Human Disease Ontology (DOID), MP, HPO, EFO and OBA amongst others, were used to map adverse events terms. For example, the effect "Glucose urine present" was mapped to "urine glucose amount" (OBA:VT0001758) and "Photosensitivity allergic reaction" was mapped to "zone of skin photosensitivity" (OBA:0003620). A total of 15 OBA terms were used in these mappings. Using these mappings, the OpenTargets database was then queried to extract the genes associated with diseases encoded by the EFO and DOID ontologies (Giblin et al. 2021).

Another important OBA use case is the online community resource Functional Trait Resource for Environmental Studies (FuTRES) (Balk et al. 2022). It contains an application ontology, FuTRES Ontology of Vertebrate Traits (FOVT) (https://obofoundry.org/ontology/fovt.html), developed to standardise measurable trait terms in vertebrates. The FOVT currently has 390 trait terms (https://futres-data-interface.netlify.app/), 65 of which are from OBA and 325 of which will be eventually incorporated into OBA (Balk et al. 2022). By standardising terms, researchers spend less time wrangling data as the harmonised terms enable interoperable data. FOVT follows and helps develop patterns developed by OBA. Using patterns helps eliminate human errors and makes for easier on-boarding of new ontology curators. FOVT also takes advantage of OBA annotation property, "measured in taxon" (OBA:2050187) and "not measured in taxon" (OBA:2050188), with 244 and 16 assertions, respectively, to increase findability of trait terms of interest to researchers studying particular groups of organisms.

OBA terms are also used in the fields of agriculture, nutrition, zoology and biodiversity. AgBioData member databases take advantage of the species-neutral nature of OBA terms to integrate agriculturally important animal and plant traits with genomics and genetics data (Harper et al.2018). The Compositional Dietary Nutrition Ontology (CDNO) uses OBA to link nutritional components found in food to their human dietary roles which include traits. This allows the integration of nutritional components like

"concentration of calcium" (CDNO:0200138) with associated traits, for example, bone strength (OBA:VT0001542) (Andrés-Hernández et al. 2022). OBA trait classes have been used for the annotation of domestic guinea pig electrophysiology data (Farrell and Bengtson 2019). The Encyclopedia of Life (EOL) TraitBank takes advantage of the well-axiomatised OBA terms to infer traits in biodiversity data and to improve their search functionality (Horn 2016; Parr et al. 2016). For example, a user looking for a body size measurement would not have to do separate searches for all the different ways body size is measured in different taxonomic groups, e.g. body length, snout vent length and fork length. The semantic features of OBA can contribute to improved named entity recognition performance when incorporated in a natural language processing (NLP) framework for biodiversity literature curation (Batista-Navarro et al. 2016). Additionally, OBA can be used to link traits and phenotypes to environments (Thessen et al. 2015). This is of particular interest in the crop science community, where researchers are working to identify specific regions of the genome that control complex traits, such as drought resistance.

Having a rich set of links between biological attributes (traits) and phenotypic abnormalities also enables a wide range of applications. For example, we can use these links to group data across species at a high level. Many databases such as the Mouse Phenome Database (MPD) have to deal with the challenge of grouping trait data from a variety of studies for meta-analyses, e.g. all trait data associated with hypertension should be grouped. Similarly, the hierarchical structure allows for a broader search in FuTRES, where a user can query for "humerus length" and have all the ways humerus length is measured with measurement values returned, rather than having to do a separate search for each measurement method to retrieve data.

OBA is a component of the successor of the Unified Phenotype Ontology (Matentzoglu et al. 2018a), uPheno 2 (Ontology Xref Service). uPheno 2 is used by the Monarch Initiative (Shefchek et al. 2020) to integrate gene-to-phenotype data across species. The integration of OBA allows grouping of phenotype data across traits without concern for the specific manifestation (e.g. "blood glucose level" instead of "abnormally increased blood glucose level" or "limb morphology" instead of "short limb").

Limitations of the approach

Identifier uniqueness

One of the key OBO Foundry principles is class uniqueness: a single term, such as "amount of lysine in the blood", should not exist in multiple ontologies. Whilst the synchronisation with EFO is unproblematic (EFO is not an official OBO Foundry reference ontology, and measurement terms are conceptually disjoint from trait or attribute terms), the synchronisation with VT may raise some questions. Whilst the class uniqueness principle is absolutely central to reference ontologies such as PATO, ChEBI, GO and PRO, it is very complicated to maintain in a cross-species context. The prevalent practice is to have one species-independent vocabulary (Uberon, uPheno and now OBA) whose goal it is to integrate species-specific ontologies (MA, VT, XAO, ZFA, FOVT, etc.). Furthermore, species-specific ontologies are typically maintained as taxonomical structures (owl:subClassOf hierarchies with little additional axiomatisation) which means that they lack the strong logical foundation that integrator ontologies provide.

Need for manual mapping curation

The integration of GWAS data with data from a more qualitative phenotyping pipeline relies to a large extent on our mappings between OBA and EFO, and VT, which is an ongoing process. Due to their lack of (logical) formalisation, alignment is largely manual, but the comparatively small sizes of the relevant branches in EFO and VT makes it feasible to curate mappings semi-automatically using automated matchers and manual curation, as described in the "Automated mapping" section (Methods). The rapid improvement in Large Language Models and other NLP techniques may be able to speed up this process in the future.

Discussion and future work

The primary objective of OBA is to break silos across data types related to characteristics (e.g. "amount" or "mass"), traits or biological attributes (e.g. "amount of lysine in blood"), phenotypic abnormalities (e.g. "Hypolysinemia") and biological entities/processes (e.g. "blood", "lysine", "mitosis" or "cardiovascular system"). Due to its rich logical definitions, OBA naturally integrates well with data focussed on links to anatomy (such as gene expression data), chemical entities, cellular components, cell type, biological process and more. This allows, for example, the integration of anatomy focussed data (such as gene expression and single cell expression data) with trait-level data which is already a significant improvement over the status quo. Existing vocabularies to capture biological attributes, such as VT and EFO, do not (aside from the provision of simple cross-references) systematically bridge the gap between PATO characteristics, reference ontologies (e.g. anatomy, chemical) and phenotype ontologies.

Phenotype ontologies such as MP and HPO that define phenotypic abnormalities have been used for over a decade in the biomedical domain for clinical and model organism

phenotyping. Due to the widespread use of the EQ design pattern (see the "Logical framework" section), we can classify phenotypic abnormalities under their respective biological attributes ("Using OBA to group phenotypic abnormalities" section). Public endpoints such as Ubergraph (see "Programmatic access to OBA" section) demonstrate how hundreds of thousands of links between biological attributes and phenotypic abnormalities can be inferred automatically without a human in the loop. Furthermore, the integration of OBA into uPheno 2 allows to easily group phenotypic effects across biological attributes, which opens up powerful possibilities for search and grouping of annotations ("Use cases" section).

Many polygenic, quantitative and GWAS traits are not in scope for the Mendelian phenotype focussed ontologies. There are ontologies that focus on or include quantitative and measurable trait terms, such as VT, EFO and the Clinical Measurement Ontology (CMO) (Shimoyama et al. 2012; Smith et al. 2013). EFO curators, for example, maintain a branch in the ontology for "measurement" terms that are used in annotation by the GWAS Catalog (Sollis et al. 2023). "Measurement" terms, such as "urinary sodium measurement" (EFO:0021522), have a broad applicability in annotation. They can be used to annotate experiments independent of any conclusion about the results and outside of any context where conclusions might be made about biological traits. The GWAS Catalog uses these terms to record something more specific—an association between the presence of an allele and some effect on the measured value of a trait. Mapping a GWAS annotation with a "measurement" term to an OBA term, such as "urine sodium amount" (OBA:VT0006274) enables recording this explicitly, and has the advantage that the terms can be integrated directly with widely used phenotype ontologies, e.g. "Hypernatriuria" (HP:0012605). Measurement terms are still useful as they can record one of many assay methods for measuring a specific trait. For example, Body Mass Index is a useful, if sometimes limited, proxy measurement of body fat levels. Using a BMI measurement term to annotate GWAS variants can record this useful information, mapping this to a trait term for body fat levels then allows this to be integrated with related traits and phenotypes. Specialised ontologies that capture the measurement method exist, for example the Ontology of Biomedical Investigation (OBI) (Bandrowski et al. 2016) or the Biological Collections Ontology (BCO) (Walls et al. 2014).

The integration of quantitative trait data (such as GWAS or QTL) with outcomes from clinical and research organism phenotyping activities is one of the most promising applications of OBA. For example, the deep integration between OBA and HPO will facilitate the use of gene-phenotype associations derived from GWAS studies in variant prioritisation software such as Exomiser (Smedley et al. 2015), which is used for clinical diagnostics. This has the potential to significantly extend the existing sources of gene-phenotype data from annotations of Mendelian disease resources such as OMIM and Orphanet as well as model organism resources such as MGI (Blake et al. 2021), IMPC (Groza et al. 2023) and ZFIN (Bradford et al. 2022).

As the space of biological attributes/traits is very large, any curation of new terms must be highly scalable. To demonstrate how defining new biological attributes can largely be automated, we rapidly aligned more than 500 terms from the Mammalian Phenotype Ontology with OBA ("OBA coverage of biological attributes relevant for cross-species data integration" section) by repurposing logical definitions used and focussing on the curation of the specific phenotypic characteristic (e.g. "amount" instead of "increased amount"). Using logical definitions for automated reasoning and templates for scalable curation enables rapid development of terms. However, not all vocabularies make use of such logical definitions, which necessitates the use of manual and automated matching approaches. Due to terminological variability (different communities use different terminologies to talk about the same concepts) and the strong need for precision when constructing ontologies, we currently use a controlled approach involving equivalent string matching and expert-reviewed pre-processing steps (to populate the DOS-DP templates), rather than relying on Ontology Matching tools like AgreementMakerLight (AML) (Faria et al. 2013) for smart or fuzzy matching. Whilst this approach is very precise, it is incomplete, which means many terms need to be matched entirely manually by a human expert. Moving forward we expect a much higher volume of new term requests that will require us to scale up our curation effort. For example, there is an increasing demand from the GWAS Catalog for new trait terms to annotate summary statistics containing hundreds of GWAS studies directly submitted by authors. We are looking into improving our approach by combining our strict approach with automated matching such as AML. The required precision for curating coherent logical axioms requires, however, a human in the loop (at least in the near future). To that end, we have also started experimenting with ChatGPT (ChatGPT) to scale up manual curation significantly..

OBA can facilitate the interpretation of trait and phenotypic findings in clinical laboratory test results, many of which are annotated with Logical Observation Identifier Names and Codes (LOINC) (Forrey et al. 1996). As part of future work, we will bridge OBA to the LOINC database via the CompLOINC project (https://github.com/loinc/comp-loinc), which decomposes the (heavily pre-coordinated) LOINC classification into an OWL ontology with is-a hierarchies for each of the 6 LOINC Part Types (Component, System, Method, Property, Time and Scalar). This OWL formalisation of LOINC allows logical reasoning,

subsumption querying by Part Type, and has the potential to provide an extensive bridge between the LOINC-dominated clinical laboratory domain and the phenotype ontology world that dominates in the area of genomics. Rather than matching LOINC codes on the level of the highly variable LOINC labels, OBA terms can be matched much more easily to LOINC Part terms (e.g. chemical entities to ChEBI, anatomical entities to Uberon). This will result in a much-needed bridge between the clinical laboratory domain and biological research and genomics.

Also as future work, we seek to integrate OBA with disease ontology terms (which are also widely used, for example, in GWAS) through phenotypic features of diseases and common links to Uberon. For example, "familial juvenile hyperuricemic nephropathy" (MONDO:0000608) is linked to "Hyperuricemia" (HP:0002149) which is logically defined as "increased amount of uric acid in the blood". It therefore is automatically classified under "blood uric acid levels" (OBA:VT0010302). This gives us a natural bridge from diseases to biological attributes which provides another layer of integration. A second level of integration that has yet to be explored is to exploit the numerous "anatomical site" relations provided by disease ontologies such as Mondo—these are already integrated with OBA through the use of a common reference ontology (Uberon), but biological attribute terms could easily be generated based on the Uberon reference for more thorough logical integration. Prior to the development of OBA, researchers hoping to retrieve mouse trait measures reflective of human disease characteristics had to know specifically which mouse traits were associated with each disease, searching trait by trait to find a complete set. Using OBA mappings, mouse trait measures in MPD, for example, are readily annotated to 3033 disease ontology terms with more than 50% coverage across all MPD trait measures, allowing researchers a simple means of retrieving all disease associated trait data using a single, intuitive disease-centric query. These data can then be used to identify preclinical mouse models collectively extreme across a set of disease-related traits.

Related work

The Vertebrate Trait Ontology (VT) (Park et al. 2013) is a cross-species, unified trait vocabulary used for the annotation of terms in vertebrates. It was created based on the structure of the Mammalian Phenotype Ontology (MP), where references to abnormalities were removed and a skeletal set of neutral trait terms was maintained. It is therefore a phenotype-neutral ontology, which, similar to OBA, describes traits that do not indicate an abnormal state or process or express any phenotypic variation. Unlike OBA, VT terms are not constructed using logical axioms, and there are no logical links to other ontologies. VT uses weak non-logical cross-references to GO and MP to indicate that a link exists, but these links are sparse and cannot be used for automated reasoning (less than 20% of VT terms have such links, compared to 100% of OBA, which are logical and therefore more meaningful).

The Plant Trait Ontology (TO) (Cooper et al. 2018) is a Planteome database reference ontology that describes phenotypic traits in plants. Comparable to OBA, it is species-neutral and many TO terms also follow the EQ pattern, drawing entities from ontologies like the Plant Ontology (PO), GO and ChEBI and quality terms from PATO to provide pre-composed descriptions of terms and logically connect TO to other ontologies. Similarly, the flora phenotype ontology (FLOPO) (Hoehndorf et al. 2016), also employing the same Entity-Quality model used by OBA, is used for describing traits in the plant flora. TO and FLOPO can be seen as complementary to OBA, covering the world of plants, where OBA is focussed on Metazoan traits.

The Animal Trait Ontology (ATO) (Meunier-Salaün 2015) is an effort to form a central, standardised repository of controlled, phenotypic trait terms for three domesticated farm animal species. It was later expanded and is now referred to as the Animal Trait Ontology for Livestock (ATOL), an ontology of traits defining phenotypes described in the Environment Ontology for Livestock (EOL). One of ATOL's objectives is to use trait terms related to industry-wide technical measurements to promote standardisation. This was realised through the adoption of PATO's Entity-Quality formalism model (Gkoutos et al. 2018), the same model used by OBA.

Conclusion

The Ontology of Biological Attributes (OBA) is a species-independent ontology with numerous links to other biological and biomedical ontologies that integrates widely used phenotype ontologies such as HPO. A scalable logical framework based on design patterns and templates allows the rapid curation of precisely defined terms which not only bridge the gap between low level characteristics (such as "weight" and "amount") and reference ontologies such as the ChEBI (chemical entity) and Uberon (anatomy) ontologies, but also the currently wide chasm between quantitative "measurement" data such as GWAS and qualitative phenotyping data from clinical or model organism phenotyping activities. OBA is an active, evolving ontology that welcomes contributions and suggestions from the trait data community. In the near future, our goal is to integrate OBA more closely with clinical laboratory data (e.g. LOINC) and disease data (e.g. Mondo).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00335-023-09992-1.

Author contributions NM, RS, MAB, AI, DO-S wrote the main manuscript text NM, RS, VdS, AI prepared figures 1-4. EJC, RLB, DOW contributed analyses related to the Mouse Phenome Database NM, RS, NLH, SG, MH, CJM, JPB, ARC, JAM, MCM and DO-S edited the manuscript. CJM, MH, NM, DO-S, JPB, RS, contributed to OBA ontology design. NM, CJM, DO-S, RS and ARC contributed to the technical development of OBA, NM, RS, JPB, MAB, SMB, LWH, AI, CJM, ES, AET, NV, RLB, DOW, EJC, PNR, SK, TP, PR, DS and DO-S contributed content and term suggestions, clarifications or edited OBA ontology terms. All authors reviewed the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by NIH National Human Genome Research Institute Phenomics First Resource, NIH-NHGRI # 5RM1 HG010860, a Center of Excellence in Genomic Science; the Office of the Director, National Institutes of Health (#5R24 OD011883); Director, Office of Science, Office of Basic Energy Sciences, of the US Department of Energy [DE-AC0205CH11231 to NLH and CJM]. EJC, RLB and DOW are supported by R01 DA028420 and U54OD030187.

Data availability The Ontology of Biological Attributes (OBA) is available from the OBO Foundry Persistent Uniform Resource Locator: http://purl.obolibrary.org/obo/oba.owl

Declarations

Competing interests The authors declare no competing interests.

References

Alliance of Genome Resources Consortium (2022) Harmonizing model organism data in the Alliance of Genome Resources. Genetics 220:iyac022

Andrés-Hernández L et al (2022) Establishing a common nutritional vocabulary—from food production to diet. Front Nutr 9:928837

Balhoff JP et al (2022) Ubergraph: integrating OBO ontologies into a unified semantic graph. https://icbo-conference.github.io/icbo2022/papers/ICBO-2022_paper_5005.pdf

Balk MA et al (2022) A solution to the challenges of interdisciplinary aggregation and use of specimen-level trait data. iScience 25:105101

Bandrowski A et al (2016) The Ontology for Biomedical Investigations. PLoS ONE 11:e0154556

Batista-Navarro R, Hammock J, Ulate W, Ananiadou S (2016) A text mining framework for accelerating the semantic curation of literature. In: Research and advanced technology for digital libraries. Springer, New York, pp 459–462. https://doi.org/10.1007/978-3-319-43997-6_44

Bello SM, Perry MN, Smith CL (2021) Know your model: a brief history of making mutant mouse genetic models. Lab Anim 50:263–266

bio-attribute-ontology. (Github)

Blake JA et al (2021) Mouse Genome Database (MGD): knowledgebase for mouse-human comparative biology. Nucleic Acids Res 49:D981–D987

Bogue MA et al (2023) Mouse Phenome Database: towards a more FAIR-compliant and TRUST-worthy data repository and tool suite for phenotypes and genotypes. Nucleic Acids Res 51:D1067–D1074

Bradford Y et al (2011) ZFIN: enhancements and updates to the Zebrafish Model Organism Database. Nucleic Acids Res 39:D822–D829

Bradford YM et al (2022) Zebrafish information network, the knowledgebase for Danio rerio research. Genetics 220:iyac016

Buttigieg PL et al (2016) The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. J Biomed Semantics 7:57

Cantelli G et al (2022) The European Bioinformatics Institute (EMBL-EBI) in 2021. Nucleic Acids Res 50:D11–D19

Clark JF, Dinsmore CJ, Soriano P (2020) A most formidable arsenal: genetic technologies for building a better mouse. Genes Dev 34:1256–1286

COB: An experimental ontology containing key terms from Open Biological and Biomedical Ontologies (OBO). (Github)

Cooper L et al (2018) The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Res 46:D1168–D1180

Dahdul WM et al (2010) Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature. PLoS ONE 5:e10708

Dececchi TA, Balhoff JP, Lapp H, Mabee PM (2015) Toward synthesizing our knowledge of morphology: using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. Syst Biol 64:936–952

Detwiler LT, Suciu D, Brinkley JF (2008) Regular paths in SparQL: querying the NCI Thesaurus. AMIA Annu Symp Proc 2008:161–165

dosdp-tools: Utility for working with DOSDP design patterns and OWL ontologies. (Github)

Faria D, Pesquita C, Santos E, Palmonari M, Cruz IF, Couto FM (2013) The AgreementMakerLight Ontology Matching System. In: Meersman R, Panetto H, Dillon T, Eder J, Bellahsene Z, Ritter N, De Leenheer P, Dou D (eds) On the move to meaningful internet systems: OTM 2013 conferences, vol 8185. Springer, Berlin, pp 527–541

Farrell B, Bengtson J (2019) Scientist and data architect collaborate to curate and archive an inner ear electrophysiology data collection. PLoS ONE 14:e0223984

Fisher ME et al (2022) The Xenopus phenotype ontology: bridging model organism phenotype data to human health and development. BMC Bioinform 23:99

Forrey AW et al (1996) Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. Clin Chem 42:81–90

Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 49:D325–D334

Giblin KA et al (2021) New associations between drug-induced adverse events in animal models and humans reveal novel candidate safety targets. Chem Res Toxicol 34:438–451

Gkoutos GV, Green ECJ, Mallon A-M, Hancock JM, Davidson D (2005) Using ontologies to describe mouse phenotypes. Genome Biol 6:R8
Gkoutos GV, Schofield PN, Hoehndorf R (2018) The anatomy of phenotype ontologies: principles, properties and applications. Brief Bioinform 19:1008–1021
Grau BC et al (2008) OWL 2: the next step for OWL. J Web Semantics 6:309–322
Groza T et al (2023) The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. Nucleic Acids Res 51:D1038–D1045
Harper L et al (2018) AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. Database 2018:bay088
Hastings J et al (2016) ChEBI in 2016: improved services and an expanding collection of metabolites. Nucleic Acids Res 44:D1214–D1219
Hoehndorf R et al (2016) The flora phenotype ontology (FLOPO): tool for integrating morphological traits and phenotypes of vascular plants. J Biomed Semantics 7:65
Horn T (2016) Integrating biodiversity data into botanic collections. Biodivers Data J 4:e7971. https://doi.org/10.3897/BDJ.4.e7971
Hsu PD, Lander ES, Zhang F (2014) Development and applications of CRISPR-Cas9 for genome engineering. Cell 157:1262–1278
Jackson RC et al (2019) ROBOT: a tool for automating ontology workflows. BMC Bioinform 20:407
Jackson R et al (2021) OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. Database 2021:baab069
Jupp S, Burdett T, Leroy C, Parkinson HE (2015) A new Ontology Lookup Service at EMBL-EBI. SWAT4LS 2:118–119
Kaldunski ML et al (2022) The Rat Genome Database (RGD) facilitates genomic and phenotypic data integration across multiple species for biomedical research. Mamm Genome 33:66–80
Kazakov Y, Krötzsch M, Simančík F (2014) The incredible ELK. J Automat Reason 53:1–61
Köhler S et al (2021) The human phenotype ontology in 2021. Nucleic Acids Res 49:D1207–D1217
Le Franc Y et al (2020) D2.2 FAIR semantics: first recommendations. https://doi.org/10.5281/zenodo.3707985
Ma L et al (2022) Database commons: a catalog of worldwide biological databases. Genomics Proteomics Bioinform. https://doi.org/10.1016/j.gpb.2022.12.004
Mabee BP et al (2012) 500,000 fish phenotypes: the new informatics landscape for evolutionary and developmental biology of the vertebrate skeleton. J Appl Ichthyol 28:300–305
Malone J et al (2010) Modeling sample variables with an Experimental Factor Ontology. Bioinformatics 26:1112–1118
Matentzoglu N et al (2018a) Phenotype Ontologies Traversing All The Organisms (POTATO) workshop aims to reconcile logical definitions across species.. https://doi.org/10.5281/zenodo.2382757
Matentzoglu N, Malone J, Mungall C, Stevens R (2018b) MIRO: guidelines for minimum information for the reporting of an ontology. J Biomed Semantics 9:6
Matentzoglu N et al (2022a) A Simple Standard for Sharing Ontological Mappings (SSSOM). Database 2022:baac035
Matentzoglu N et al (2022b) Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies. Database 2022:baac087
Meunier-Salaün M-C (2015) ATOL: Animal Trait Ontology for livestock. In: Scientific conference (unknown, 2015)
Moses L, Niemi S, Karlsson E (2018) Pet genomics medicine runs wild. Nature 559:470–472
Motik B, Grau BC, Horrocks I (2009) OWL 2 web ontology language profiles, 2nd edn. https://www.w3.org/TR/owl2-profiles/
Mungall CJ et al (2010) Integrating phenotype ontologies across multiple species. Genome Biol 11:R2
Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA (2012) Uberon, an integrative multi-species anatomy ontology. Genome Biol 13:R5
Musen MA, Protégé Team (2015) The Protégé project: a look back and a look forward. AI Matters 1:4–12
oba_alignment.ipynb at master · obophenotype/bio-attribute-ontology. (Github)
OBO foundry. https://obofoundry.org/principles/fp-003-uris.html
Ochoa D et al (2023) The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. Nucleic Acids Res 51:D1353–D1359
Ong E et al (2017) Ontobee: a linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic Acids Res 45:D347–D352
Ontology Xref Service. Unified phenotype ontology (uPheno2) < ontology lookup service < monarch initiative. https://ols.monarchinitiative.org/ontologies/upheno2
ontology-access-kit: Ontology Access Kit: a python library and command line application for working with ontologies. (Github)
Osumi-Sutherland D, Courtot M, Balhoff JP, Mungall C (2017) Dead simple OWL design patterns. J Biomed Semantics 8:18
Park CA et al (2013) The Vertebrate Trait Ontology: a controlled vocabulary for the annotation of trait data across species. J Biomed Semantics 4:13
Parr C et al (2016) TraitBank: practical semantics for organism attribute data. Semantic Web 7(6):577–588
Pathak RK, Kim J-M (2022) Vetinformatics from functional genomics to drug discovery: insights into decoding complex molecular mechanisms of livestock systems in veterinary science. Front Vet Sci 9:1008728
Rehm HL et al (2021) GA4GH: International policies and standards for data sharing across genomic research and healthcare. Cell Genom 1:100029
Schriml LM et al (2022) The human disease ontology 2022 update. Nucleic Acids Res 50:D1255–D1261
Shefchek KA et al (2020) The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res 48:D704–D715
Shimoyama M et al (2012) Three ontologies to define phenotype measurement data. Front Genet 3:87
Slater LT, Gkoutos GV, Hoehndorf R (2020) Towards semantic interoperability: finding and repairing hidden contradictions in biomedical ontologies. BMC Med Inform Decis Mak 20:311
Smedley D et al (2015) Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc 10:2004–2015
Smith CL, Eppig JT (2009) The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdiscip Rev Syst Biol Med 1:390–399
Smith JR et al (2013) The clinical measurement, measurement method and experimental condition ontologies: expansion, improvements and new applications. J Biomed Semantics 4:26
Sollis E et al (2023) The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res 51:D977–D985
src/patterns/dosdp-patterns at master · obophenotype/bio-attribute-ontology. (Github)
Stephens ZD et al (2015) Big data: astronomical or genomical? PLoS Biol 13:e1002195
Thessen AE et al (2015) Emerging semantics to link phenotype and environment. PeerJ 3:e1470
Thessen AE et al (2020) Transforming the study of organisms: phenomic data models and knowledge bases. PLoS Comput Biol 16:e1008376
Vasilevsky N et al (2020) Mondo Disease Ontology: harmonizing disease concepts across the world. In: CEUR workshop proceedings, vol 2807 (CEUR-WS, 2020)

Walls RL et al (2014) Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies. PLoS ONE 9:e89606

Washington NL et al (2009) Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biol 7:e1000247