Classifying Animal Breeds with the Vertebrate Breed Ontology (VBO)

Sabrina **Toro** *1*, Nicolas **Matentzoglu** *2*, Kathleen R. **Mullen** *1*, Nicole **Vasilevsky** *1*, Halie M. **Rando** *1*, Melissa **Haendel** *1*, Christopher J. **Mungall** *3*, Zhi-Liang **Hu** *4*, Gregoire **Leroy** *5*, Imke **Tammen** *6*, and Frank W. **Nicholas** *6*

1 University of Colorado Denver, Anschutz Medical Campus, Aurora, CO USA
2 Semanticly, Athens, Greece
3 Lawrence Berkeley National Laboratory, Berkeley, CA, USA
4 Department of Animal Science, Iowa State University, Ames, IA 50011, USA
5 Food and Agricultural Organization of the United Nations, Rome, Italy
6 The University of Sydney, Sydney School of Veterinary Science, NSW 2006, Australia.

Abstract

In the current era of biomedical big data, advances in diagnostics and treatments can leverage a wealth of information from research and health records. This approach requires integrating and comparing data related to genes, phenotypes, and diseases reported in disparate sources, such as health records and animal model databases. Integrating and comparing these data sources requires data standardization, including the use of ontologies.

Currently, some resources for non-human animals, such as Online Mendelian Inheritance in Animals (OMIA), report data using their own terminologies, and therefore comprehensive computable vocabularies for naming non-human animals, their phenotypes, and other associated data are needed. The Monarch Initiative, in collaboration with OMIA, created the Vertebrate Breed Ontology (VBO) as a means of incorporating all breed names into a single computable resource.

VBO is based on the Domestic Animal Diversity Information System (DAD-IS) breeds list created and maintained by the United Nations' Food and Agriculture Organization (FAO), and has been created using standard semantic engineering tools including the Ontology Development Kit (ODK). VBO terms refer to breeds defined by their name, country of existence, and species. Annotations such as synonyms and cross-references allow bridging between VBO and existing breed lists and ontologies. Other annotations, such as the domestication status and description of origin, incorporate additional information from DAD-IS.

The adoption of VBO as the source of breed names in databases such as OMIA and other non-human health records is one step in making information more computable and consistent, therefore enhancing data interoperability. This, in addition to the adoption of ontologies for phenotypes and diseases, will support data integration and comparison, and ultimately diagnosis and treatments, for non-human animals.

Keywords

Ontology, animal breeds, vertebrate animals, disease diagnostics

1. Introduction

Throughout history, humans have formed close relationships with other animals as food, transportation, labor, and companionship. Human influence had a major impact through selective breeding which, along with natural adaptation and/or genetic isolation, led to the development/evolution of breeds. By its very nature, a breed cannot be defined in strict scientific terms. The Food and Agriculture Organization of the United Nations (FAO), which has collected more information on breeds than any other organization, provides two concepts for breeds: a group of domestic animals with definable and identifiable external characteristics that enable it to be separated by visual appraisal from other similarly defined groups within the same species; or a group of domestic animals for which geographical and/or cultural separation from phenotypically similar groups has led to acceptance of its separate identity [1]. Both definitions suggest breeds are concepts well-matched to ontological structuring.

Advances in diagnostics and treatments for human health conditions can leverage a wealth of information from biomedical research and from medical and veterinary records; however, even though diagnostics and treatments of animal breeds have advanced alongside their human counterparts, progress in the aggregation and integration of data from non-human animal diseases and phenotypes that could be leveraged for this purpose has lagged. The Online Mendelian Inheritance in Animals (OMIA) database catalogs inherited disorders, traits, and genes in 346 animal species (excluding human, mouse, and rat) and is, therefore, a prime resource for non-human animal information [2] (https://omia.org). The information available in OMIA has been recorded by referring to breeds as they are reported in publications. Consequently, breed names are often used in OMIA inconsistently. For OMIA to be interoperable with other relevant resources, a comprehensive, internally consistent, computable breed list is required, including all synonyms.

The initial intention was to base the OMIA breed ontology on the Livestock Breed Ontology (LBO) (https://www.animalgenome.org/bioinfo/projects/lbo/), which was created in the context of genomic research in leading livestock breeds. However, the overlap in the LBO and OMIA breed list was lower than expected, reflecting OMIA's preoccupation with rare, inherited diseases in (sometimes) rare livestock breeds and in companion animals. By far the most comprehensive global resource is the list of breed names and synonyms compiled by FAO in its Domestic Animal Diversity Information System (DAD-IS, https://www.fao.org/dad-is/). This list contains information on more than 15,000 national breed populations representing more than 8,800 breeds across 38 species. It is assembled, contributed to, and updated by country-nominated National Coordinators from 182 countries, making the DAD-IS breeds list the most exhaustive and up-to-date resource available. Although this list resides in a powerful online database, the list itself is not interoperable.

The Monarch Initiative [3] and the OMIA teams, in consultation with FAO, collaborated to create the Vertebrate Breed Ontology (VBO, https://github.com/monarch-initiative/vertebrate-breed-ontology). This ontology builds on the information in the DAD-IS breeds list, incorporates additional breed names from OMIA and other sources (especially for dogs and cats), and integrates all this information into an ontological format. VBO aims to be a comprehensive source for breed names across all vertebrate animals, including livestock and companion animals.

The use of VBO will improve non-human animal information, such as that in OMIA, enabling all relevant breed information, including the occurrence of models for human diseases, to be accurately related across resources. This will ultimately support the integration of non-human animal data and allow the more effective comparison of genes, phenotypes and diseases across species, including humans. VBO will also enhance the contribution of DAD-IS by facilitating its reciprocal hyperlinking with other resources relating to breed characterization and the management of animal genetic resources, including germ-plasm preservation programs.

2. Vertebrate Breed Ontology

2.1. Identifying components for unique "breeds" and "national breed populations" in DAD-IS

VBO defines "breed" as a group of animals within a species that have been selected for specific characteristics (such as traits, behavior,...), or which cultural or geographical separation has led to

general acceptance of its separate identity [1]. Breeds included in VBO have been characterized, defined, and determined by recognized organizations such as the National Coordinators at FAO who compiled livestock breed-related information into DAD-IS.

Livestock breeds represented in DAD-IS include "local breeds" (breeds reported in a single country), and "transboundary breeds" (breeds reported in several different countries). In addition, DAD-IS includes "national breed populations", representing instances of the transboundary breed in a particular country. We reviewed the DAD-IS list and determined how to represent individual breeds and national breed populations in VBO. The information in DAD-IS is centered on the following attributes: country of the breed/national breed population (referring to the country in which the breed/national breed population is reported), species, and the breed/national breed population's "most common name". Different records can share some of the same attributes listed above. For example, breeds from different species can share the same "most common name" in the same country (e.g. "Tibetan" in China is the most common name for a breed of each cattle, chicken, pig, goat, etc.). Similarly, the same most common name can refer to the same breed in different countries (e.g. "Jersey Giant" is a breed of chicken that exists in Canada, Ireland, Luxembourg, etc.). Therefore, defining unique records with unique labels in VBO required the inclusion of the breed/national breed population's species, country, and most common name; hence, term labels were created by concatenating these components, following the format: `'Most common name, Country; Species'`, in which country and species are the English names (e.g. 'Jersey Giant, Canada; Chicken' VBO:0006068). Records for "Transboundary breeds", which exist across several countries, often with a different most common name in each country, do not include the "country" attribute, and can be uniquely represented with their internationally recognized name and the species of the breed, following the format: `'Transboundary name; Species'` (e.g. 'Jersey Giant; Chicken' VBO:0000547). The instance of the transboundary breed in a particular country, the national breed populations, are represented using the name/country/species format reported above (e.g. 'Jersey Giant, Canada; Chicken' VBO:0006068) and are subclasses of their corresponding transboundary breed (Figure 1).

2.2. Use of NCBITaxon as the top-level classification of vertebrate breeds

Breeds are identified by National Coordinators as distinctive groups within a species. Therefore, breeds are a subclass of a species. Species are represented by the NCBITaxon entities; hence, we integrated VBO terms within the NCBITaxon hierarchy using the *is_a* relation (Figure 1).
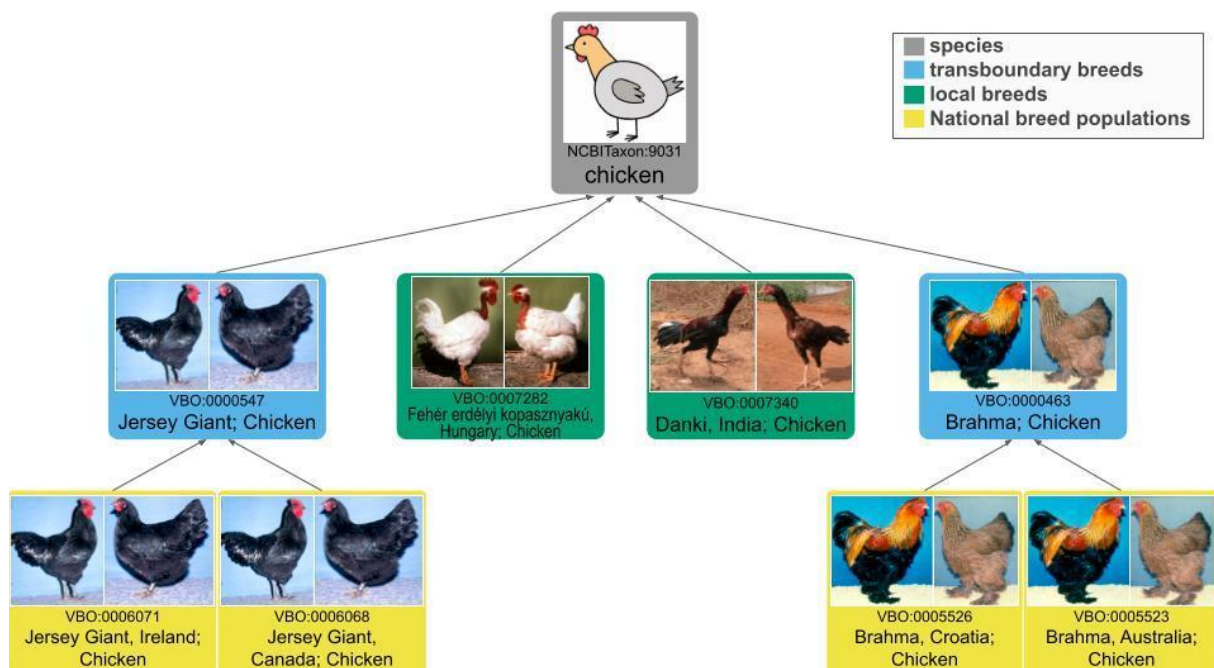
Figure 1: Classification of chicken breeds in VBO.
Some chicken breeds (green boxes), such as 'Fehér erdélyi kopasznyakú, Hungary; Chicken' (VBO:0007282) or 'Danki, India; Chicken' (VBO:0007340), exist in just one country and hence are direct children of the 'chicken' species represented in NCBITaxon '*Gallus gallus*' (NCBITaxon:9031, grey box). Other chicken breeds exist in more than one country; these "transboundary breeds" (blue box) are also direct children of the 'chicken' species. The corresponding national breed populations (yellow boxes) share a "transboundary breed" (blue box) parent. Arrows represent the '*is_a*' relation. It should be noted that other existing breeds of chicken are not displayed here.
Photo credits: https://www.fao.org/dad-is/image-browser/en, Mr Hans L. Schippers ('Jersey Giant; Chicken', 'Jersey Giant, Ireland; Chicken', 'Jersey Giant, Canada; Chicken', 'Brahma; Chicken', 'Brahma, Croatia; Chicken', 'Brahma, Australia; Chicken'), and NBAGR, Karnal, India ('Danki, India; Chicken').

2.3. Creation of VBO using the Ontology Development Kit (ODK)

The Ontology Development Kit (ODK, https://github.com/INCATools/ontology-development-kit) provides a framework for creating ontologies, including both executable workflows for managing ontologies such as release workflow and continuous integration as well as ontology-processing tools such as ROBOT. The ODK is used to automatically check the VBO for errors whenever changes are proposed (e.g. new classes are added), and to release new ontology versions. The ontology is managed and openly available on GitHub at https://github.com/monarch-initiative/vertebrate-breed-ontology.

We used the DAD-IS list as the foundation for VBO terms and attributes. We manually mapped species to their corresponding NCBITaxon record and are mapping breed/national breed population names from other resources to the relevant VBO classes. VBO was created and is currently maintained via ROBOT templates [4].

2.4. VBO term metadata

VBO term labels follow the format described above, based on the breed/national breed population's most common name, country, and species as reported in the DAD-IS list. However, other names are used to refer to breeds/national breed populations in publications and databases, and the DAD-IS list includes many of these; we, therefore, record alternative names as "exact synonym" annotations (Table 1), including names in the language of the breed/national breed population country of origin when applicable. These synonyms are annotated with their source. For instance, we manually mapped the breed names used in OMIA to the relevant VBO terms. These breed names used in OMIA were added as 'exact synonyms', which received "OMIA" as a "source" annotation to indicate that the synonym is used in the OMIA database.

Data interoperability relies on the ability to determine equivalent terms between terminologies. Therefore, in addition to 'exact synonyms' and their source annotation, VBO terms were annotated with 'database cross-reference' annotations that include the identifier (id) for the same concept in other terminologies. These corresponding mappings are being done manually by a curator. To date, we have added cross-references to the internal OMIA ids, which will facilitate the incorporation of VBO into OMIA.

The country where the breed/national breed population is reported was recorded with the annotation 'exists_in_country', using the Wikidata id to refer to the country. Many resources, including DAD-IS, refer to country using the UN ISO standard three-letter country codes (https://www.iso.org/iso-3166-country-codes.html). We, therefore, added the ISO country codes to the VBO term using the annotation "exists_in_country_ISOcode" in addition to the Wikidata country id. Other metadata associated with a VBO term include the domestication status of the breed (whether it is domestic, wild, or feral) and the description of the breed origin. All metadata includes the source where the information originated, which, in these early stages of the ontology, is mostly DAD-IS/FAO.

Table 1 VBO term metadata

Metadata	Annotation ID	Description	Examples

rdfs:label	http://www.w3.org/2000/01/rdf-schema#label	Single label	'Ayrshire, Finland; Cattle'
has_exact_synonym	http://www.geneontology.org/formats/oboInOwl#hasExactSynonym	Optional, any other name referring to the breed	Finnish Ayrshire
database_cross_reference	http://www.geneontology.org/formats/oboInOwl#hasDbXref	Optional, identifier referring to the same breed in another ontology, database, or the source	OMIA:923
exists_in_country	http://purl.obolibrary.org/obo/vbo#existsincountry	Wikidata entry for the country where the breed exists	wikidata:Q33
exists_in_country_ISOcode	http://purl.obolibrary.org/obo/vbo#existsincountryiso	ISO country code for the country where the breed exists	FIN
has_domestication_status	http://purl.obolibrary.org/obo/vbo#hasdomesticationstatus	Optional	domestic
in_subset	http://www.geneontology.org/formats/oboInOwl#inSubset	Optional, tags indicating transboundary breed	transboundary
description of origin	http://purl.obolibrary.org/obo/vbo#origindescription	Optional	(...) based on animals imported at the end of the 19th and the beginning of the 20th century. The breed has been crossed with other Nordic red breeds.
source	http://www.geneontology.org/formats/oboInOwl#source	Added to an annotation, indicates the annotation provenance	https://www.fao.org/dad-is

3. Conclusion and future work

VBO was created out of a need for a comprehensive, computable compendium of breed names in order to make OMIA data interoperable with other resources. The first version of VBO as reported in this paper focused on livestock breeds/national breed populations included in the DAD-IS list. We were able to determine criteria to ensure term uniqueness (as required for ontologies); however, it necessitated the concatenation of breed/national breed population common name, country, and species, making the VBO term label unwieldy and cumbersome. Even though this term label format is less than ideal, we could not find an alternative that would satisfy both term label uniqueness and practicality.

The livestock breeds/national breed populations in VBO are in accordance with their characterization and determination by the FAO national coordinators. As a result, the country in which

a breed is reported, which can differ from the country where the breed originated, is a central characteristic of a livestock breed in VBO, as it is in DAD-IS. It should be noted that the current DAD-IS system is limited, and more specific cases such as breed sub-populations are currently not represented. For example, sub-populations of a single breed (e.g. Lacaune; Sheep (VBO:0001474) bred for dairy versus meat uses) might be maintained differently in different parts of the same country (e.g. France); even though these sub-populations should refer to the same breed, they are currently recorded as distinct breeds (e.g. 'Lacaune Lait, France; Sheep' (VBO:0014841) and 'Lacaune Viande, France; Sheep' (VBO:0014842) are two different local sheep breeds) with no obvious relation to each other.

The first version of VBO was created in early 2022, and although much progress has been made in developing VBO, several immediate improvements are in development. We will join a consortium comprising people who are knowledgeable about breeds and breed ontologies, with the aim of reaching the widest possible consensus on VBO terms, record keeping and annotation. We will enhance workflows for users to add synonyms and to map breed names and/or breed ids from other resources to VBO classes. In addition to the ongoing work to refer to OMIA records, we will refer to other resources such as LBO. It should be noted that the creation of non-ambiguous mappings between VBO and LBO is our short-term strategy to reconcile both ontologies; discussions for the long-term plans are ongoing, with VBO and LBO potentially merging or LBO potentially becoming a part of VBO. We are also working on submitting VBO to the OBO Foundry, and we will work with FAO to review the ontology and determine what other requirements they would need from VBO. We are very aware that there is a continual need for breeds to be added to VBO, e.g. breeds that have not yet been incorporated into DAD-IS. Also, work is in progress to add breeds for companion animals such as dogs (which DAD-IS covers only partially) and cats (which are outside DAD-IS' terms of reference), as well as laboratory animals. Also, given that the DAD-IS list is a compilation of contributions from 182 National Coordinators, it is recognized that inconsistencies within the list, especially to do with links to transboundary breed names, are likely to occur from time to time. Strategies will be developed with FAO to identify and rectify the inconsistencies and to incorporate DAD-IS updates into VBO. DAD-IS includes much information on breeds/national breed populations that is not currently incorporated into VBO, e.g. risk of extinction. Adding more information from DAD-IS and from other sources to VBO is relatively straightforward, and we will work with FAO and the consortium in deciding to what extent other information should be included.

VBO provides an ontological compendium for breeds. When adopted by databases and in non-human animal health records, VBO will improve data interoperability by disambiguating the breeds associated with traits/diseases, support data comparison, and ultimately advance diagnosis and treatment for both human and non-human animal patients.

4. Acknowledgements

5. References

[1] Food and Agriculture Organization of the United Nations. The Global Strategy for the Management of Farm Animal Genetic Resources: Executive Brief. Food and Agriculture Organization of the United Nations; 1999.

[2] Nicholas FW. Online Mendelian Inheritance in Animals (OMIA): a record of advances in animal genetics, freely available on the Internet for 25 years. Anim Genet. 2021;52: 3–9.

[3] Shefchek KA, Harris NL, Gargano M, Matentzoglu N, Unni D, Brush M, et al. The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. 2020;48: D704–D715.

[4] Jackson RC, Balhoff JP, Douglass E, Harris NL, Mungall CJ, Overton JA. ROBOT: A Tool for Automating Ontology Workflows. BMC Bioinformatics. 2019;20: 407.