ICBO: International Conference on Biomedical Ontologies · Buffalo, NY, USA
Facilitating Anatomy Ontology Interoperability Workshop, July 27, 2011

# The Vertebrate Bridging Ontology (VBO)

Ravensara Travillian[1], James Malone[1], Chao Pang[2], John Hancock[3],
Peter W.H. Holland[4], Paul Schofield[5], Helen Parkinson[1]

[1]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK
[2]Genomics Coordination Center, Groningen Bioinformatics Center, University of Groningen
and Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands
[3]MRC Harwell, Harwell, Oxfordshire, UK
[4]Department of Zoology, University of Oxford, UK
[5]Department of Physiology, Development and Neuroscience, University of Cambridge, UK

**Abstract.** The recent proliferation of ontologies for organizing and modeling anatomical, phenotypic, and genetic information is a welcome development, with a great deal of potential for transforming the way scientists access and use knowledge. Realization of this potential calls for effective ways of integrating and computing on various information sources. In this paper, we introduce the Vertebrate Bridging Ontology (VBO), which permits the transfer of information about homologous anatomical structures between species – a first step towards the integration of species-specific anatomical ontologies. We present the ontology, design patterns, and methodology, and discuss how it can be applied to use-cases to meet the information needs of the scientific user community.

**Keywords:** anatomy, ontology, vertebrate, evolutionary biology, homology

**Availability:** http://sourceforge.net/projects/vbo/
http://bioportal.bioontology.org/projects/102
http://www.ebi.ac.uk/ebiwiki/VBO/index.php/Main_Page

## 1 Introduction

The problem of integrating diverse single-species anatomy ontologies is well-documented [1]. Comparison of conserved and divergent patterns of gene expression and mutant phenotypes between species has become a powerful approach for investigating gene function and its evolution, particularly as more and more data accumulates from a wide range of species. In order to facilitate a computational approach to cross-species comparisons it is necessary to formalize the description of anatomy in each species, but this then leaves us with the problem of crossing between evolutionarily homologous structures in separate species. Two existing approaches have been attempted: lexical matching and the generation of a "universal" vertebrate anatomy ontology. The former is, for reasons discussed in [1] and below, always going to be intrinsically flawed. The latter has met with some success with the development of the CARO upper level anatomy ontology, and the Uberon multi-species metazoan anatomy ontology [2, 3]. However neither take full account of the evidence-based inferred evolutionary relationships between anatomical structures in different taxa. In this paper, we introduce the Vertebrate Bridging Ontology (VBO), an evidence – based approach which permits the transfer of information about homologous anatomical structures across species – a first step towards the integration of species-specific anatomical ontologies.

## 2 Development and Implementation of VBO

The VBO is developed in the Web Ontology Language (OWL) using Protégé 4, in order to provide a common representation compatible with that of the single-species ontologies it is intended to integrate. The OBO (Open Biomedical Ontologies) recommendation of unique namespaces and identifiers has been adhered to in its development.

Use cases collected at a VBO community workshop in June 2010 include key questions the evolutionary-biology and biomedical research communities might wish to address:

1. Gene driven: Compare expression of (a) a named gene or (b) gene family or (c) combination of genes between species in homologous tissues. The queries from this use case will take such forms as: Which anatomical structures are involved in the expression of this {gene | gene family | combination of genes}? Are these structures the same or different in different species? Is expression conserved between species only in homologous structures?
2. Anatomy driven: Compare transcriptomes in a particular homologous anatomical structure between species. The queries from this use-case will take forms such as: For this specific structure, are the same genes or different genes are expressed? What are the differential expression patterns among homologous structures in different species?
3. Compare gene expression similarity and/or difference in particular tissues between species to test a hypothesis of homology. The queries from this use-case will take forms such as: Is Tissue A in Species 1 likely to be homologous to Tissue B in Species 2 as assessed through transcriptome similarity?

Data for these use cases comes from user annotations of model organisms within ongoing human disease mechanism studies, comparative gene expression studies for functional genomics and evolutionary biology, and phenotype/genotype association studies in adult and developing organisms.

**Approach.** The VBO is based only on anatomical homology – that is, evolutionary relatedness of structures by uninterrupted descent from a common ancestor. The other types of structural similarity in classical comparative anatomy – analogy (similarity of function), and homoplasy (similarity of appearance independent of common descent) – are not part of the VBO scope.

Homology is symmetric, reflexive, and transitive, and thus the homologous nodes for a particular structure form a maximally-connected graph for the relation *homologous-to*. The combinatorial complexity of the possible axioms linking anatomical entities of even a few species requires a programmatic approach to populating the classes and relationships within the VBO framework. There are two ways to leverage evolutionary anatomical relationships to programmatically populate VBO: a most recent common ancestor (MRCA, "top-down") approach and a homology chain ("bottom-up") approach [4], illustrated in Fig. 1.

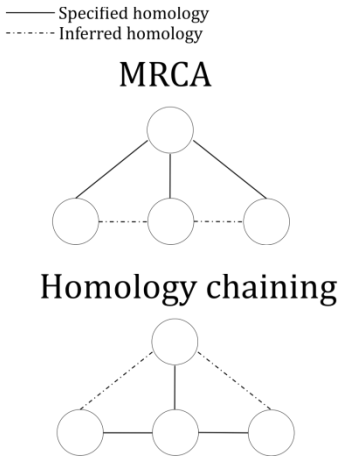


**Figure 1.** The MRCA approach (*top*) specifies homologies from the MRCA to its descendants, and homologies among the descendants are inferred. The homology chains approach (*bottom*) specifies homologies among the descendants, and requires one explicit connection to the MRCA for that characteristic in order to infer all the other homologies from the descendants to the MRCA.

The two approaches are similar in efficiency, but in principle we favored the MRCA approach as it is more similar to the way biologists reason over evolutionary relationships. In practice, we ended up using a hybrid approach, because the data often were available for one approach but not the other.

**Entities.** There are two types of entity in VBO: anatomical structures and taxa. An anatomical structure consists of the following data structure (Fig. 2), where the surrounding circles represent annotation properties that link the structure to the homologous structure in other ontologies and taxonomies:

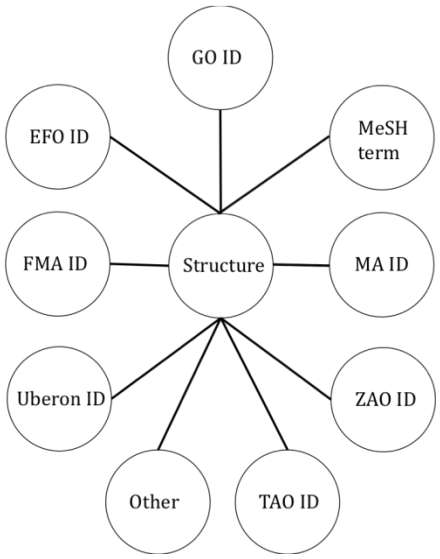


**Figure 2.** The data structure of an anatomical entity in the VBO (*center*), with annotation properties (*surrounding*).

The corresponding structure(s) in the Experimental Factor Ontology (EFO) [5] is/are linked via the EFO ID, the corresponding structure(s) in the Foundational Model of Anatomy (FMA) [6] are linked via the FMA ID, the corresponding structure(s) in the Teleost Anatomy Ontology (TAO) [7] are linked via the TAO ID, and so forth. The annotation property "Other" represents additional IDs that can be added as the VBO is aligned with additional species anatomy ontologies.

For VBO 1.0, we selected the adult skeletal system for demonstration and proof-of-principle, as it is a relatively straightforward example to model: it tends to be bilaterally symmetrical and highly conserved, with relatively little sexual dimorphism. However, data for other systems became available during the course of the project, so VBO also contains structures outside the adult skeletal system.

Taxon entities can be at any level of phylogenetic ranking, because anatomical structures can be characteristic of any level of ranking. For example, jaws are characteristic of the infraphylum Gnathostomata, while hair, sweat (eccrine) glands, and mammary glands are characteristics of the class Mammalia, and hypertrophied manus digits supporting wings are characteristic of the order Chiroptera. While the scope of the VBO is vertebrate structures, many structures that are characteristic of vertebrates actually originate further back in evolutionary history, so a rigorous modeling of the VBO requires the ability to model structures as differentia at the appropriate taxon ranking. The current VBO phylogeny is consistent with the NCBI taxonomy for vertebrates.

Taxon entities in VBO have the following data structure (Fig. 3):

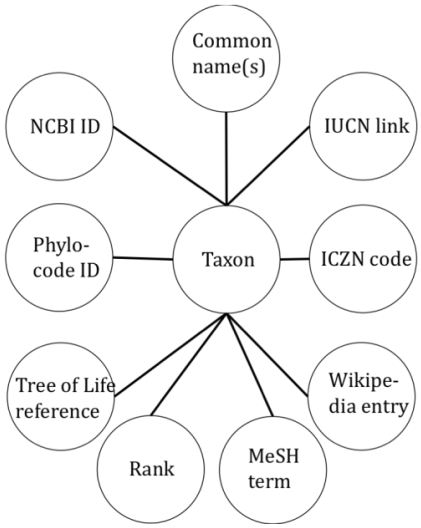


**Figure 3.** The data structure of a taxon entity in the VBO (*center*), with annotation properties (*surrounding*).

A compound entity represents a structure in a species, as represented in Fig. 4.

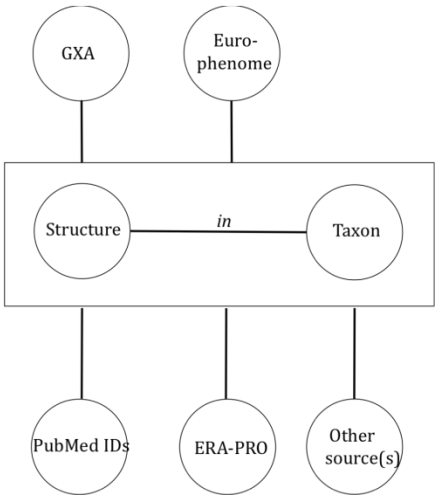


**Figure 4.** The data structure of a compound entity, representing a structure *in-a* taxon, and the annotation properties that document the evidence of existence of that compound entity: PubMed, ERA-PRO, Gene Expression Atlas (GXA), Europhenome, and so forth.

Compound entities also have annotation properties representing the source of the assertion that

$$\forall \text{Structure-in-Taxon} \rightarrow \{\exists(x \in \text{Taxon}) : \quad x \textit{ has-part } \text{Structure}\} \qquad (1)$$

The following relationships operate on compound entities:

**Relationships.** These relationships in the VBO describe homology relationships among compound entities.

**1. Homologous-to**. The relationship homologous-to describes a 1:1 (injective) and onto (surjective) (thus, bijective) structural similarity based on evolutionary relationship between a structure in one species and a structure in a second species, as in Fig. 5.

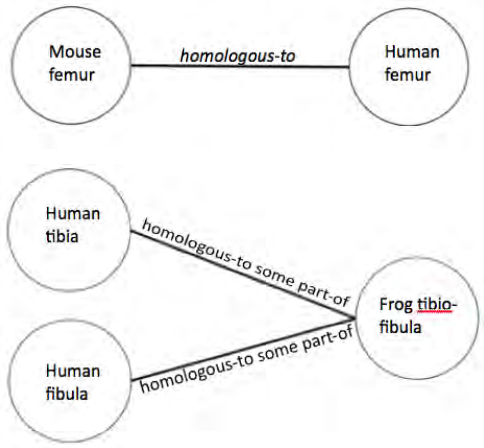


**Figure 5.** Mouse femur *homologous-to* human femur (*top*), human tibia *homologous-to* some part of frog tibiofibula and human fibula *homologous-to* some part of frog tibiofibula. (*bottom)*.

While not definitively ruling out a genetic event that occurred after the species' separation from the MRCA, a 1:1 and onto mapping tends to be indicative of evolutionary conservation. When the mapping by term name or structure is not itself 1:1 and onto with a homologous structure (which can indicate an evolutionary event), there may be a 1:1 and onto mapping from a structure in one species to some part of the homologous structure in the second species.

**2. Not-homologous-to.** The need to explicitly encode a negative relation in VBO is a consequence of the combination of open-world reasoning and the history of comparative anatomy. The *not-homologous-to* relationship can be one-to-many.

The naming of structures in one species, based on analogy ("wing" in insect, pterosaur, bird, and bat) or homoplasy (panda's "thumb") to a non-related structure in a different species, muddies the waters tremendously for determining homology based on lexical matching. Haendel *et al* (accessed 10 April 2011) have remarked upon the case of the frontal bone in the zebrafish being homologous to the prefrontal bone, and not the frontal bone, in humans. The problem is magnified tremendously by the use of important vertebrate skeletal terms to refer to segments in insects, and that is in turn magnified by the importance of those insects, such as *Drosophila*, in the comparative medical research community. Table 1 presents an illustration of the problem for some representative skeletal structures.

| Structure | Invertebrate taxa | Refers to | Vertebrate taxa | Refers to |
|---|---|---|---|---|
| acetabulum | parasitic worms (trematodes), leeches | the sucker (feeding) | tetrapods (4-limbed vertebrates) | concave pelvic surface meeting femur at hip joint |
| femur | insects | leg segment | tetrapods (4-limbed vertebrates) | long bone in leg |
| trochanter | insects | leg segment | tetrapods (4-limbed vertebrates) | part of thigh bone |
| coxa | insects | leg segment | tetrapods (4-limbed vertebrates) | hip (either joint or anatomical region) |
| tibia | insects | leg segment | tetrapods (4-limbed vertebrates) | long bone in leg |

**Table 1.** Representative identically-named vertebrate and invertebrate non-homologous structures in PubMed.

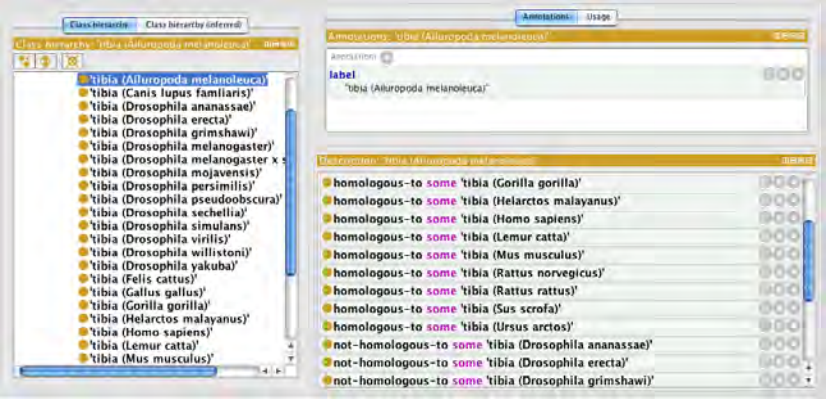


**Figure 6.** Sample entities and relations in the VBO.

The open-world assumption means that any lexical-matching tool used to populate VBO or any other homology-based ontology will create a high number of false positives based on lexical matches such as these, since – under that assumption – there could, in future, be insect structures that are homologous to their vertebrate homonyms. This possibility, permitted under the open-world assumption, actually violates a biological constraint on homology. To prevent those false positives, to provide metaknowledge for future data mining tools, to mitigate human error in creating axioms containing NOT and a vast number of disjoints in Protégé, and to make reasoning more tractable, we have explicitly encoded the *not-homologous-to* relationship, along with any necessary invertebrate species, in the VBO in order to definitively rule out that possibility. Although it is not an ideal solution, it is a workable compromise, given the state of the art and the scope of the problem. We do not represent a phylogeny of invertebrates, nor do we make any statements about the relationships among *not-homologous-to* relationships, as those are clearly out of scope, so *not-homologous-to* forms a simply-connected graph, and not a maximally-connected one.

Entities and relationships as described above provide the content of VBO; Fig. 6 shows representative entries for a vertebrate tibia, and its relationships to other vertebrate and invertebrate tibiae.

VBO was initially populated by a combination of manual and automated approaches. Annotations from the Gene Expression Atlas [7, 8], ERA-PRO [9], Europhenome [10, 11], and Phenoscape [12] databases provided anatomical structures and species for the ontology. Additionally, Uberon and FMA provided structures for VBO. These structures and species were manually added to the OWL file in Protégé. For VBO 1.0, inclusion of a taxon or structure class in one of the above databases or ontologies was considered sufficient evidence of existence to include it in the ontology. The use of these sources also uncovered some major discrepancies between how major ontologies, such as FMA and Uberon, represent anatomical classes versus the way the terms corresponding to those classes are used in real-world contexts [1]. Those considerations influenced how we developed composition of compound entities, for example, and will continue to inform future versions of VBO.

Some preliminary data-mining of PubMed abstracts was carried out to populate VBO. Python scripts which searched PubMed iteratively through a list of structures from FMA and Uberon were used to collect abstracts of articles that contained musculoskeletal terms with references to non-human vertebrate species. Reference to a structure in a species in an abstract was considered evidence of a compound entity (Equation [1]), and the compound entity was evaluated for homology to that structure in humans or another species. This evaluation was carried out on the basis of available evidence – reference material, journal articles, and so forth. The provenance of the evidence was recorded as well. This direct connection to evidence for homology statements is a unique strength of VBO.

When sufficient evidence established the homology between the compound entities, the triple

<Compound-entity-1>*relationship*<Compound-entity-2>

was recorded as a "pairwise mapping" in a spreadsheet. A set of Java tools was developed to transform the spreadsheet's pairwise mappings into classes and relationships in Protégé, and to create the relationships among the nodes of the maximally-connected graph. These generated relationships are marked evidentially as inferred from homology.

A beta version of VBO has been successfully integrated into the EFO to support cross-species comparisons of orthologous genes in homologous tissues through the Gene Expression Atlas interface.

## 3 Future work

We plan to continue integrating VBO into the Gene Expresson Atlas via EFO, and improving the functionality and the interface. We will add more sophisticated analysis of evidence that can work with the Phenoscape taxonomy of evidence model for easier integration and sharing of data. More complex systems which present more complicated modeling challenges, and incorporating developmental structures as well as adult structures are also areas into which we plan to extend VBO.


### Acknowledgements

We thank the members of the VBO Scientific Advisory Board, Jonathan Bard, Claudio Stern, Martin Ringwald, and Monte Westerfield, who guided and supported this project. In addition, we thank Hilmar Lapp, and the participants in our community workshops, who provided valuable feedback and suggestions.

**Funding:** Biotechnology and Biological Sciences Research Council (grant #BB/G022755/1), and European Molecular Biological Laboratory core funding.


### References


1. Travillian RS, Adamusiak T, Burdett T, Gruenberger M, Hancock J, Mallon AM, Malone J, Schofield P, Parkinson H. Anatomy Ontologies and Potential Users: Bridging the Gap. Biomed Semantics, forthcoming.
2. Haendel M, Gkoutos G, Lewis S, Mungall C. Uberon: towards a comprehensive multi-species anatomy ontology. Available from Nature Preceedings (2009).
3. Haendel MA, Neuhaus F, Osumi-Sutherland D, et al. CARO - The Common Anatomy Reference Ontology. In: *Anatomy Ontologies for Bioinformatics, Principles and Practice* Albert Burger, Duncan Davidson and Richard Baldock (Eds.), 2007.
4. Osumi-Sutherland D. personal communication, 2010.
5. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 2010 Apr 15;26(8): 1112-8.
6. Rosse C, Mejino JL. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform. 2003 Dec;36 (6):478-500.
7. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A. Gene Expression Atlas at the European bioinformatics institute. Nucleic Acids Res. 2010 Jan;38 (Database issue):D690-8.
8. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U, Brazma A. ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res. 2009 Jan;37(Database issue):D868-72.
9. Birschwilks M, Gruenberger M, Adelmann C, Tapio S, Gerber G, Schofield PN, Grosche B. The European radiobiological archives: online access to data from radiobiological experiments. Radiat Res. 2011 Apr;175(4):526-31.
10. Morgan H, Beck T, Blake A, Gates H, Adams N, Debouzy G, Leblanc S, Lengger C, Maier H, Melvin D, Meziane H, Richardson D, Wells S, White J, Wood J; EUMODIC Consortium, de Angelis MH, Brown SD, Hancock JM, Mallon AM. EuroPhenome: a repository for high-throughput mouse phenotyping data. Nucleic Acids Res. 2010 Jan;38(Database issue):D577-85.
11. Mallon AM, Blake A, Hancock JM. EuroPhenome and EMPReSS: online mouse phenotyping resource. Nucleic Acids Res. 2008 Jan;36(Database issue):D715-8.
12. Dahdul WM, Lundberg JG, Midford PE, Balhoff JP, Lapp H, Vision TJ, Haendel MA, Westerfield M, Mabee PM. The teleost anatomy ontology: anatomical representation for the genomics age. Syst Biol. 2010 Jul;59(4):369-83.