



# **Virtual Physiological Human Network of Excellence**

**Grant Agreement: 223920**

## **VPH ToolKit Guideline Document**

**Topic: Ontological Annotation**

**Version 1.0**

**22-Mar-11**

## Document Information

<b>IST Project Num</b>	FP7 – 2007 - ICT - 223920	<b>Acronym</b>	VPH NoE
<b>Full title</b>	Virtual Physiological Human Network of Excellence		
<b>Project URL</b>	http://www.vph-noe.eu		

<b>Document</b>	<b>Number</b>	G04	<b>Title</b>	Guidance (Ontological Annotation)
-----------------	---------------	-----	--------------	-----------------------------------

<b>Status</b>	Version. 1.0	Final <input checked="" type="checkbox"/>
<b>Dissemination Level</b>	Public <input checked="" type="checkbox"/> Consortium <input type="checkbox"/>	

<b>Authors (Partner)</b>	UCL	Nash	
	UOXF	Cooper	
	CNRS	Friboulet, Cervenansky	
	INRIA	Sermesant, Bleuzé	
	UPF	Martelli, Omedas	
	UOA	Britten	
	EBI	de Bono	
	USFD	Fenner	
	IMIM	De Fabritiis, Giorgino	
<b>Responsible Author</b>	B de Bono	<b>Email</b>	bdb@ebi.ac.uk
	<b>Partner</b> EBI	<b>Phone</b>	0044 7743949508

<b>Abstract (for dissemination)</b>	This document provides guidance on the attributes required of VPH NoE ToolKit content.
-------------------------------------	--

*The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. Its owner is not liable for damages resulting from the use of erroneous or incomplete confidential information.*

<b>Version Log</b>			
<b>Issue Date</b>	<b>Version</b>	<b>Author</b>	<b>Change</b>
30-Jul-10	0.1	Bernard de Bono	First Issue
18-Feb-11	0.2	Bernard de Bono	Second Issue
22-Mar-11	1.0	WP3	First public release

## Table of Contents

<b>EXECUTIVE SUMMARY .....</b>	<b>4</b>
<b>Introduction .....</b>	<b>5</b>
<b>Guideline Topic: VPH Resource Documentation.....</b>	<b>6</b>
<b>Underlying Concepts.....</b>	<b>7</b>
<b>Interactions and Dependencies .....</b>	<b>27</b>
<b>Applicable Legislation.....</b>	<b>29</b>
<b>Standards and Standards Bodies .....</b>	<b>30</b>
<b>Characteristics .....</b>	<b>31</b>
<i>Overview.....</i>	<i>31</i>
<i>Knowledge Representation Languages .....</i>	<i>31</i>
Logic-based knowledge representation languages .....	31
Web languages .....	32
Classes of tools.....	32
Table.....	34
<b>Methods of Verification .....</b>	<b>38</b>
<b>Ownership .....</b>	<b>39</b>
<b>Training.....</b>	<b>40</b>
<b>Maintenance.....</b>	<b>41</b>
<b>Ranking .....</b>	<b>42</b>
<b>Documentation, Reporting Templates.....</b>	<b>43</b>
<b>Further Information .....</b>	<b>44</b>
<b>References .....</b>	<b>45</b>

## EXECUTIVE SUMMARY

There are several such guideline documents in this series, covering the full range of issues affecting content providers. They are being developed over a period of time and, once finalized, these guideline may be bound together into a single VPH NoE resource. The present guidelines concern knowledge representation in relation to the VPH and, in particular, as an overall context for ontological annotation. These guidelines were developed in collaboration with the RICORDO VPH STREP (<http://vph-ricordo.eu/>).

## Introduction

This document is one of a series that together build to form a complete guide to the ideal content and presentation of materials offered for distribution via the Virtual Physiological Human Network of Excellence ToolKit Portal. The full set of Guideline Documents is summarised below.

<b>Guidance Area</b>	<b>Description</b>
Tool characterisation	The attributes important for inclusion in the documentation of Tools, including performance validation
Model characterisation	The attributes important for inclusion in the documentation of Models, including performance validation
Data Characterisation	The attributes important for inclusion in the documentation of Data
Ontological Annotation	The significance, benefits and methods of ontological annotation of ToolKit content
Interoperability	Key attributes concerning the additional specification of predominantly tools and models that will allow operation in a multistage workflow alongside other items of ToolKit content
Ethico-legal issues, provenance	The inherited responsibilities that are attached to any item of ToolKit content – perhaps particularly data – including legal, ethical and territorial restrictions
Licensing	The conditions that apply to the legitimate use of the content from a commercial and intellectual property standpoint
Usability & Training	The factors that are important for the easy use and ready acceptance of ToolKit content, taking into account the environment, the likely users and the need for interoperability. Additionally, the nature of training facilities of all types appropriate to particular content categories.

## Guideline Topic: VPH Resource Documentation

The following guidelines address the fundamental VPH objective to support the verifiability and re-use of data and model resources (DMRs). A key requisite for this goal is to have a coherent and communal strategy for DMR documentation that is both human- and machine-readable. Two key aspects of DMR documentation will be discussed in these guidelines, namely: (i) DMR metadata to hold resource annotation, and (ii) the use of Knowledge Representation and Ontologies in annotation.

**Motivation:** In practice, the effort employed by the community in providing detailed annotation to a DMR is closely influenced by the expectation of a resource being shared. Therefore, the limits imposed on the distribution of a resource (typically for commercial, legal, confidentiality, but also interoperability, reasons) tend to curb directly the quality and machine readability of the corresponding documentation: after all, why document a DMR if the resource cannot (or will not) be accessed by third parties? In response to this type of obstacle, these guidelines identify ways by which the distribution of, and access to, DMR documentary metadata is uncoupled from that of the DMR *per se*. In consequence, therefore, this approach allows VPH users to make well-defined details of their work known to the community, while satisfying the constraints and obligations of confidentiality that sensitive clinical or commercial work often entails. For example, this approach will make it easier for the community to be aware of the presence of datasets or models that may be relevant to some biomedical objective, despite the fact that the actual DMRs themselves may not be in the public domain.

**Overall approach:** In the context of the VPH effort, a key role for Knowledge Representation (KR) is to support the semantic interoperability (SI) between data and model resources. Methodologies in biomedical SI aim to provide a standardized and machine readable context to scientific data resources of biological importance. In so doing, such methods ensure the consistent and automated interpretation of different resources in a co-ordinated manner. The aim of this document is to provide guidelines to improve model integration and data re-utilization through a dedicated SI framework. The SI framework referred to in this document is the one developed through the RICORDO VPH project ([www.vph-ricordo.eu](http://www.vph-ricordo.eu)), in collaboration with the VPH Network of Excellence.

## Underlying Concepts

In response to obstacles to interoperability there is a growing investment by the VPH community in the knowledge-based automation of (scientific and clinical) data and model resource (DMR) management. A key objective for this effort is to allow experts to search, classify, integrate and share information about DMRs based on the biomedical knowledge they represent. A requisite for this goal is to develop a common representational framework for physiological knowledge. Such a framework can provide both the vocabulary and means for formally characterizing DMRs and thereby allow the DMR's knowledge to be represented and managed in an automated and biomedically meaningful manner. This section illustrates some of the basic concepts involved in the development of semantic interoperability for VPH DMRs.

VPH resources carry information that is fit for some purpose. At the same time, they also tend to bear significant amounts of implicit knowledge: for example, the intended biological interpretation of a free-text label associated with a DMR element. Examples of an element in a DMR include (i) a data column in a spreadsheet or database table of clinical trial data, (ii) a variable in a model, (iii) a specific region in a radiology image, or (iv) a pathology term in a list of disease names. (See **Figure 1**: DMRs are depicted in the lower band of the diagram – each green DMR rectangle shows a single element as a grey box, and the associated label is depicted as a pink dot within the element. In this diagram, each element is also associated with a URI, shown as a yellow dot).

For a set of DMRs to achieve SI it is necessary that such implicit knowledge is rendered explicit and machine readable. For instance, SI allows the meaning of a column on a spreadsheet to be automatically compared to the meaning of a variable in a model.

As part of the DMR documentation effort, a DMR element is typically associated with a free-text label (e.g. the text in the topmost cell of a column in a spreadsheet, or the column name of a database table schema). In this section the basis for SI is discussed in the context of interpreting label descriptors associated with elements in biomedical DMRs in a more formal manner.

For knowledge within a DMR element to become more explicit, therefore, the meaning of the free-text label associated with an element needs to be rendered machine readable. This machine-readable meaning is stored in the metadata associated with the DMR element (see **Figure 2** for an example). In this section, we discuss the nature of metadata and the methods by which metadata is imbued with meaning through the process of annotation with ontology terms (see Figures 1a and 1b for an overview).

The basis for SI between DMRs is the ability to automatically read metadata and compare the meaning it bears. The explicit knowledge that allows such comparisons to be made automatically will be drawn from standard reference ontologies.

## 1 What is Metadata?

In computational biology, discussions about DMR documentation typically allude to the notion of metadata. One type of documentary metadata refers to documentation material that is linked to a corresponding DMR element indicating how the actual content of that element should be interpreted: such metadata will be referred to as semantic metadata because it conveys meaning. By explicitly representing the meaning implicitly held by free-text labels associated with DMR elements, this type of metadata adds semantic features to a resource, and provides a formal and independent guide as to what a particular DMR element represents.

Semantic metadata has two key characteristics, namely (i) semantic content (*i.e.* the concepts used to provide machine readable meaning)<sup>1</sup> and (ii) encoding (*i.e.* the manner in which metadata content is rendered machine readable).

## 2 What is Semantic Interoperability?

A set of DMRs is said to be semantically interoperable when the elements of all members of this set can be consistently related and navigated through their explicit meaning. For this to be achievable, all DMRs in this set should share the same machine readable metadata standard.

Such a standard would therefore entail the use of the same (i) set of reference ontologies as a source for semantic content, and (ii) machine-readable language for its metadata encoding. Figures 1a and 1b illustrate the key relationships between ontologies, DMRs, as well as the semantic metadata that conveys the link between the two.

The goal of achieving SI for a set of DMRs is motivated by the need to automate the coherent interpretation of DMR content over a large number of diverse DMRs. A key result of attaining this goal is the ability to automatically identify DMRs that are related to each other solely on the basis of their metadata documentation notwithstanding any differences in format or accessibility the various DMRs may have.

These SI aims, therefore, require that metadata is fully machine readable and interpretable.

---

<sup>1</sup> It is important to distinguish DMR information content (*i.e.* the actual data stored in a resource) from the semantic metadata content (*i.e.* the ontology terms providing explicit meaningful descriptions *about* the DMR element). In practice, there is no formal relationship between semantic metadata for a DMR element, and any free-text label that may implicitly describe its meaning in human readable form (see Figure 1 for an illustrative example).

To this end, the community in the systems biology domain, for instance, is addressing this goal by introducing a common format for annotating their data and models. The Minimal Information Required In the Annotation of Models (MIRIAM) is a set of guidelines for annotation and curation processes of computational models to facilitate their exchange and reuse [1]. A number of VPH resources are already annotated using MIRIAM, such as SBML [2] and CellML [3]. The Model Format OWL (MFO) is another effort within the systems biology community that is focused on data integration by capturing the SBML structure of biological annotations in OWL-DL to support reasoning, validation, and querying of SBML models [4].

### **3 Why is Semantic Interoperability important for scientific and clinical research?**

The practice, research and industrialization of biomedicine generate large quantities of data, often at great risk or expense. In addition, the study of this data typically employs the use of mathematical models based on discrete (*e.g.* statistical) or continuous (*e.g.* calculus) methods.

In turn, the validity and robustness of a model, and the results it produces, largely depend on the quality and quantity of data that is applied in its construction and usage. One of the key biomedical research applications of SI, therefore, is to help the community find datasets that are relevant to their modelling goals (*e.g.* see Box 1). Ideally, having found the relevant datasets, the same SI framework would be transferable to the workflow that handles data and model interaction. When the same semantic metadata standards are applied across the board, both datasets and models achieve SI.

The cardiac scenario described in Box 1 provides an insight into the importance of SI for data resources relevant to drug discovery and, in particular, the ability to map to reference concepts (such as anatomy) in a standardized manner. There are two key requirements associated with this standardization, namely: the standardization of (i) semantic content, and (ii) the encoding of associations between terms in metadata (*i.e.* semantic metadata content) and DMR elements.

In order to describe more fully the role SI plays in biomedical data management, an example from cardiology research is outlined below in the context of drug discovery. This idealized scenario is set in an institution with a centralized data management system.

In this example, the drug discovery process starts with a compilation of molecular pathway data that is relevant to the causes of congestive heart failure (CHF). A considerable number of pathologies, affecting different cardiac structures, may lead to CHF. The main categories of cardiac pathology relevant to CHF are: autoimmunity, myopathy, electrophysiological interference, accumulation disease, ischaemia, infection, malignancy and toxicity.

A first step of this compilation, therefore, could involve manually selecting from a list (DMR1 in Figure 1b) of names of pathologies that are germane to this study. A second step would focus on the assembly of available protein pathway and molecular interaction data based on available gene expression data (DMR2 in Figure 1b) derived from cardiac structures affected by these pathologies.

In practice, the key data management feature sought in this particular case is SI between a list of pathologies (DMR1) and gene expression data (DMR2). The SI goal is to find relationships between the knowledge in the two DMRs on the basis of a shared cardiac anatomy representation. In short, the anatomy knowledge that applies to elements in DMR1 should be comparable to the one applied to elements in DMR2. However, if the documentation of the two resources only relies on the use of free-text labels, it is unlikely that the mapping to anatomical concepts may be inferred in a straightforward manner. For example, Libman–Sacks endocarditis (found in the list of pathologies of DMR1) is a manifestation of Systemic Lupus Erythematosus that typically involves the Mitral Valve. As this relation is not explicit in the free-text name of the pathological condition, the mapping between the label of the condition and the corresponding anatomical concept is not inferable.

Box 1. Cardiac drug discovery SI scenario

#### **4 Why are biological ontologies useful as reference knowledgebases for semantic content in DMR metadata?**

A number of biomedical vocabularies are supported by the community to provide standardized free-text labels to describe clinical data elements (e.g. SNOMED-CT, ICD-10, MedDRA, and the CDISC Terminology). In some cases, such vocabularies are primarily oriented to support human readability, and largely consist of either a flat list (e.g. CDISC) or a single hierarchy (e.g. MedDRA) of free-text terms controlled via some editorial process to

avoid semantic redundancy (hence the use of the phrase “controlled vocabulary”).

While useful for human assessment, such vocabularies provide limited scope for automated processing of machine readable knowledge. For example, it is difficult to automatically infer that the MedDRA Lower-Level Term ‘Itchy Rash’ and the CDISC CodeList Name ‘Skin Classification’ both relate to some property of the skin. A DMR that bears CDISC free-text labels and a DMR that has MedDRA labels are not semantically interoperable, and comparisons between the two DMRs may not be made automatically.

A second limitation to consider applies to the ability to automatically infer relationships between free-text terms within the same vocabulary. For example, a clinical expert may immediately spot the link between the CDISC term ‘Amniotic Fluid’ (from the ‘Specimen Type’ CodeList) and the term ‘Intra-amniotic’ (from the ‘Route of Administration’ CodeList). The implicit knowledge of these two free-text terms, however, is not amenable to machine processing. Therefore, the ability to compare DMR elements described by the same vocabulary set is still somewhat restricted. The type of SI such vocabularies afford to a set of DMRs is typically limited to the exact matching of labels that use precisely the same term (*i.e.* it is possible, of course, to match two DMR elements that are labelled with exactly the same vocabulary term).

Reference ontologies are able to provide (a) human- and machine-readable terms to convey biological interpretation of DMR content, as well as (b) a corresponding stable identifier (ID) space for machine processing<sup>2</sup>. However, in contrast to most biomedical vocabularies, ontologies provide explicit machine readable knowledge. An example of machine readable knowledge is the classification of types of anatomical organ (also known as a subsumption graph – see Figure 1c). Another is the hierarchy of parts of an organ such as the heart (also known as a partonomy graph, also shown in Figure 1c – for a discussion about composite terms see section 6). By explicitly representing knowledge as well-defined concept nodes and relation edges between such concepts, it is possible to compare two concepts from the same ontology with precision. The anatomy ontology example in Figure 1c shows a machine-readable graph that represents how some anatomical structures are components of other structures. For instance, the term “heart” (with identifier 71) has a part called “left ventricle” (ID: 75) which, in turn, has a part called “Mitral Valve” (ID: 77). The standard relation in this

---

<sup>2</sup> Stable IDs are essential to deal with synonyms (*i.e.* the syntactic variability of a semantic entity represented by an ontology term) and ensures that, for example, the synonymous cardiology free-text labels “mitral valve”, “bicuspid valve” and “left atrioventricular valve” map to the same semantic entity ID. While the syntax of the label of a semantic entity is not relevant to ID-based machine matching, synonyms are a crucial for ease of human interpretation

case is *part\_of*. While the terms labelling the concept nodes are also human readable, the graph structure *per se* provides a rich source of machine-processable knowledge (*i.e.* such a graph provides a well-defined structure to meaning).

In the context of the cardiac scenario in Box 1, therefore, a data management system that makes use of an ontology-based knowledgebase standard could simplify and expedite the search for all cardiac structures by automatically parsing the graph of parts and all their synonyms. The process of automated traversing of, and inference from, this type of graph is known as **reasoning over an ontology**. This type of automated reasoning is simply not possible with list-based vocabularies. Indeed, one step to achieving SI for a set of list-based vocabularies is to map their terms onto the same set of standard reference ontologies.

## 5 Which biological ontologies are relevant to the VPH?<sup>3</sup>

The recommendations cover the following areas:

- I. Biological structure
- II. Biological processes
- III. Qualities
- IV. Classes of entities in Experiments, Modeling and Simulation

These four key categories are discussed below.

### I. Biological Structures

#### a. The Foundational Model of Anatomy [1-2]

The Foundational Model of Anatomy (FMA) is an ontology of human anatomy and will serve as the VPH Core Vocabulary for annotating and mapping macroscopic anatomical entities (*i.e.*, multicellular entities, and larger). It is a reusable and generalizable resource of deep anatomical knowledge which is designed to meet the needs of any knowledge-based application that requires structural information. It is widely used in many informatics research projects as well as in clinical applications. To date, we have documented over 2100 registrations to download the FMA from its website, drawing interest from both private and public sectors, including local and international government agencies and institutions, from >150 universities in 51 countries, >90 US universities and colleges, large commercial companies such as General Electric, Siemens, GSK and Philips, non-profit organizations such as the Mayo Foundation and the Allen Brain Institute, health care providers such as Cleveland Clinic, Texas Healthcare Network and private

---

<sup>3</sup> This section is based on the RICORDO deliverable report 2.1 [[LINK](#)]: Establish the CORDO set of dictionaries

citizens. The FMA has served to provide and continues to provide 1) ontological framework to a number projects such as the Virtual Soldier Project, RadLex and Terminologia Anatomica (FICAT), 2) ontology template for other ontologies such as Common Anatomy Reference Ontology (CARO), Cell ontology (CL) and Aneurist; 3) anatomy content enhancements to RadLex, the Disease Ontology (DO), the Human Phenotype Ontology (HPO) and Skin Ontology; 4) ontology support for various applications such as BodyParts3D (Japan), AnatomyLens (search engine from IBM), and Autodesk (ergonomics modeling); 5) ontology for alignment and mapping projects involving the anatomy axis of SNOMED-CT, GALEN anatomy and neuro-imaging terminologies such as Talairach, AAL and FreeSurfer; and 6) ontology test bed for auditing methodology (NJIT), tool development and knowledge engineering by projects such as K-CAP (knowledge capture challenge), LexGrid (distributed network of resources), Microsoft (mappings) and TopQuadrant (scalability of tools). The FMA has also been adopted as the anatomy standard by the European working group (CEN) of ISO for health informatics. The FMA is the primary anatomy reference ontology source for the composite annotations to be used in the VPH.

**b. The Edinburgh Mouse Atlas (EMAP) and the Mouse Anatomical (MA) Dictionary**

The domain of the MA is the adult mouse, while that of EMAP is mouse embryo anatomy. EMAP and MA have a close, formal collaboration and are working towards a unified anatomy of the mouse. EMAP was one of the first anatomy ontologies – perhaps the first to be implemented and used widely. The MA was built on the EMAP model and discussions continue between the two groups about all aspects of anatomical representation. Both are used to annotate gene expression in the GeneExpression database (GXD) and are widely used elsewhere in the mouse community, in EMAGE, EUREXPRESS, GUDMAP, EuReGene, etc. Some other gene expression resources use anatomy ontologies related to or based on these (for example, GenePaint).

The EMAP anatomy ontology is organized into 26 developmental stages, referred to as Theiler stages (TS1–TS26) [8]. Each stage is primarily organized as a structural part–of tree. The tissues represented by subnodes of a node in the tree are intended to be non-overlapping (exclusive) and complete, i.e. they describe all distinct parts of the parent tissue. (The term ‘tissue’ is used in a very generic way, meaning both: whole anatomical structures as well as specific tissues.) For example, the trophectoderm consists of the mural trophectoderm and the polar

trophectoderm, which are distinct from each other and are the only parts of the trophectoderm. The requirement for the anatomy ontology to be non-overlapping and complete was a design decision by the biologists who developed it, rather than a constraint needed for its computational formalisation.

In addition to the ontology, the EMAP atlas contains 3D voxel-based reconstructions of the embryos at the different developmental stages and mappings between concepts in the ontology and their corresponding voxel sets in the 3D model. This enables the integration of spatially mapped data with data labelled using the terms from the anatomy ontology.

EMAP was built on a developmental stage-by-stage basis. The currently ongoing revision of the EMAP ontology aims to (i) thoroughly incorporate class (Is\_a) relations in order to represent the development and adult state of each structure by recognizing the identity of the continuant structure rather than focusing on the snapshot views through the separate windows of successive developmental stages, and (ii) incorporate properly-defined temporal relationships.

In the long term, the goal is to have a single ontology that includes everything in the mouse, including the adult.

#### **c. The Cell Type Ontology [9]**

The Cell Ontology is an ontology for cell types that covers the prokaryotic, fungal, animal and plant worlds aiming to provide descriptions of other biological objects, such as gene-expression data and facilitate interoperability. It includes over 700 cell types. These cell types are classified under several generic categories and are organized as a directed acyclic graph. The ontology is available in the formats adopted by the Open Biological Ontologies umbrella and is designed to be used in the context of model organism genome and other biological databases.

#### **d. The Gene Ontology Cell Component [10]**

The cellular component ontology is one of the three vocabularies of the Gene Ontology that provides structured controlled vocabularies for the annotation of gene products with respect to their cellular location at the levels of subcellular structures and macromolecular complexes. Examples of cellular components include nuclear inner membrane, with the synonym inner envelope, and the

ubiquitin ligase complex, with several subtypes of these complexes represented. Generally, a gene product is located in or is a subcomponent of a particular cellular component. The cellular component ontology includes multi-subunit enzymes and other protein complexes, but not individual proteins or nucleic acids. Cellular component also does not include multicellular anatomical terms.

**e. The Protein Ontology [11]**

PRO is a formal representation of protein objects, providing both descriptions of these objects and the relationships between them. PRO encompasses a sub-ontology of proteins based on evolutionary relatedness (ProEvo) and a sub-ontology of the multiple protein forms produced from a given gene (ProForm). PRO is interoperable with other OBO Foundry ontologies--such as the Sequence Ontology (SO) and the Gene Ontology (GO)--that provide representations of protein qualities. This interoperability facilitates cross-species comparisons, pathway analysis, disease modeling, and the generation of new hypotheses through data integration and machine reasoning.

**f. Chemical Entities of Biological Interest [12]**

Appearing in a wide variety of contexts, biochemical 'small molecules' are a core element of biomedical data. Chemical ontologies, which provide stable identifiers and a shared vocabulary for use in referring to such biochemical small molecules, are crucial to enable the interoperation of such data. One such chemical ontology is ChEBI (Chemical Entities of Biological Interest), a candidate member ontology of the OBO Foundry. ChEBI is a publicly available, manually annotated database of chemical entities and contains around 18000 annotated entities as of the last release (May 2009). ChEBI provides stable unique identifiers for chemical entities; a controlled vocabulary in the form of recommended names (which are unique and unambiguous), common synonyms, and systematic chemical names; cross-references to other databases; and a structural and role-based classification within the ontology. ChEBI is widely used for annotation of chemicals within biological databases, text-mining, and data integration.

**II. Biological Events**

**a. GO Biological Process [10]**

The Biological Process (BP) ontology is one of the three vocabularies of the Gene Ontology that describes biological goals accomplished by one or more ordered assemblies of molecular functions. High-level processes such as 'cell death' can

have both subtypes, such as 'apoptosis', and subprocesses, such as 'apoptotic chromosome condensation'. We note here that A biological process is not equivalent to a pathway; at present, GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

**b. Mammalian Pathology [13]**

The mammalian pathology ontology provides a description mammalian pathology phenotypes and incorporates known mouse pathologies hierarchically organised as "instances of" pathological processes. It distinguishes between pathological anatomical entities and pathological process and can be used in conjunction with the PATO ontology to provide pathological descriptions.

**III. Qualities**

**a. GO Molecular Function [10]**

The Molecular Function (MF) is one of the three vocabularies of the Gene Ontology that describes activities, such as catalytic or binding activities, at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when or in what context the action takes place. Examples of individual molecular function terms are the broad concept 'kinase activity' and the more specific '6-phosphofructokinase activity', which represents a subtype of kinase activity.

**b. PATO [14-17]**

RICORDO aims at the standardization of the representation of physiological variables and data. There has been a substantial effort in the general domain of phenotype data representation over the years and a system has been developed to allow for a high degree of expressivity to capture the wide range of phenotypes observed across a variety of organisms and types of investigation. This system is based on the PATO framework that provides an integration platform that allows consistent phenotype descriptions and data (including physiology data) by allowing a so-called post-composition methodology that involves the combination of a set of domain ontologies with their appropriate qualities that are provided from an ontology of qualities termed PATO. The PATO framework has the ability of expressing both qualitative and quantitative descriptions as well as providing

logical definitions for so-called precomposed complex descriptions, such as the ones available from the various phenotype ontologies that exist for several domains. Within the field of human physiology a set of domain specific and non specific ontologies will be required to be combined with the PATO ontology.

**c. Ontology for Physics in Biology [3-7]**

The Ontology of Physics for Biology (OPB) is an ontology of physical concepts, properties, and laws that a VPH Core Vocabulary for physics-based biosimulation modeling and data analysis. It will be used for annotating and mapping analytical model variables and experimental datasets (e.g., pressure of blood in aorta, MPK-gene expression rate in renal tubule cells) across all structural levels (i.e. molecular to whole organism) and time-scales. A relatively new ontology, the OPB is being co-developed in conjunction with and “stress-tested” against advanced model annotation software (SemGen, SemSim) for annotating, merging, and encoding biosimulation models using composite annotations. The OPB has been developed by the UoW team in the Web Ontology Language (OWL) based on knowledge representational principles developed and proven by the UoW FMA development experience. OPB is unique amongst ontologies with biophysical content in that its content is based strictly, and deeply, on the principles the principles of classical physics and thermodynamics. Such a principled approach is critical to resolving ambiguities of usage and naming across multiple biomedical and biophysical disciplines.

**IV. Classes of entities in Experiments, Modeling and Simulation**

**a. Units Ontology [14]**

The Units Ontology (UO) was developed in order to allow for the PATO framework to provide a means to record both qualitative and quantitative information. Although ontologies provide qualitative partitions of the kind of entities we find in nature, in biology we also need to record quantitative information relating to phenotypic observations (e.g. a quality such as "size"). Therefore, as a necessary adjunct to PATO, a Units Ontology [UO] has also been built and published to the OBO site. The UO was assembled from existing text-based sources and constructed according to OBO Foundry principles. It is maintained in OBO format. The UO includes 264 terms, all of which are defined. These definitions are consistent with those of the Unified Code for Units of Measure [UCUM]. Wherever possible definitions from the National Institute of

Standards and Technology [NIST] are used.

#### **b. Ontology for Biomedical Investigation**

The Ontology for Biomedical Investigations (OBI) addresses the need for controlled vocabularies to support integration of experimental data, a need originally identified in the transcriptomics domain by the Microarray Gene Expression Data Society (MGED), which developed the MGED Ontology as an annotation resource for microarray data. In response to the recognition of convergent needs in areas such as protein and metabolite characterization, this effort was broadened to become what was initially known as FuGO (Functional Genomics Investigation Ontology). FuGO was further expanded in 2006 to include clinical and epidemiological research, biomedical imaging and a variety of further experimentation domains to become what is today OBI, an ontology designed to serve the coordinated representation of designs, protocols, instrumentation, materials, processes, data and types of analysis in all areas of biological and biomedical investigation. OBI employs the PATO and UO ontologies for the description of phenotypes.

RICORDO is also examining the use of upper level ontologies that describe very general concepts across all biological domains, thus supporting very broad semantic interoperability between a large number of ontologies. Candidates include BFO, DOLCE, GFO. A further key goal of RICORDO, in the context of WP5, is the use of predefined well formed ontological relationships. To this end, RICORDO is examining the use of the Relation Ontology proposed by the OBO foundry.

## **6 What are composite ontology terms?<sup>4</sup>**

While well-developed reference ontologies are readily available to describe basic biological concepts (*e.g.* structure, processes and their qualities) in a consistent manner, most biomedical data and models tend to represent more complex concepts as well. An example of a complex concept from physiology is “venous return”, which refers to the rate of blood flowing from the central systemic veins back to the right atrium of the heart. In such a case, no single ontology from the above reference sets can provide a term that completely represents the precise meaning of that semantic entity.

In order to maintain a link between basic biological concepts and complex one, terms from

---

<sup>4</sup> This section is based on the RICORDO deliverable report 5.1 [\[LINK\]](#): Report on composite annotation architecture

basic reference ontologies may be combined into a composite structure that conveys a complex meaning. Such a composite would still be used for annotation and query purposes. To this end, the VPH RICORDO effort is developing a grammar (and implementing a corresponding composite editor) that draws upon terms from basic reference ontologies to create composite representations of complex biological concepts (see Figure 2c for an illustration of the grammar as applied to “venous return”). The key advantage of the composite approach is that complex concepts retain a mapping to reference ontology terms in a systematic and consistent manner.

## 7 Why is standard metadata encoding essential for Semantic Interoperability?

The choice of reference ontologies as knowledge representational standard for a set of DMRs does not necessarily render this set semantically interoperable *per se*. A second key requisite for SI is the standardization of the manner by which an ontology term is associated to an element in a DMR. Therefore, for a set of DMRs to be fully semantically interoperable, it is also required it adopts the same standard:

- a) set of relations (see **Box 2**) to relate a uniquely identifiable element in a DMR (typically identified by a **Uniform Resource Identifier (URI)**) and the URI of a term from an ontology. The triplet involving the two URIs and a relationship term URI is known as an **annotation** (as shown in Figure 1b);
- b) method to encode the annotation triplet – *i.e.* the standardized annotation *semantics* are to be encoded in a standardized *syntax* to ensure coherent machine readability and processing.

The link between anatomical and pathology terms (discussed in the context of DMR1 in Box 1) may require a number of relations to convey the full meaning of individual associations. For example, the anatomical term “pancreas” may be related to the cardiac pathology term “diabetic cardiomyopathy” as a **site of primary cause (relation 1)**, while the anatomical term “cardiac muscle” may be associated as a **site of primary effect (relation 2)**. In the case of the gene expression dataset (DMR2) the experimental sample from which the data was derived would be linked to the anatomical term via a more direct ‘**is a portion of**’ relation (**relation 3**).

**Box 2.** Examples of relations relevant to the cardiac drug discovery scenario

The basis for SI for a set of DMRs, therefore, relies on a usage consensus of (i) ontology terms, (ii) relations, and (iii) encoding standards for embedding annotation triplets in DMR metadata. The choice of encoding methodology may itself be informed by factors such as

community SI objectives and the associated expectation of DMR sharing discussed in the Introduction.

## **8 DMR integrity, security and metadata sharing**

A key advantage that metadata standardization offers to collaborative communities in biomedicine is that free and unfettered sharing of DMR metadata information may take place without compromising security restrictions of the DMRs themselves. In other words, depending on the choice of metadata encoding, annotations may become accessible as a catalogue for querying by third parties, without having to provide access to the original models or datasets being catalogued. For example, within a Pharma company, a clinical department may serve a catalogue *describing* clinical trial data holdings without necessarily *providing* access to the actual data repositories to unauthorised personnel.

The uncoupling of metadata from their corresponding resource has the additional benefit of protecting the integrity of DMRs. No significant change to the format of a DMR may be required if related metadata can be stored in a separate file as long as it holds a mapping to the DMR element URIs. This approach may therefore provide a viable SI solution despite the inevitable heterogeneity of resource formats: for instance, cardiac physiology models written in different programming (or markup) languages may share the same metadata standard along with radiological datasets of the heart (which may also be stored over a number of heterogeneous formats).

## **9 What are the key VPH priorities and objectives for annotation?**

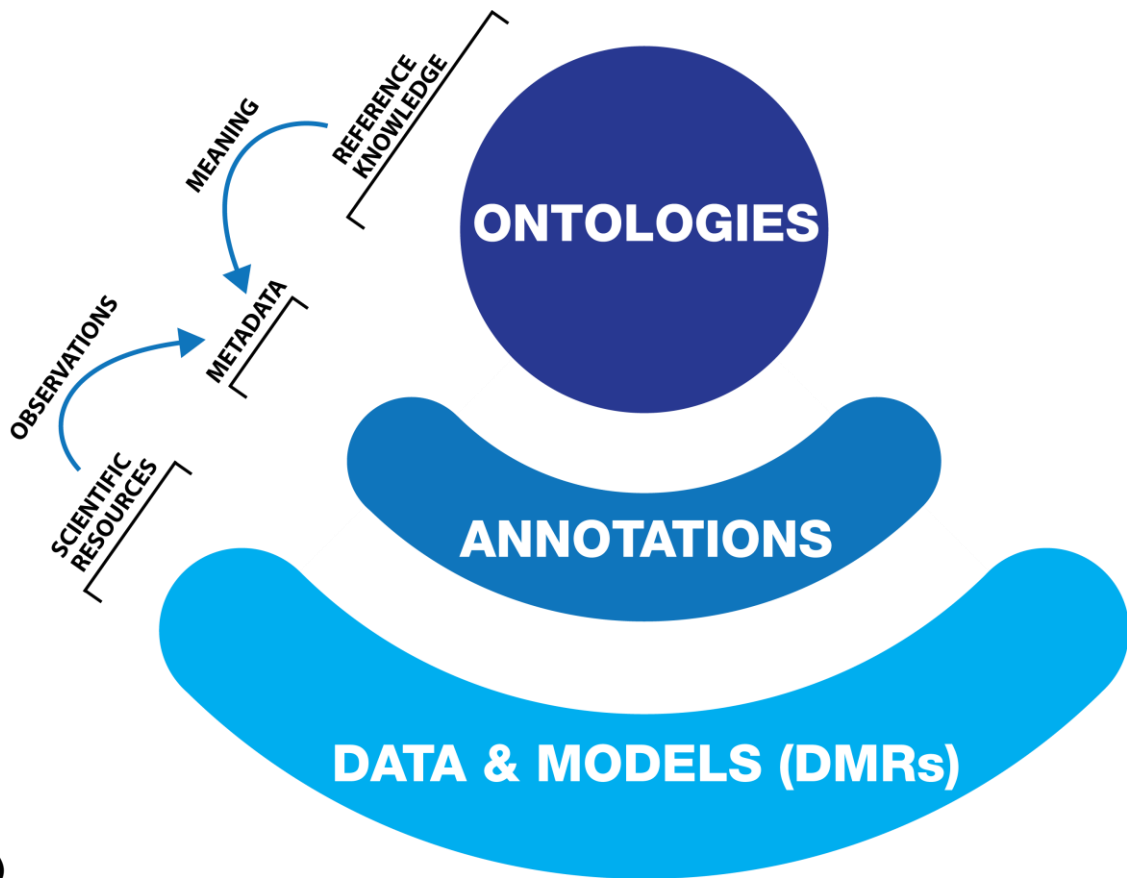
The biologically meaningful co-ordination of mathematical modelling and data resource management in the domains of systems biology, bioengineering and pharmacometrics requires SI between the systematic description of models and the documentation of pre-clinical and clinical datasets. To this end, the VPH NoE and RICORDO effort is designing and implementing a SI framework over two fronts:

- a) The first priority is to build a community standard for:
  - i. the use of reference ontologies from the community as a source of unambiguous and uniquely identifiable terms and relations for DMR element annotation;
  - ii. the well defined encoding of uniquely traceable metadata in which annotations are embedded.

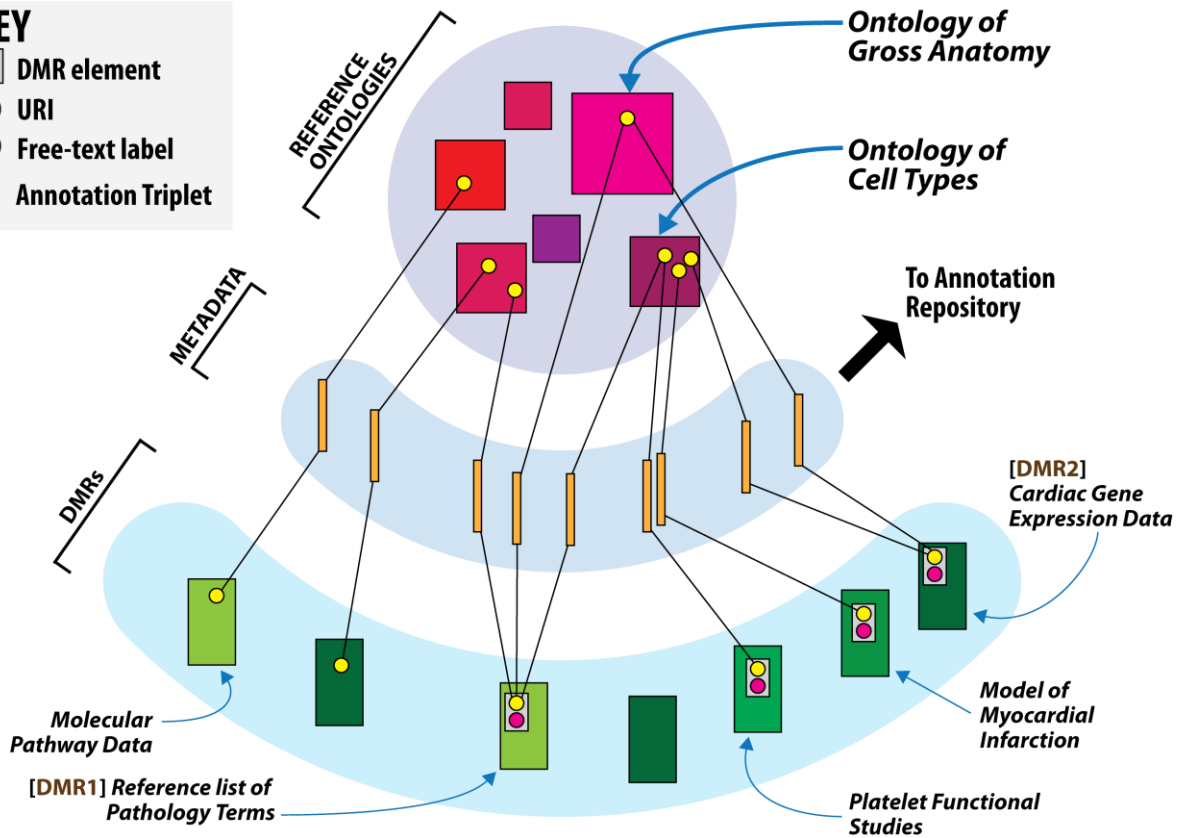
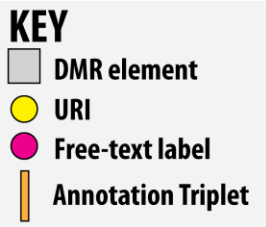
- b) The second aspect of this effort addresses the development of an open toolkit to:
- i. annotate DMRs and to share annotation triplets. Annotation sharing is done in such a way that their distribution may be uncoupled from the (accessibility or format) restrictions that may be applicable to their original DMRs;
  - ii. provide services in support of querying repositories of annotations (see **Figure 3**) through the automated reasoning over the reference ontologies from which terms are derived.

In short, part of the VPH approach to a SI plan for a set of biomedical DMRs is to create a new data resource: a shareable and query-able catalogue of annotation triplets describing the other DMRs (see Figures 2c and 3). The infrastructural objective of this effort is to provide the tools with which to generate, house, access and search these annotations and, crucially, to reason over the ontologies they refer to (Figure 3).

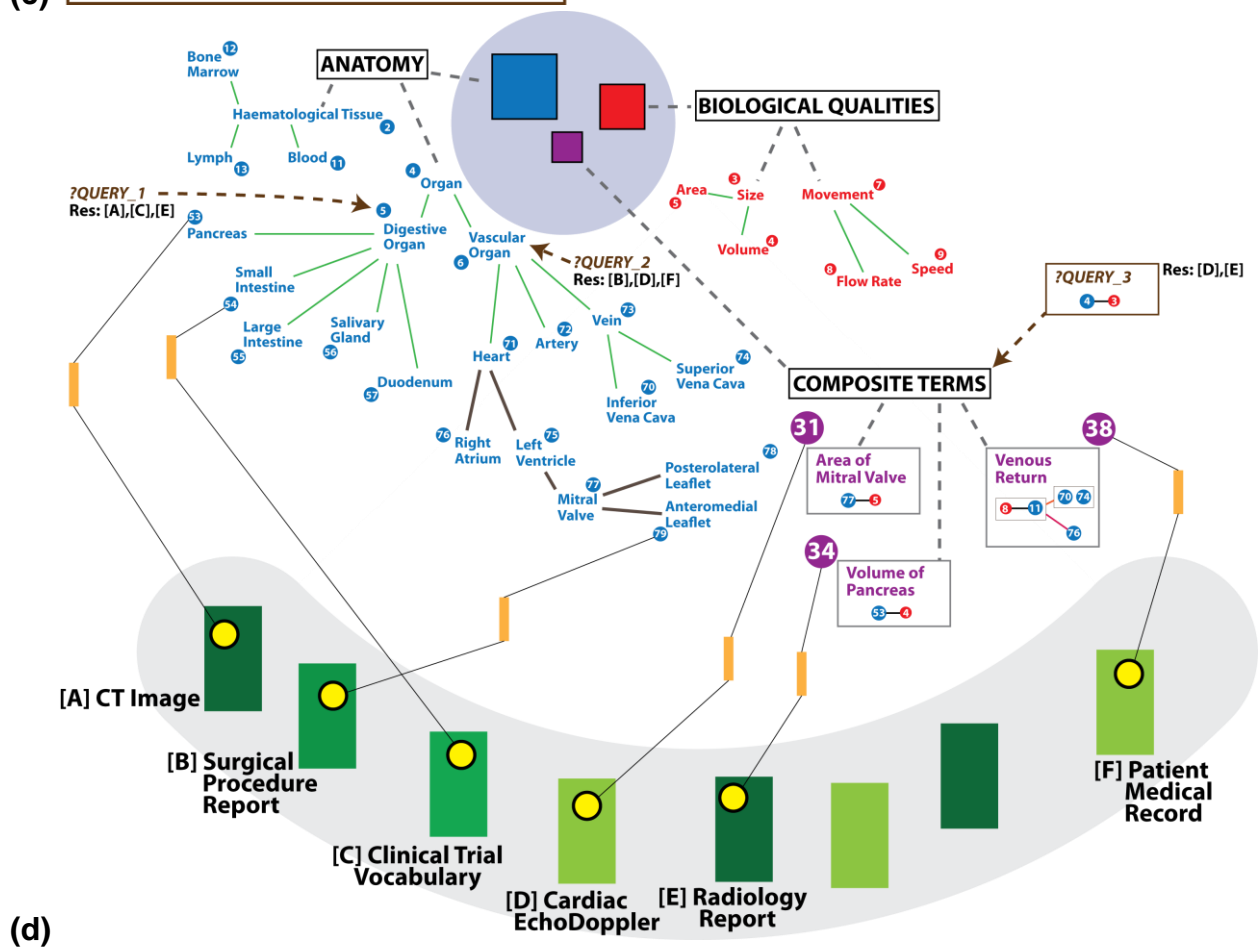
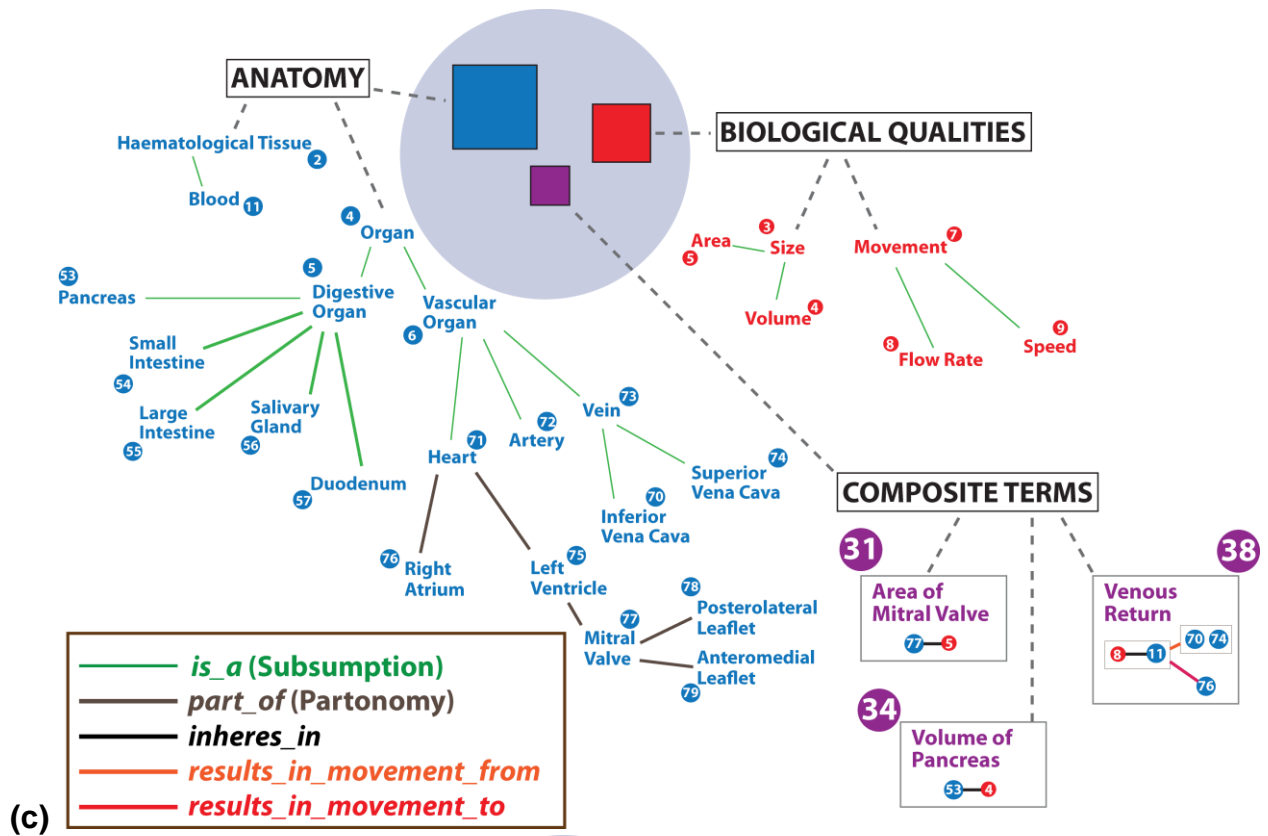
More specific technical details about the ongoing developments on the SI framework are available through the VPH RICORDO deliverables: [D2.2](#), [D2.3](#), [D3.1](#), [D4.1](#), [D4.2](#) and [D4.3](#).



(a)



(b)



**Figure 1.**

**(a)** Overall schematic representation of the key aspects of semantic interoperability (SI) in which annotations provide a link between DMR observations and ontology-based meaning.

**(b)** An example illustrating the role of semantic metadata in support of the SI for a set of DMRs. Note that the DMRs may have different formats to encode their *scientific* content. However, for SI to be possible between this set of DMRs, the encoding of the *semantic* metadata content must be in the same language/format. A second key requisite for SI is that the DMR metadata must make use of the same set of reference ontologies for its semantic content. (**URI**: Uniform Resource Identifier.)

**(c)** A detail of reference ontology structure representing explicit knowledge. The section of the Biological Qualities ontology only makes use of the subsumption relation. The Anatomy ontology also uses the parthood relation. Note that, while composite terms have their own unique identifier, they still explicitly refer to URIs of standard reference ontologies.

**(d)** Annotations of DMRs using terms from standard reference ontologies and their composites may be created using the RICORDO VPH framework. The same framework also supports annotation searching.

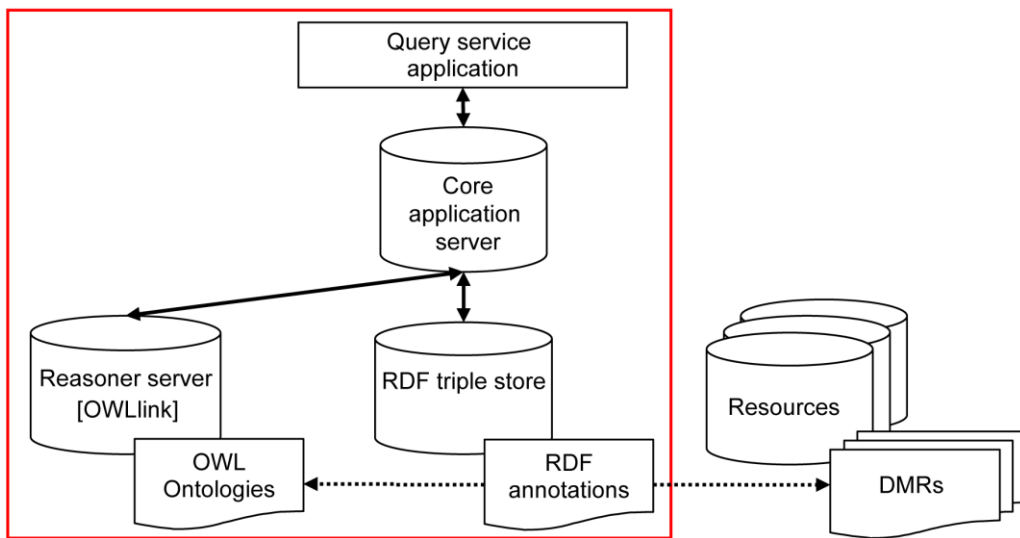
For example, in **Query\_1**, the RICORDO infrastructure supports the search for any annotation that involves ontology terms that are a type digestive organ by reasoning over the subsumption edges of the graph. This query takes as input the anatomy ontology URI for Digestive Organ (URI: 5). The result of this annotation search returns three hits: (i) an annotation to a DMR [A] element that represents a region on a CT image equivalent to the pancreas; (ii) an annotation to DMR [C] element that bears a clinical trial vocabulary term (*e.g.* the ‘Small Intestine’ term from ‘Anatomical Location’ CDISC codelist); and (iii) DMR [E] element that reports on the volume of the pancreas calculated from a radiology image.

**Query\_2** supports queries that search for annotations to all known vascular organs and their parts. This query takes as input the anatomy ontology URI for Vascular Organ (URI: 6) and returns the following annotations: (i) DMR [B] element pertaining to the a surgical procedure report to the anteromedial leaflet of the mitral valve; (ii) DMR [D] element that holds data about a cardiac echo Doppler that measures the area of the mitral valve; and (iii) DMR [F] in which a medical record stores information about the flow of blood from the patient’s central veins to the right atrium. The latter concept is represented as a composite of standard reference ontology terms.

**Query\_3** asks the following questions: ‘Are there annotations that explicitly describe the size of organs?’ The *ad-hoc* creation of a composite construct that refers to the size (as the quality ontology URI:3) that pertains to (or *inheres\_in*) all known organs (*i.e.* their subclasses and parts) (anatomy ontology URI: 4) allows the RICORDO framework to ask that question in a machine interpretable manner. This query returns the annotations to DMR [D] (area of mitral valve) as well as DMR [E] (the size of pancreas). Both ‘Area’ and ‘Volume’ are subclasses of ‘Size’ in the quality ontology. ‘Pancreas’ is a subclass of ‘Organ’, while ‘Mitral Valve’ is a part of an ‘Organ’ subclass.



**Figure 2.** Example illustrating the structure of an archetypal model resource **A** (in this case, an SBML model) in which element **B** is shown to bear (i) a human-readable text label **G**, as well as (ii) machine-readable metadata **C**. The annotation in the metadata above conveys the meaning that the reaction with the Unique Resource Identifier (**URI**) 230655 (**D**) that occurs in (**F**) the anatomical location identified by URI **E**. The latter URI represents the sinoatrial node term in the Foundational Model of Anatomy.



**Figure 3.** The current implementation of the RICORDO infrastructure (enclosed by a red margin). The RICORDO core application is deployed in a tomcat application server. It interacts with (i) the OW knowledge base which is deployed in a Pellet OWLlink server, and (ii) the RDF repository which resides in a Virtuoso server. RDF annotations carry URI references to OWL ontology terms and DMR elements. Note how the DMRs themselves need not be directly available to the users of the query service application.

**Note:**

OWLlink – <http://OWLlink-owlapi.sourceforge.net/>;

RDF – <http://www.w3.org/RDF/>;

OWL – <http://www.w3.org/TR/owl-features/>

Tomcat - <http://tomcat.apache.org/>

Virtuoso - <http://virtuoso.openlinksw.com/>

Pellet - <http://clarkparsia.com/pellet>

## Interactions and Dependencies

There is a number of guidelines developed within the VPH and to which the present Knowledge Representation guideline are more or less directly related. The commented list below may require updating:

### Tool Characterisation

Knowledge representation tools developed within the VPH context ought to take into account generic tool characterisation guidelines.

### Model Characterisation

Knowledge representation efforts in the VPH context ought to take into account Model Characterisation guidelines. The relation here is not necessarily one of inheritance. Developing knowledge representation artefacts is not necessarily developing models and models may be characterised in ways which are not immediately applied to knowledge-based approaches. However, VPH models constitute one main resource that the VPH knowledge representation tools are intended to leverage upon. Therefore, in carrying out related knowledge representation tasks, awareness of model specifications and the dimensions involved in those specifications is desirable in the same way that knowledge about a subject matter is desirable. Knowledge representation tools intended to cater for the needs of model management will also demand awareness in the form of compliant mapping of model schemas.

### Data Characterisation

*Mutatis mutandis*, the potential relationship between the present guideline and the Data Characterisation guideline is the same as that between the present guideline and the Model Characterisation guideline.

### Ontological Annotation

This is the present guideline. It is titled Knowledge Representation guideline to take into account the broader scientific and technological context of ontological annotation.

### Interoperability

The VPH Interoperability guideline has broader scope than the present guideline. It is not only best practice in developing knowledge representation tools and artefacts to aim for interoperability. Indeed, knowledge representation is motivated, partly by interoperability requirements and objectives. Nevertheless, general interoperability questions arise in the development of knowledge representation tools which need to be cross-platform and ideally handle diverse sources and formats.

### Ethico-legal issues, provenance

No specifics.

### Licensing

It is notable that there is a significant divide between open source and proprietary software in knowledge representation. Community specific software tends to be open source and free. Also, significant, reusable knowledge representation technologies, however, even when developed on a proprietary basis do have open versions and in most cases accessible academic licenses. The present guidelines defer to the more general licensing guidelines but the strategic importance of this aspect must be highlighted.

### Usability and training

Despite its wide range of application, knowledge representation remains a specialist area. It is envisioned that most contributions to the VPH in this area will contain a backend and a frontend, with the frontend being as user friendly and as accessible as possible. This ought to be completed with readily available and simple training or tutoring material. It is credible that training events would be organized in connection to VPH contributions, if only as part of the development process of educational resources. Backends, especially when in the open source realm, that are suitable for being accessed by more experienced users or developers will also demand documentation that follows best practices. As best practices are arguably context and community specific, the present guidelines defer to the VPH Usability and Training guidelines for generic guidelines but it is worth emphasizing the essential importance of this aspect.

## Applicable Legislation

The present guideline has nothing specific to contribute to this topic.

## Standards and Standards Bodies

The Open Biomedical Ontologies (OBO) knowledge representation community which supports the development of biology-related ontologies with the goal of creating a suite of orthogonal (i.e. non-overlapping) interoperable reference ontologies. The role of the OBO for the VPH community is to advise on the most appropriate ontologies to use for annotations, and to support the management of ontology term requests.

## Characteristics

### **Overview**

This section presents a selected list of knowledge representation tools that are relevant to the VPH community. Although this list includes a number of references to knowledge representation resources such as languages and tools which are of broad and generic interest, it emphasizes references to resources particularly relevant to and sometimes simply developed by the biomedical community.

### **Knowledge Representation Languages**

The root of knowledge representation is in the languages that are used to develop formal knowledge representation artifacts, such as ontologies, and carry out the ground task of representing knowledge in a computer system.

#### **Logic-based knowledge representation languages**

A number of languages have evolved which are historically rooted in logic-based knowledge representations and its logic programming side in computer science. Often time, these languages are very similar in ways but differ in their native implementation platforms which secure for them reasoning capabilities and programmatic software environment.

Two main families of closely related logic programming approaches to knowledge representation are:

- Prolog-based. A concrete instance of which is embedded in the SWI-Prolog system.
- LISP-based. A concrete instance of it is embedded in the Cyc system (and OpenCyc variant of it). Another example is the open source OCML.

CommonLogic (CL). Common logic is the result of standardisation of logic for knowledge representation, It can be seen at a certain level as an abstraction over various logic based knowledge representation languages.

OBO format. OBO Format is a flat file format for Open Biomedical Ontologies. It has emerged from efforts in the GO community. It has been given a specification using the CL standard, namely OBOLog.

## Web languages

Web languages have been developed and designed for their suitability to standardised Web applications to knowledge representation.

RDF. The Resource Description Framework can be seen as a simple knowledge representation language. Many of the logic-based knowledge representation languages will also endeavor to support some level of interoperability, thus extending the purview of their applications to the Web. Such links oftentimes come with trade-off which have to do with computational constraints imposed by Web-based applications.

RDF-Schema is an extension of RDF that allows to specify schemas that are very close to p SPARQL (a query language for RDF)

OWL is a W3C recommendation that is used to specify ontologies for the Web. It is supported by a number of specific tools and most systems pay attention to allow for integration with those tools, for example through plug-ins, and intended support extends to many knowledge representation languages and the systems implementing their framework. Concretely, OWL extends RDF(S) and support is generally attentive to computational optimisation thus creating a trade-off between expressiveness and reasoning capabilities.

## Classes of tools

The second column of the table below provides an indication of the class of tool - tools may be classified over two main distinguishing axes, namely:

### **By functionality:**

Tools may be classified according to the functionality they provide. In the present document we do not make a distinction between monofunctional tools (e.g. a specific Web service) and systems (e.g. an overall knowledge management system integrating multiple components and exhibiting multiple functionalities). We will, however, mention relationships between simpler and more complex tools listed.

#### *a) Authoring.*

Authoring tools allow for creating or editing ontologies and knowledge bases. They may be specific to a particular language or they can support a variety of language, although the multi-support may only be implemented through import/export functionalities.

*b) Browsing and Visualisation.*

Visualisation tools provide a graphical user interface to browse knowledge resources (ontologies or datasets). They can present the resources as a browsable tree structure or implement more sophisticated visual effects.

*c) Reasoning.*

Reasoning tools supporting inferencing or querying over ontologies at various degrees of sophistication. These tools can rely on plug-ins (e.g. a reasoner plug-in in Protege) or have the capability integrated in their system ().

*d) Management.*

Management tools allow for manipulating multiple ontologies. For example, they are ontology repository browsers. They can also provide some degree of integration such as by allowing to merge ontologies. Stores are included under this heading.

**By deployment:**

*a) Desktop vs Web applications:* While desktop applications are tools that can be installed on a local machine and run in an autonomous fashion, Web applications sit on the Web and can be accessed through the internet.

*b) Web service vs library:* Many Web applications providing a user interface will also be complemented by Web services which allow for programmatic access. Similarly, there will be a number of libraries that allow for developing applications.

**Systems:**

For the purpose of this document a system is either an environment or an application that combines multiple functionalities. We do not distinguish strongly systems in relation to the user interface they present, therefore there is a spectrum of systems in the present sense. At

one end of this spectrum are systems are software that present users with a graphical interface to which functionalities may be added through a number of plug-ins. At the other end of this spectrum would be command line systems that allow functionalities through access to libraries.

#### ToolBox:

A toolbox is a set of resources combining different tools, services or libraries and which are developed by an identified community or for a given specific domain of applications. The most significant toolboxes identified in the present document are the NCBO, GO, and RICORDO toolboxes.

### Table

\* = based on proprietary software with versions having academic licences

Name	Class	Availability	Key features
Protege	System Desktop Web-based	<a href="http://protege.stanford.edu/">http://protege.stanford.edu/</a>	Authoring tool Reasoning, Visualization (plug-ins) RDF(S), OWL, OBO Format
OLS	Web-based Ontology Browser Web services	<a href="http://www.ebi.ac.uk/ontology-lookup">http://www.ebi.ac.uk/ontology-lookup</a>	Ontology browser OBO ontologies
NCBO Bioportal	Repository Browser Web-based	<a href="http://bioportal.bioontology.org/">http://bioportal.bioontology.org/</a>	Ontology repository access Ontology browser OBO format, OWL
OBOEdit	Authoring Desktop	Part of GO toolkit	OBO ontologies OBO format

Phenote	Annotation Web-based Desktop	<a href="http://www.phenote.org">http://www.phenote.org</a>	Annotation with ontologies OBO format
Phenomizer	Authoring Web-based	<a href="http://compbio.charite.de/Phenomizer/Phenomizer.html">http://compbio.charite.de/Phenomizer/Phenomizer.html</a>	Human Phenotype Ontology Uses statistical correlations to support clinical diagnostic
ISA tool suite	Authoring Desktop	<a href="http://www.isa-tools.org">http://www.isa-tools.org</a>	Ontological Annotation of Experiments
RICORDO VPH toolkit	ToolBox	<a href="http://www.vph-ricordo.eu">http://www.vph-ricordo.eu</a> (prototype available in 2010, release version available in 2011)	Online browsing of VPH RDF dataset Authoring tool combining OBO
GO ToolBox	ToolBox	<a href="http://www.geneontology.org/GO.tools.shtml">http://www.geneontology.org/GO.tools.shtml</a>	List of MolBio analysis tools built upon the Gene Ontology standard
OpenCyc*	System	<a href="http://opencyc.org/">http://opencyc.org/</a>	Authoring and reasoning knowledge-based system Browse and edit knowledge base Create CycL ontologies Inference Export and map to Web and Linked Data
NCBO PURL Server	Web-based browser Web service	<a href="http://purl.bioontology.org/">http://purl.bioontology.org/</a>	Cross-referencing of bioontological identifiers

NCBO BioPortal REST services	Web services library	<a href="http://www.bioontology.org/wiki/index.php/NCBO_REST_services">http://www.bioontology.org/wiki/index.php/NCBO_REST_services</a>	Reference tracking
NCBO ToolBox	ToolBox	<a href="http://www.bioontology.org/technology">http://www.bioontology.org/technology</a>	Bioportal, Protege, PURL
MIRIAM Resources	Browser Web-base Webservice	<a href="http://www.ebi.ac.uk/miriam/">http://www.ebi.ac.uk/miriam/</a>	Cross-referencing of bioontological identifiers
SWI-Prolog	System Desktop	<a href="http://www.swiprolog.org/">http://www.swiprolog.org/</a>	Prolog-based Authoring Reasoning RDF and semantic Web support
OCML	System Desktop	<a href="http://technologies.kmi.open.ac.uk/ocml/">http://technologies.kmi.open.ac.uk/ocml/</a>	Lisp-based Authoring Reasoning OCML
Joseki	Management	<a href="http://joseki.sourceforge.net/">http://joseki.sourceforge.net/</a>	Triple store Supports RDF and SPARQL queries
OpenRDF	Management	<a href="http://www.openrdf.org/">http://www.openrdf.org/</a>	RDF triple store Supports RDF and SPARQL queries
OWLIM*	Management	<a href="http://www.ontotext.com/">http://www.ontotext.com/</a>	RDF triple store

		t.com/owlim/	Supports RDF and SPARQL queries
Jena Framework	System	http://jena.sourceforge.net	Authoring, reasoning RDF, OWL, SPARQL
OWL API	Authoring and management library	http://owlapi.sourceforge.net/	Authoring and manipulation services for OWL ontologies  RDF, OWL, OBO format (parser)
Anatomy API	Reasoning	Release available in 2011 (Contact: bdb@ebi.ac.uk)	Supports graph manipulation of anatomical connectivity ontology
ONTOCAT	Browsing library	http://ontocat.sourceforge.net/	Integrated queries over NCBO Bioportal and OLS  OWL, OBO

## Methods of Verification

Verification is multi-faceted and context-dependent. Methods and aims vary across technology, communities and application domains.

The present guideline defers in general to best practices. In particular it is best practice to have ontological artifacts enduring review processes. Also, communal exposure and feedbacks ought to be relevant to validation processes. In some cases, generic technologies are accompanied by validation tools, e.g. syntax validation. Finally, in general and in particular for tools and software developed within the VPH context, an important instrument is the existence of tests and test suits, possibly using communally designed test datasets.

## Ownership

The present guideline has nothing specific to contribute to this topic - however, it is suggested that awareness of the following should be borne in mind in relation to the theme of the present guideline insofar as it connects to the topic of ownership. The respect of intellectual property and due credit is an essential, if only sometimes tacit rule in the academic community. It is also noteworthy that open source and free communal efforts are well in tune with the knowledge representation ecosystem and are a factor of productivity. They also have rightful significance in light of the public benefit that the VPH promises to bring about.

## Training

Knowledge representation is a possible demanding activity in terms of skills and background knowledge. These aspects, however, comes in degrees. The current trend is to develop front end and user interface which impose the less demand and promote greatest usability. But there is always a trade-off which is dependent upon aims, tasks, and resources. It is not incredible to have different levels of user according to their background and proficiency. In this connection, it is also best practice to consistently develop training material which is suitable for the intended user basis. Modalities for the dissemination of such training are open and may depend upon circumstances. Training can occur during workshop or special tutorial sessions in communal events. Nevertheless, persistent and adequately updated resources that are independently accessible must be provided. It is also important that any element of the VPH toolkit be accompanied with basic training material in aspects that are necessary for full operation at no extra cost for the user.

## Maintenance

[If there is a requirement for the currency of this topic to be maintained, perhaps due to changing standards, this should be discussed here]

Knowledge representation artifacts, ontologies, standards, tools and technology evolve. Maintenance is, however, not risk free. VPH toolkit elements should be supported and maintained. This can be done as a result of transfer, such as outsourcing maintenance to a community or making resources and tools adopted by a established and significantly active bodies. For example, an ontology may be submitted to the OBO Foundry or tools may be open sourced, provided this is an effective and not merely formal transfer, when the original developer discontinues maintenance.

It must be clear that VPH may not commit to technology that will not be robustly maintained over the relevant period. It must also be clear that components may not be made essential that will not come with a satisfactory maintenance strategy. Tools and components which claim to remain relevant must be such that their maintenance can be picked up by active developers if their being unattended creates functionality breakdown within the VPH toolkit. Risks relating to these aspects ought to be regularly evaluated.

## Ranking

[Here the various possible combinations of attributes are described that might be considered to form a basis for the comparative ranking of tools, models, or data.]

It is difficult, if not impossible, to provide definitive ranking scales in relation to the relevant technologies. This is because there are a number of dimensions that are relevant to the adjudication of the usefulness of artifacts and tools in the context of the VPH Toolkit. Ranking may vary according to aims, tasks, the characteristics of the intended user basis and priorities. Also, most of the dimensions that shape this guideline are possible criteria even if they cannot be uniquely ordered.

The best approach is to identify as genuinely as possible any aspect of VPH Toolkit component according to a number of basic criteria which may require themselves further evaluation. The proposed list is indicative and not exhaustive, the segmentation of the list and the ordering of its items is not significant and does not suggest order of priority.

### ***Absolute generic criteria:***

Availability, e.g. proprietary versus open

Usability, e.g. as mapped against required accessibility capabilities, technical proficiency, domain knowledge, familiarity with specific resources

Documentation and training

Robustness of validation procedures

### ***Context-dependent criteria:***

Integration and interoperability

Adoption of communal standards

Strength and commitment of developer community and user basis

Fitness for purpose (quality)

Generic application, e.g. generally reusable component versus specific ad hoc component

## Documentation, Reporting Templates

It is best practice in the knowledge representation community to develop self-documented artifacts. There are inherent reasons for which formal specifications of ontologies, for example, cannot do without adequate documentation. This aspect is also made more or less critical depending on the choice of knowledge representation framework. Tools ought to follow best practices of software engineering with self-documented code and embedded documentation. In the context of the VPH, user manuals when required should be pragmatic, adapted to the intended audience and of outstanding quality. Training material and test suites can also be seen as parts of documentation for the present purpose.

For the purpose of cataloging, Knowledge Representation VPH Toolkit components ought to be provided with a simple fact-sheet that replaces the component within the relevant dimensions emerging from the present guideline.

Recommendation for template for such fact-sheet could emerge as these guidelines evolve.

## Further Information

The present guideline is open to update, modification and changes.

Version number	Date	Comment
0.1	2010/07/30	Previous version.
0.2	2011/02/18	This document.

## References

- [1] C. Rosse and J. L. Mejino, Jr., "A reference ontology for biomedical informatics: the Foundational Model of Anatomy," *J Biomed Inform*, vol. 36, pp. 478-500, 2003.
- [2] C. Rosse and J. L. V. Mejino, Jr., "The Foundational Model of Anatomy Ontology," in *Anatomy Ontologies for Bioinformatics: Principles and Practice*, vol. (in press), A. Burger, D. Davidson, and R. Baldock, Eds. New York: Springer, 2007.
- [3] J. H. Gennari, M. L. Neal, J. L. V. Mejino, D. L. Cook. *Using multiple reference ontologies: Managing composite annotations*. In: *Proceedings of the International Conference on Biomedical Ontology*. Buffalo, N.Y., 2009. p. 83-86. NIHMSID: 164625.
- [4] D. L. Cook, J. L. V. Mejino, M. L. Neal, and J. H. Gennari, *Composite Annotations: Requirements for Mapping Multiscale Data and Models to Biomedical Ontologies*, presented at *IEEE Engineering in Medicine and Biology Conference*, Minneapolis, MN, 2009.
- [5] M. L. Neal, J. H. Gennari, T. Arts, and D. L. Cook, "Advances in semantic representation for multiscale biosimulation: a case study in merging models," *Pac Symp Biocomput*, pp. 304-15, 2009.
- [6] D. L. Cook, J. L. Mejino, M. L. Neal, and J. H. Gennari, "Bridging biological ontologies and biosimulation: the Ontology of Physics for Biology," *AMIA Annu Symp Proc*, pp. 136-40, 2008.
- [7] J. H. Gennari, M. L. Neal, B. E. Carlson, and D. L. Cook, "Integration of multi-scale biosimulation models via light-weight semantics," *Pac Symp Biocomput*, pp. 414-25, 2008.
- [8] Richard Baldock and Duncan Davidson The Edinburgh Mouse Atlas, in "Anatomy Ontologies for Bioinformatics: Principles and Practice" (Springer, 2008) 249-266.
- [9] Bard J, Rhee SY, Ashburner M. An Ontology of Cell types, *Genome Biol*. 2005;6(2):R21.
- [10] The Gene Ontology (GO) database and informatics resource *Nucleic Acids Res*. 2004 January 1; 32(Database issue): D258–D261
- [11] D.A. Natale et.al., *Framework for a Protein Ontology*, *BMC Bioinformatics* 2007, 8(Suppl 9):S1doi
- [12] J. Hastings, P. de Matos, M. Ennis & C. Steinbeck, *Towards automatic classification within the ChEBI ontology*, *Nature Proceedings*, doi:10.1038/npre.2009.3525.
- [13] Schofield PN, Bard JBL, Rozell B, Sundberg JP: *Computational pathology: challenges in the informatics of phenotype description in mutant mice*. In: *Handbook on Genetically Engineered Mice*, ed. Sundberg JP, Ichiki T, pp. 61–81. CRC Press, Boca Raton, FL, 2005.
- [14] G. V. Gkoutos, E.C.J. Green, A-M Mallon, J.M. Hancock and D. Davidson, *Using ontologies to describe mouse phenotypes*. *Genome Biology*, 2004, 6, R8. [<http://genomebiology.com/2004/6/1/R8>]
- [15] Gkoutos, G.V., Green, E.C.J., Mallon, A.M., Blake, A., Greenaway, S., Hancock, J.M., and Davidson, D., *Ontologies for the Description of Mouse Phenotypes*, *Comparative and Functional Genomics*, 2004, 5, 545-551
- [16] C. J. Mungall, G. V. Gkoutos, C. L. Smith, M. A. Haendel, S. E. Lewis, and M. Ashburner, "Integrating phenotype ontologies across multiple species," *Genome Biol*, vol. 11, pp. R2, 2010.
- [17] G. V. Gkoutos, C. Mungall, S. Dolken, M. Ashburner, S. Lewis, J. Hancock, P. Schofield, S Kohler and P. Robinson, *Entity- Quality-Based Logical Definitions for the Human Skeletal Phenome using PATO*, *Conf Proc IEEE Eng Med Biol Soc*. 2009;1:7069-72.