RECIIS
Electronic Journal of Communication Information & Innovation in Health
[www.reciis.cict.fiocruz.br]
ISSN 1981-6286

Original Articles

# Survey of current terminologies and ontologies in biology and medicine

DOI: 10.3395/reciis.v3i1.239en

Fred Freitas
Informatics Center, Federal University of Pernambuco, Recife, Brazil
fred@cin.ufpe.br

Stefan Schulz
Institute of Medical Biometry and Medical Informatics, University Medical Center, Freiburg, Germany
stschulz@uni-freiburg.de

Eduardo Moraes
Informatics Center, Federal University of Pernambuco, Recife, Brazil
ecm2@cin.ufpe.br

## Abstract

This paper provides a survey of the state of the art in terminologies and ontologies applied to Biology and Medicine. Not intending to be fully comprehensive, we describe some of the most relevant resources that currently attract interest from industry and academia. We introduce a description framework and compare the systems in terms of their architectural elements, their expressiveness, and coverage, as well as analyze the nature of entities they denote. In particular, we scrutinize the International Classification of Diseases (ICD), the Medical Subject Headings (MeSH), the Gene Ontology (GO), the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), the Generalized Architecture for Languages, Encyclopaedias and Nomenclatures (openGALEN), the Foundational Model of Anatomy (FMA), the Unified Medical Language System (UMLS), and the Open Biomedical Ontologies (OBO) Foundry.

## Keywords

Terminology; ontology; biology; medicine

## Introduction

### Background

The growing digital availability of huge amounts of biomedical data and knowledge resources has burdened researchers and practitioners with the task of managing terabytes of semantic content, which is, by nature, subtly interwoven and needs to be aggregated and manipulated. A deluge of data used to solve complex tasks requires more and more sophisticated techniques of intelligent information and knowledge management, enhancing the interoperability of content in large repositories supported by different types of automated reasoning. This challenge has increasingly been addressed by biologists, clinical and public health researchers, health economists and clinical practitioners. A practical outcome of this effort is the emergence of an increasing set of semantic reference systems often characterized as vocabularies, thesauri, terminologies, and ontologies (Rubin 2007).

The current developments in biomedical knowledge management have essentially two roots:

- the establishment of indexing vocabularies and classification systems such as the International Classification of Diseases and the *Index Medicus*, dating back to the 19th century, driven by public health and epidemiology interests on the one hand, and by library science on the other hand; and
- the research on medical decision support and expert systems, starting in the seventies of the last century, driven by the emerging research field of Artificial Intelligence and inspired by the idea of creating knowledge-based computer tools to assist the complex process of medical decision making.

Motivated by the vision of the Semantic Web, the term "ontology" has become one of the most fashionable terms in Computer Science. Ontologies are advertized to precisely describe domains in detail and to employ these descriptions in many types of applications, ranging from natural language processing to logic reasoning and decision support systems. Many application areas currently take advantage of ontologies, but the field of life sciences is gaining more and more visibility in this picture, since very few scientific domains, if any, contain such impressive and rapidly growing amounts of terms, concepts, and definitions.

## Ontologies

The term "ontology" has become very popular since the mid nineties but, unfortunately, no universally accepted definitions exist (Kuzniersky 2006). Since the seventeenth century it has been used for the discipline of general metaphysics in the tradition of Aristotle's "first Philosophy" as the science of being qua being. It is often seen as complementary to the notion of Epistemology (the science of knowledge).

In Computer Science, the definition of ontology as the explicit specification of a conceptualization (Gruber 1995) prevails. Conceptualization is here meant as an abstract, simplified view of the world that we wish to represent for some purpose, e.g., to draw inferences, to perform automatic classification, etc. A conceptualization usually includes concepts (also called classes or types, e.g., Heart), individuals as instances of concepts (e.g. the individual Fido is an instance of Dog), binary relations between concepts or individuals (e.g. *Dog is-a Vertebrate*), logic-based restrictions (all instances of *Herbivore* eat *only* vegetables while all instances of *Carnivore* eat *some* instances of *Animals*), and axioms (sentences that are always true in a domain – e.g., every instance of *Living Person* has some instance of *Heart*). The link to connect these entities is clearly given by ontological relations. They will represent the different aspects in which concepts relate to each other. The most relevant and used relation types are subclass (*Heart* is a subclass of *Organ*, since all instances of the former are instances of the latter, with some special features that distinguish them from others), and partonomic relations (every instance of *Heart Ventricle* is a part of some *Heart*). But there are other definitions of ontology, such as "representation of a domain of discourse, consisting of a list of terms, the relationships among them and the axioms which are always valid in the domain" (Antoniou & Harmelen 2004), or a "representational artifact whose representational units are intended to designate classes or universals in reality and their interrelations" (Smith 2005).

The notion of ontology is often specialized to what is named "formal ontology" (Guarino 1998). This means that the content of an ontology is described using mathematical logics which can endow computer systems with the ability of logical inference. It can also support autonomous discovery over recorded data, as well as reuse and exchange of knowledge.

The rise of ontologies in the Computer Science mainstream has spread to many other branches of knowledge: Motivated by the vision of the Semantic Web (Berners-Lee 2001), many groups from academia and industry throughout the world became interested in ontologies, and the number of tools, standards, and users grew accordingly. Indeed, some goals to produce standard ontologies in some areas were accomplished, particularly in Medicine and Biology.

## Terminologies vs. ontologies

Especially Medicine is characterized by a wealth of so-called terminologies, best described as language-oriented artifacts that relate the various senses or meanings of linguistic entities with each other. Terminologies are generally built to serve well-defined purposes like document retrieval, resource annotation, the recording of mortality and morbidity statistics, or health services billing. Biomedical terminologies do not use formal and well-defined descriptions; they rather define the terms (if ever) by human language expressions, and express the associations between terms by informal, close-to human language relations. Words or multiword terms are the basic building blocks of terminologies, which generally organize them in hierarchies that relate their meanings in terms of synonymy (same meaning), hyperonymy (broader meaning), hyponymy (narrower meaning). Although terminologies can be successfully used in representing abstract meaning, e.g. in natural language processing or in the annotation of resources (e.g. literature abstracts, experimental results), they are not precise and expressive enough for more knowledge-intensive applications.

Whereas one use case may require knowledge on *how* and on *what* some terms differ from others, another one may demand more precise relations between terms (for example that every instance of a normal *Arm* has some instance of *Forearm* as its part. To meet these requirements, a language-centered resource is not expressive enough. Here, a reality-centered resource is better suited in order to capture the subtleties of which entities (objects, qualities, processes, etc.) are related to others, under which circumstances these relations hold, and how these relations should be exactly interpreted (e.g. of whether the relation part-of between a body part and a body still holds after the body part like a kidney is removed). That is where ontologies come into play. Ontologies are expressed in logic-based formalisms, which provide (meta-) definitions of classes (concepts), relations, instances and axioms. Therefore, ontologies can represent a domain in a form that computers can handle the definitions according to the semantics of the definitions instead of employing only terms or semantic identifiers. Thus, a system can check whether some interpretation is correct or not, if a given statement is true according to some ontology, among other related tasks. Ontologies can also encompass different dimensions that a domain should embrace: for instance, in organisms, the degree of canonicity of organs (whether an organism functions as usually supposed or not), the degree of development (e.g. embryo vs. adult), the place of an organism or organic matter in the biological taxonomy (e.g. fly vs. mouse), or the granularity by which biological structure is described (e.g. macroscopic vs. microscopic), to mention a few (Schulz 2004).

However, the classical terminological approach is increasingly blended with principles of modern ontology design, with ontology languages from the Computer Science domain and with the emerging discipline of applied ontology embedded in the field of Analytical Philosophy.

What we intend to describe in this paper is the broad range of these very heterogeneous artifacts, for which an overarching term is still missing (the often used term "biomedical vocabularies" is misleading as it stresses too much the language aspect). In the remainder of this article we therefore use the acronym BMTOs for "biomedical terminologies and ontologies". It is organized as follows: In the next section, the main BMTOs are explained in detail. Section 3 is devoted to foundations and efforts that integrate many of these systems. Section 4 discusses some important topics from each BMTO, while Section 5 addresses open issues and challenges for integration of BMTO.

## Important examples of biomedical terminologies and ontologies (BMTOs)

### Description scheme

Several efforts have been made in the biomedical field for the development of semantic standards such as medical terminologies, ontologies, and coding systems. In this section, we will analyze a set of BMTOs which reflects the broad variety of this genre. We will address the International Classification of Diseases (ICD), the Medical Subject Headings (MeSH), the Gene Ontology (GO), the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), the Generalized Architecture for Languages, Encyclopaedias and Nomenclatures (openGALEN), the Foundational Model of Anatomy (FMA) and, as examples of overarching initiatives, the Unified Medical Language System (UMLS) and the Open Biomedical Ontologies (OBO) Foundry. We will describe and compare them by identifying common features and differences. Moreover, we will discuss what these systems represent and which architecture they use. To this end, we introduce the architectural elements we encounter in all BMTOs as follows:

- **Nodes**: the primary identifiers of meaning
- **Links**: the connections between nodes
- **Codes**: alphanumeric identifier for a node or a link.
- **Hierarchies**: network of links that constitute a partial order, thus defining trees or directed graphs
- **Attributes**: seen as further descriptions of nodes and links
- **Axioms**: sentences expressed in logic which are always true in the domain.

We furthermore describe the systems in terms of

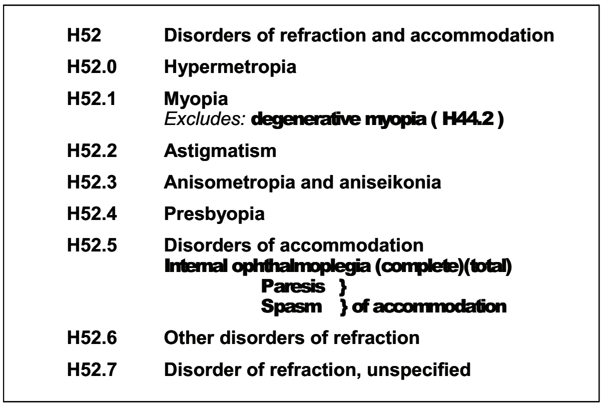
- **Purpose**: why they were built and where they were used
- **Scope**: the knowledge domain they represent
- **Reference**: what nodes and links denote

### The International Classification of Diseases

Terminological standardization in Medicine has a long history. In 1880, the International Classification of Diseases (ICD) (WHO 2008) was created, based on the London Bills of Mortality which distinguished about 200 causes of death providing codes for all known diseases at that time. For many years, the ICD was the only medical terminology resource. Its current (10th) edition is maintained by the World Health Organization (WHO) and translated into 42 languages. ICD-10 provides about 13,000 classes for the encoding of diseases and reasons of encounter. Originally created for epidemiological purposes, ICD now constitutes the most widely used disease encoding system and is globally used as a common basis for health statistics. In many countries, the ICD is also employed as a basis for Diagnosis Related Groups (DRG) used for billing. DRGs group patients that are clinically similar and are therefore expected to use the same healthcare resources.

ICD has a simple but efficient architecture. Partitioned into 22 chapters (*Infections*, *Neoplasms*, *Blood Diseases*, *Endocrine Diseases*, etc.), its nodes denote classes of diseases and related problems. This means that each individual disease falls into a category that has a unique code, e.g. the myopia of the second author of this paper can be encoded by H52.1. ICD classes are hierarchically arranged into up to five levels. The hierarchy-building relation is the *is-a* (subclass) relation, expressing that each member of a class is also member of any parent class. ICD axiomatically assumes that sibling classes do not overlap. This warrants that no class has more than one parent class and that there is exactly one terminal class for each entity to be classified, hence its characterization as a "classification". The simple cause for this is to prevent that one disease is counted twice. In order to avoid gaps, residual categories ("not elsewhere classified") were created. Additional attributes of ICD classes are inclusion and exclusion statements, and in one chapter also glossary-like free text definitions. Inclusion statements list more specific diseases that are contained in the same class, while classes with exclusion statements segregate certain conditions from a class, thus assigning them to a different class.

ICD's scope extends the realm of diseases as it also includes injuries and external causes of health problems, signs and symptoms, and any kind of conditions that justify the encounter with health professionals. Figure 1 displays an excerpt of ICD relating to certain types of eye disorders, which are subclasses of the three-digit category H52. Note the exclusion under H52.1 and the inclusions under H52.5. The former must be coded in a different branch, while the latter names more specific disorders for which no separate codes are available. Note also that H52.6 constitutes the complement to H52.0-H52.5, and that H52.7 corresponds to H52 and expresses that the coder lacks details that would enable to use a more specific code.



| | |
|---|---|
| H52 | Disorders of refraction and accommodation |
| H52.0 | Hypermetropia |
| H52.1 | Myopia<br>*Excludes:* degenerative myopia ( H44.2 ) |
| H52.2 | Astigmatism |
| H52.3 | Anisometropia and aniseikonia |
| H52.4 | Presbyopia |
| H52.5 | Disorders of accommodation<br>Internal ophthalmoplegia (complete)(total)<br>Paresis }<br>Spasm } of accommodation |
| H52.6 | Other disorders of refraction |
| H52.7 | Disorder of refraction, unspecified |

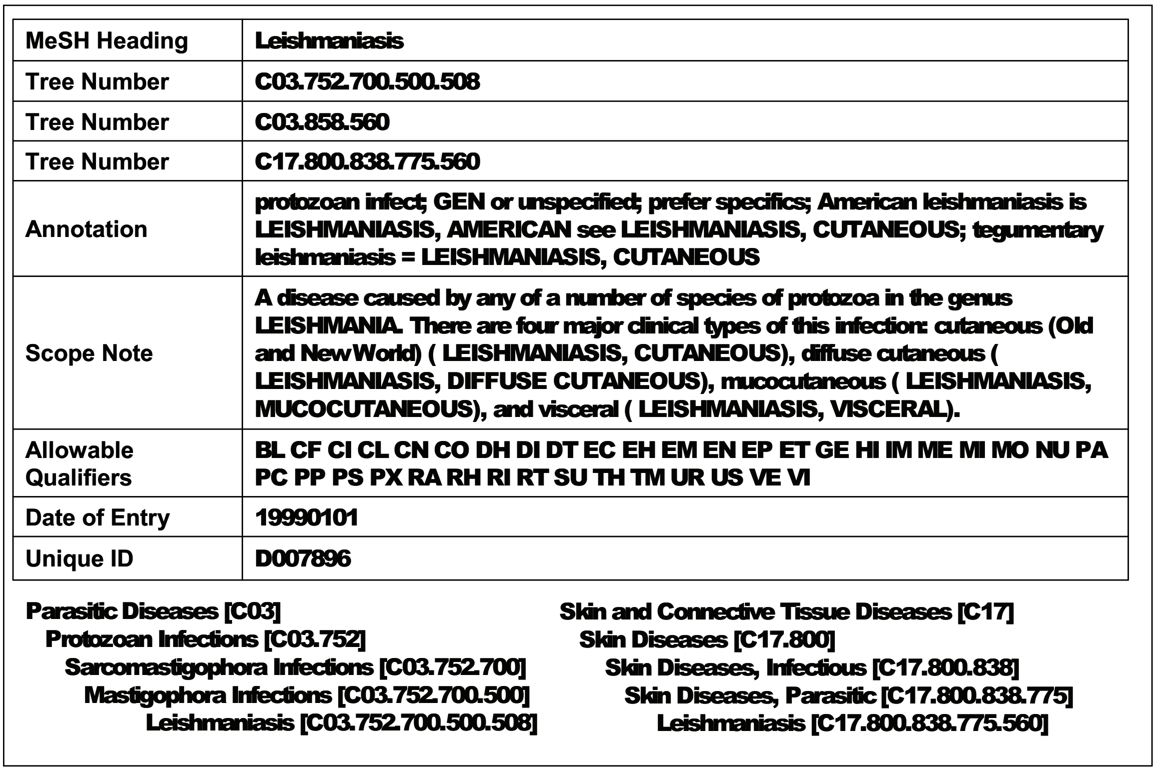
**Figure 1 -** Excerpt of the International Classification of Diseases, 10th version (ICD-10).

## The Medical Subject Headings (MeSH)

The Medical Subject Headings (MeSH) (Nelson 2007, MESH 2008), edited and maintained by the U.S. National Library of Medicine (NLM), consist of a controlled vocabulary used for indexing the content of health related documents, above all literature abstracts in the life science literature database MEDLINE with nearly 20 Million citations (Nelson 2007, PubMed). MeSH is available in 41 languages.

MeSH is partitioned at its uppermost level into 16 branches (*Anatomy*, *Organisms* and *Diseases*, among others). MeSH's nodes are named "headings" and denote a standardized meaning of a group of medical terms. In contrast to the tree-like hierarchy of ICD, MeSH headings are placed in multiple hierarchies. The hierarchical order is based on the principle that all documents indexed by a given heading are also relevant for any parent descriptor. These informal links are also characterized by the name "broader/narrower"). So is the MeSH heading *Leishmaniasis* both part of the hierarchy *Parasitic Diseases* and the hierarchy *Skin and Connective Tissue Diseases*, as depicted by Figure 2. Thus, documents on leishmaniasis are found in a MEDLINE query for parasitic diseases just as in a query for skin diseases. MeSH headings have, in addition to their unique identifier, a so-called tree number for each hierarchical context.

Headings are furthermore specified by a textual definition, a so-called scope note. Additional attributes are entry terms (synonyms or more specific terms) and allowable qualifiers, such as prevention, therapy, and others in the case of diseases, pathogenicity in case of organisms.



| MeSH Heading | Leishmaniasis |
|---|---|
| Tree Number | C03.752.700.500.508 |
| Tree Number | C03.858.560 |
| Tree Number | C17.800.838.775.560 |
| Annotation | protozoan infect; GEN or unspecified; prefer specifics; American leishmaniasis is LEISHMANIASIS, AMERICAN see LEISHMANIASIS, CUTANEOUS; tegumentary leishmaniasis = LEISHMANIASIS, CUTANEOUS |
| Scope Note | A disease caused by any of a number of species of protozoa in the genus LEISHMANIA. There are four major clinical types of this infection: cutaneous (Old and New World) ( LEISHMANIASIS, CUTANEOUS), diffuse cutaneous ( LEISHMANIASIS, DIFFUSE CUTANEOUS), mucocutaneous ( LEISHMANIASIS, MUCOCUTANEOUS), and visceral ( LEISHMANIASIS, VISCERAL). |
| Allowable Qualifiers | BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH TM UR US VE VI |
| Date of Entry | 19990101 |
| Unique ID | D007896 |

Parasitic Diseases [C03]
Protozoan Infections [C03.752]
Sarcomastigophora Infections [C03.752.700]
Mastigophora Infections [C03.752.700.500]
Leishmaniasis [C03.752.700.500.508]

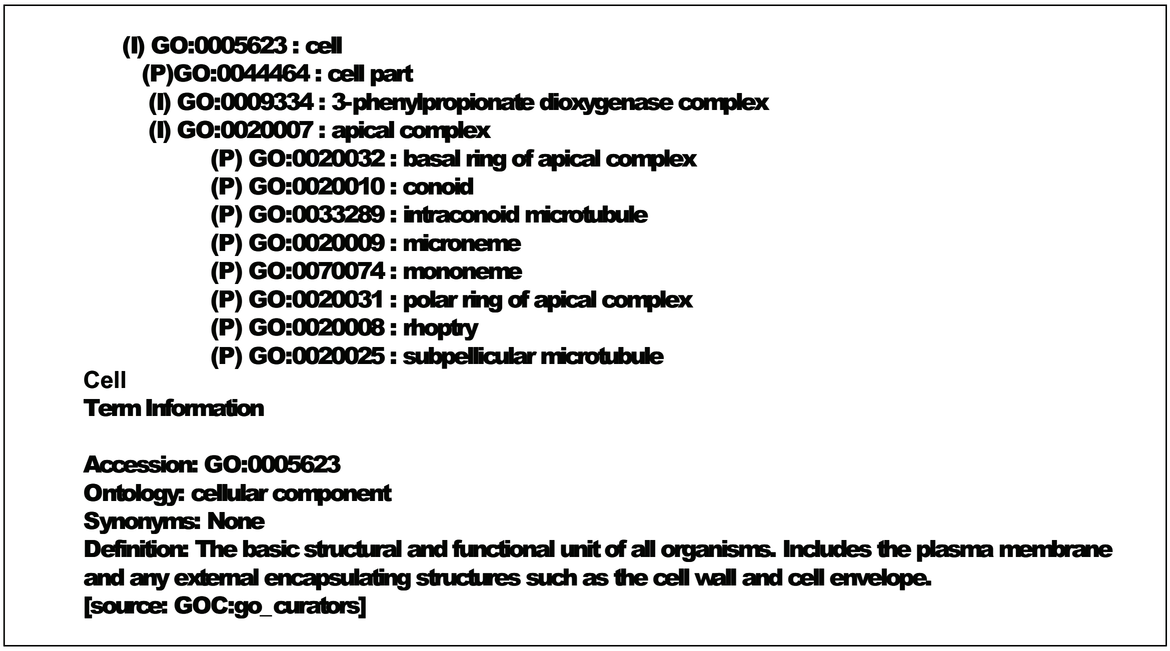
Skin and Connective Tissue Diseases [C17]
Skin Diseases [C17.800]
Skin Diseases, Infectious [C17.800.838]
Skin Diseases, Parasitic [C17.800.838.775]
Leishmaniasis [C17.800.838.775.560]

**Figure 2 -** The MeSH entry for "Leishmaniasis". The table provides definition and attributes. Two of the "trees" in which this heading is inserted are displayed at the bottom.

## The Gene Ontology

The Gene Ontology (GO) (GO 2008) is maintained by the Gene Ontology Consortium, which originally created it to support shared annotations of genomic data in three model organism (Drosophila, Yeast, Mouse) databases. Since then, its scope has been broadened so that it now encompasses all biology independent of the characteristics of specific organisms. In contrast to its name, GO is not an ontology of genes. Instead, it provides semantic identifiers that standardize the description of data on genes or gene products (e.g., proteins) along three dimensions: (i) in which cell compartment a gene is expressed (e.g. the mitochondrium), (ii) with which functions a protein is associated (e.g. signaling), and (iii) in which biological processes a protein participates (e.g. mitosis). Thus GO is able to support queries across the databases consortium members maintain, thus facilitating the access to the knowledge discovered by them.

Like MeSH, the Gene Ontology is partitioned in disjoint branches at its uppermost level. The three branches *Cellular Component*, *Biological Process*, and *Molecular Function* outline its scope. Each branch consists in a multiple hierarchy, of a totality of 24,500 nodes, called GO *terms*. As much as GO's architecture may resemble MeSH at first sight, there are crucial differences that may justify its qualification as an ontology. First of all, its nodes are more than semantic descriptors. In contrast to MeSH headings, GO terms represent classes of real entities. For instance, the (abstract) class *Cell Nucleus* has all (material) cell nuclei in the world as members. GO terms are characterized by identifiers, so-called accession numbers, and have synonyms and definitions as additional attributes. Another difference compared to MeSH is the semantic explicitness of links. Instead of "broader / narrower", GO provides two precisely labeled relations: *is-a* and *part-of*. The former signifies that every entity that is member of one class is also member of all parent *is-a* classes, just as in ICD. *Part-of* has to be interpreted in the sense that every entity that is member of one class is part of some entity that is member of all of its *part-of* classes. Figure 3 presents an entry from GO referring to the class *Cell*.



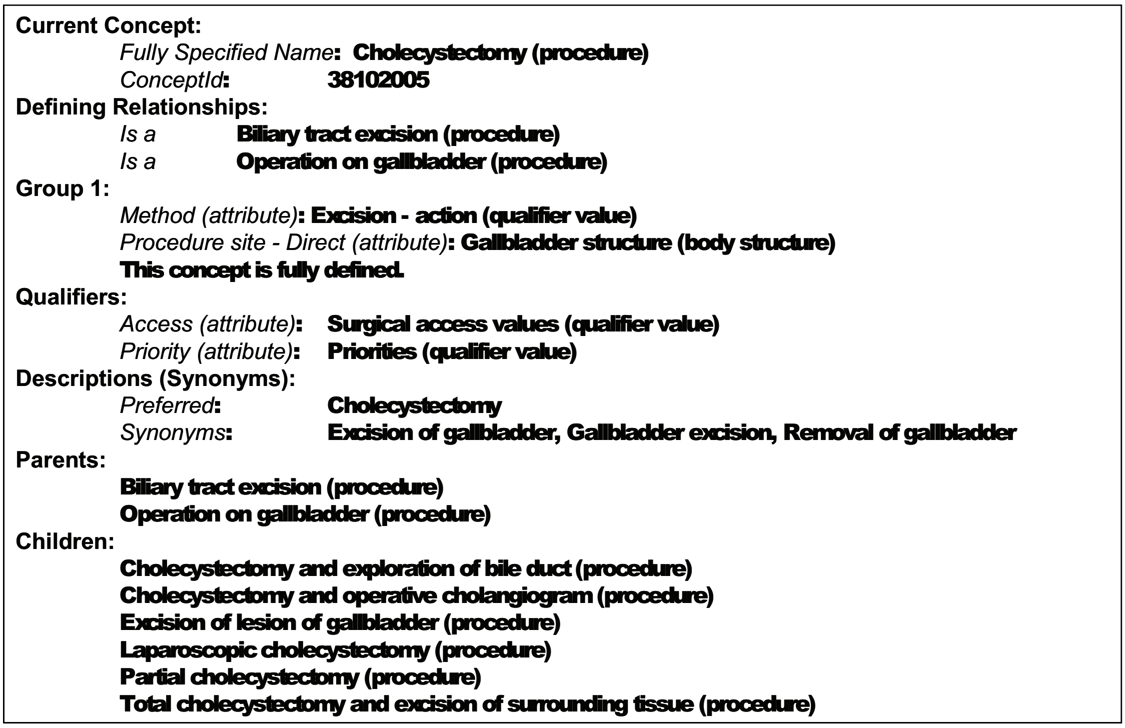
```
        (I) GO:0005623 : cell
          (P)GO:0044464 : cell part
           (I) GO:0009334 : 3-phenylpropionate dioxygenase complex
           (I) GO:0020007 : apical complex
                  (P) GO:0020032 : basal ring of apical complex
                  (P) GO:0020010 : conoid
                  (P) GO:0033289 : intraconoid microtubule
                  (P) GO:0020009 : microneme
                  (P) GO:0070074 : mononeme
                  (P) GO:0020031 : polar ring of apical complex
                  (P) GO:0020008 : rhoptry
                  (P) GO:0020025 : subpellicular microtubule
Cell
Term Information

Accession: GO:0005623
Ontology: cellular component
Synonyms: None
Definition: The basic structural and functional unit of all organisms. Includes the plasma membrane
and any external encapsulating structures such as the cell wall and cell envelope.
[source: GOC:go_curators]
```

**Figure 3 -** Entry of the class *Cell* in the Gene Ontology (GO).
(I) stands for is-a hierarchies, (P) for part-of hierarchies.

## SNOMED-CT

The Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) (Spackman 2004, IHTSDO 2008) is a comprehensive terminology, created to cover the whole patient record. It also comprises body structures, procedures and relevant health-related aspects, including also social context. SNOMED CT is the result of the merger of the UK Clinical Terms version 3 (also called Read Codes) with SNOMED RT (Reference Terminology) (Spackman 1997), the latter being built on several generations of precursor versions (Cornet 2008). Since April 2007, SNOMED CT is owned, maintained, and distributed by the International Health Terminology Standards Development Organization (IHTSDO), a non-for-profit association based in Denmark. SNOMED CT products and services are open for researchers but its use for clinical coding or other commercial usages is restricted to its licensees (currently ten countries and some companies). SNOMED CT is officially available in English and Spanish, while other translations (e.g. Dutch, Danish, Swedish) are currently in the works.

From a structural point of view, SNOMED CT provides multiple *is-a* hierarchies containing about 310,000 nodes. SNOMED CT nodes, referred to as *concepts*, denote mostly classes of individual entities (such as diseases, procedures, lab results, drugs etc., but also particulars like geographic entities), although there is still some controversy of whether the referents of, e.g., the concept *Chest Pain*, are the objects themselves (e.g. the pain in the chest of a given patient) or their mention in the health record (e.g. the entry "chest pain"). SNOMED CT concepts are uniquely identified by numeric keys together with their fully specified names. Most SNOMED CT concepts include several synonyms (named "descriptions"), and, in just a few cases, also free-text definitions. Additional attributes are SNOMED qualifiers, which provide optional refinements for concepts, e.g. *Laterality* for anatomy or *Severity* for diseases.



**Current Concept:**
*Fully Specified Name*: **Cholecystectomy (procedure)**
*ConceptId*: **38102005**
**Defining Relationships:**
*Is a* **Biliary tract excision (procedure)**
*Is a* **Operation on gallbladder (procedure)**
**Group 1:**
*Method (attribute)*: **Excision - action (qualifier value)**
*Procedure site - Direct (attribute)*: **Gallbladder structure (body structure)**
**This concept is fully defined.**
**Qualifiers:**
*Access (attribute)*: **Surgical access values (qualifier value)**
*Priority (attribute)*: **Priorities (qualifier value)**
**Descriptions (Synonyms):**
*Preferred*: **Cholecystectomy**
*Synonyms*: **Excision of gallbladder, Gallbladder excision, Removal of gallbladder**
**Parents:**
**Biliary tract excision (procedure)**
**Operation on gallbladder (procedure)**
**Children:**
**Cholecystectomy and exploration of bile duct (procedure)**
**Cholecystectomy and operative cholangiogram (procedure)**
**Excision of lesion of gallbladder (procedure)**
**Laparoscopic cholecystectomy (procedure)**
**Partial cholecystectomy (procedure)**
**Total cholecystectomy and excision of surrounding tissue (procedure)**

**Figure 4 -** SNOMED CT's definition of Cholecystectomy. Note that this concept is fully defined, i.e. the combination *Method – Excision Action* with *Procedure Site – Gallbladder Structure* is a sufficient condition for *Gallbladder*

SNOMED CT offers also 50 link types, called *linkage concepts*. They are used in what can be considered the most important distinctive criterion of SNOMED CT, *viz.* the use of a rich ontology representation language compatible which the Semantic Web standard OWL-DL (description logics) (Bechhofer et al. 2004). Description logics allow the definition of new classes using existing classes and relations. As shown in Figure 4, *Cholecystectomy* is fully defined as a new class, using the existing classes *Excision* and *Gallbladder*, together with the links (relations) *Method* and *Procedure Site*. This means that each and every excision procedure at some gallbladder is a cholecystectomy and *vice versa*.

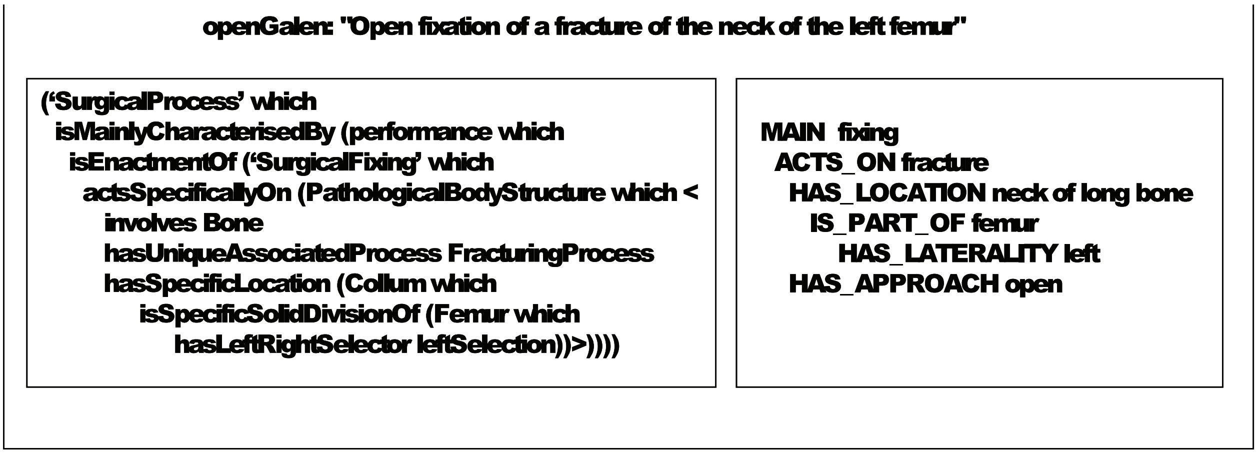
The creation of complex expressions based on SNOMED concepts and obeying a formal syntax and semantics is called coordination. This can be done at the moment of coding (pre-coordination) or beforehand, by introducing new concepts into the terminology (post-coordination) (Chen 2005).

## openGALEN

the Generalized Architecture for Languages, Encyclopaedias and Nomenclatures (openGALEN) provides an open-source clinical ontology which had been developed in the nineties as an outcome of a series of European projects (GALEN) (Rector 2003). It is aimed at clinical applications and contains about 25,000 nodes (concepts) and 26 link types (relations). openGALEN concepts are arranged in multiple *is-a* hierarchies, too. It uses a description logic language called GRAIL (GALEN Representation and Integration Language), which allows the definition of classes similar as in SNOMED CT but provides a richer syntax, as can be seen in the example of Figure 5 which describes a fixation of the left femur neck fracture. The GALEN model is split into the following items:

- a high level ontology, which provides an overall categorization framework,
- the common reference (CORE) model, containing reusable definitions from anatomy, diseases, surgical procedures, symptoms, etc.,
- detailed extensions for specific subdomains, such as surgery.

Its purpose is therefore similar to SNOMED CT, but it has never reached its scope and granularity. However, openGALEN can be regarded as the pioneer of the use of formal logics in biomedical terminologies. Its most important use case was the development of the French medical procedure classification CCAM (Trombert-Paviot 2000).



openGalen: "Open fixation of a fracture of the neck of the left femur"

```
('SurgicalProcess' which
  isMainlyCharacterisedBy (performance which
    isEnactmentOf ('SurgicalFixing' which
      actsSpecificallyOn (PathologicalBodyStructure which <
        involves Bone
        hasUniqueAssociatedProcess FracturingProcess
        hasSpecificLocation (Collum which
          isSpecificSolidDivisionOf (Femur which
            hasLeftRightSelector leftSelection))>))))
```

```
MAIN  fixing
  ACTS_ON fracture
    HAS_LOCATION neck of long bone
      IS_PART_OF femur
        HAS_LATERALITY left
    HAS_APPROACH open
```

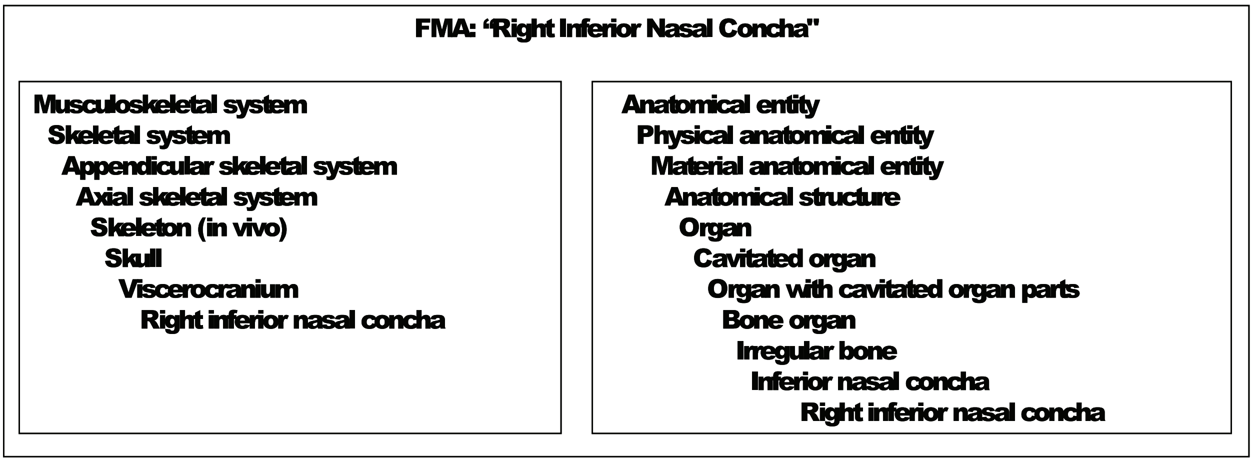
**Figure 5 -** OpenGALEN detailed entry defining a type of fracture fixation. Left: description logics like representation (GRAIL syntax). Right: close-to-user syntax devised for facilitating the definition of surgery concepts.

## Foundational Model of Anatomy

The Foundational Model of Anatomy (FMA) (FMA 2008) is a biomedical ontology that provides declarative knowledge on the macroscopic structure of the human body. It was originally developed for describing anatomy images for didactic purposes. Like in GO, FMA nodes are arranged in two hierarchies, the *Anatomy Taxonomy*, which is an *is-a* monohierarchy, and the multihierarchical *Part-Whole Network* that employs *part-of* as an ordering relation. Additional attributes are identifier, synonyms, and additional relations (e.g. *has-dimension*, *has-mass*, *adjacent_to*, etc.). The FMA is represented in the frame formalism, which makes less rigid ontological assumptions and therefore can only be incompletely translated to Description Logics.

FMA nodes are named *classes* or *types*, which underlines its commitment to real word entities rather than to term meanings. However, FMA explicitly states that its classes extend to *canonical* anatomical entities, just as in anatomic atlases, which results in the description of an ideal human body without any deficiency or anatomical alteration of malformation. This sometimes causes inconsistencies such as the one with the FMA axiom that states that "*Lower gastrointestinal tract has-part Appendix*". It clearly conflicts with frequent clinical situations.

Figure 6 shows the class *Right Inferior Nasal Concha*, stating that it is part of *Skull* which is on its turn part of *Skeleton* and so on. Another entry defines it as a subtype of *Inferior Nasal Concha* which is a *Bone Organ* which is a subtype of many other classes, including the most general class *Anatomical Entity*.



FMA: "Right Inferior Nasal Concha"

```
Musculoskeletal system
  Skeletal system
    Appendicular skeletal system
      Axial skeletal system
        Skeleton (in vivo)
          Skull
            Viscerocranium
              Right inferior nasal concha
```

```
Anatomical entity
  Physical anatomical entity
    Material anatomical entity
      Anatomical structure
        Organ
          Cavitated organ
            Organ with cavitated organ parts
              Bone organ
                Irregular bone
                  Inferior nasal concha
                      Right inferior nasal concha
```

**Figure 6 -** The Foundational Model of Anatomy's definition of the Right Inferior Nasal Concha.

## Efforts to gather different sources of biomedical knowledge

### Rationale

Considerable efforts have been devoted on the one hand to align the numerous and largely overlapping biomedical terminologies and ontologies, but also to prevent the anarchic proliferation of BMTOs by establishing principles for the coordinated development of interoperable resources on the other hand. We will describe the Unified Medical Language System (UMLS) and the OBO (Open Biological Ontologies) Foundry. Whereas UMLS is an example for the former strategy, OBO embodies the latter approach.

### The Unified Medical Language System UMLS Metathesaurus

The richest source of biomedical terminologies, thesauri, classification systems and ontologies is constituted by the Unified Medical Language System (UMLS) Metathesaurus (Nelson 2006, UMLS 2008), initiated in 1986 by the U.S. National Library of Medicine (NLM), with the purpose to integrate information from a variety of disparate terminological sources. The UMLS now covers over 2 million names for about 1 million biomedical concepts from more than 120 BMTOs, as well as 12 million relations among these concepts (Bodenreider 2004). Apart from openGALEN, all the above described systems are included in the UMLS Metathesaurus, together with many others, covering organisms, drugs, chemicals, devices, procedures etc.

Besides facilitating transparent access to the sources (through the provision of raw files and online services), the main achievement of the UMLS Metathesaurus lies essentially in the following:

• each node of the source BMTO is retrospectively mapped to a Metathesaurus concept, each of which has a unique identifier, called CUI (*Concept Unique Identifier*). These mappings are regularly updated by manual effort. They enable the bridging between different source BMTOs. As a consequence, links between source nodes are mapped to links between CUIs, called semantic relations. Applications using them can therefore take advantage of concept linkages from both directions;

• each Metathesaurus concept is categorized by at least one semantic type from the UMLS Semantic Network, an overarching conceptual umbrella over the biomedical domain (McCray 2003). A tree of 135 semantic types, linked by *is-a* relations forms the backbone of this Semantic Network. Additionally, the network includes a hierarchy of 53 associative relationships (e.g., *location_of*, *treats*) which are used to form 612 triples (e.g., *Tissue, Diagnostic Procedure,* etc.) from which 6,252 additional triples can be inferred. These triples are interpreted as domain / range restriction of the relations.

### The Open Biomedical Ontologies (OBO) Foundry

Created in 2003, OBO, the Open Biomedical Ontologies (OBO 2008) platform evolved as a library of online, public-domain biomedical ontologies. On this basis, the OBO Foundry initiative developed a set of shared principles regulating the development of biomedical ontologies (Smith 2007). The coverage of the OBO foundry comprises several anatomy ontologies (including the FMA), the Gene Ontology, as well as specialized ontologies of biochemistry (ChEBI), phenotypes (PATO), sequences (SO), and investigation techniques (OBI). Currently, more than 50 ontologies are listed as candidates for the OBO Foundry.

The OBO Foundry propagates two representation languages. Besides OWL-DL there is a proprietary format (OBO-EDIT 2009) in which most OBO ontologies are encoded.

Just as in the Gene Ontology, nodes in OBO ontologies denote classes of entities in the real world. Links between these classes are interpreted as existentially quantified links; for instance, *A part_of B* means that every instance of *A* is part of some instance of *B* (but not *vice-versa*). OBO main relations (*is_a*, *part_of*, *integral_part_of*, *proper_part_of*, *located_in*, *contained_in*, *adjacent_to*, *transformation_of*, *derives_from*, *preceded_by*, *has_participant*, *has_agent*, *instance_of*) have been provided with consistent and unambiguous formal definitions (Smith 2005).

## Discussion

We have described a sample of BMTOs which *pars pro toto* represent the variety of semantic standards in biology and medicine. Our purpose was to give the readers an overview of the substantial efforts being carried out to describe terms and the entities they denote in order to support querying and intelligent data and knowledge processing in general as well as specific applications. Moreover, we present these efforts according to their expressivity in an increasing sequence. One aspect directly linked to expressivity is scaling and coverage, since BMTOs encoded in expressive formalisms should be employed in more restricted domains, while for informal terminologies this constraint is not relevant.

Though it seems straightforward in theory to distinguish terminologies from formal ontologies, in practice the distinction is less clear. The key idea is that terminologies are much more related to organizing domain terms only (as a huge amount of terms is at the core of any subfield of Biomedicine) – while ontologies give a more precise account which is based on formal logic and as much as possible independent of human language. A typical instance for this is SNOMED CT. Its predecessors have their roots in a compositional standardized nomenclature (SNOMED Int.) and a clinical coding system (NHS Clinical Terms Version 3) but its current redesign is being increasingly guided by ontological principles. On the contrary, BMTOs such as ICD and MeSH can be considered more established as important and globally successful use cases have existed for decades. ICD has the longest history and most widespread dissemination due to its simple architecture and the early need for health or disease statistics. Endorsed by the WHO and by national bodies, its objective has then increasingly included clinical epidemiology, health management, quality assurance and billing in many countries, including Brazil. MeSH, on the other hand, has a complex multihierarchical structure tailored to querying in biomedical text collections.

A clear trend that can be observed is the increasing adoption of Semantic Web languages and formalisms, particularly the ontology language OWL and its subset OWL-DL, the latter being adapted to the needs of machine reasoning. The main advantages of using inferencing machinery such as the ones available for description logics is to be able to check the entailments of the axioms contained in the ontology, to support knowledge-intensive queries, to calculate semantic equivalences of syntactically different expressions and to disambiguate natural language utterances. Although the currently available classifiers run into scalability problems with more expressive (and therefore more interesting) formalisms, the fact that standards like description logic and OWL exist pays off for applications that require in-depth knowledge about a small number of subfields. As could be seen in the previous section, many of the BMTOs presented have undergone endeavors to shift from their original format to description logic: SNOMED was a pure terminology in the past; FMA has already partially shifted from frames to OWL, and there is a tendency for OBO ontologies to adopt OWL-DL, although a proprietary format had been developed in the past and is still largely used. Interestingly, openGALEN had been conceived from the very beginning to use a logic-based, DL-like language. It therefore can be proud of having first axiomatized significant amounts of medical terms, and the lessons learned are highly valuable for biomedical ontology engineering till this date.

The sheer amount of BMTOs describing partly overlapping domains for similar or different use cases based upon different formalisms, philosophies and (tacit) assumptions has been identified as a problem already in the eighties. Since then, large efforts have been invested into the UMLS Metathesaurus by which an increasing number of heterogeneous sources are annually cross-mapped and categorized. Two constraints must, however, be stated. Firstly, the mapping cannot be more expressive than the least expressive source BMTO, and secondly, the usefulness of the UMLS for practical applications is hampered by the fact that many of its sources are subject to individual licensing.

In contrast, the OBO sources are completely in the public domain and can be accessed by everyone. This, at least partly, explains their success and the high level of biological expertise being invested in their construction and maintenance.

In Figure 7, some key features of the described BMTOs and gathering efforts are summarized, showing their scope, coverage, volume, formalism and usages.

| Name | Scope | Formalism | Number of Nodes | Applications | URL |
|---|---|---|---|---|---|
| ICD | Diseases | Classification, strict is-a | Around 13,000 classes | Health Statistics, Epidemiology, Health Reporting Billing | www.who.int/classifications/apps/icd/ |
| MESH | Medicine, Nursing, Dentistry, Veterinary Medicine, Health Care Systems, Preclinical Sciences | Terminology Semantic Networks | 24,767 (2008) terms | Indexing articles from 4,800 of the world's leading biomedical journals for the MEDLINE/ PubMED® database | www.pubmed.gov |
| SNOMED | Everything encoded in the electronic health record | Description Logic | 311,000 concepts (2008) | Information about a patient's medical history, illnesses, treatments, and laboratory results | www.ihtsdo.org |
| GO | Cellular components, molecular functions, biological processes | OBO/OWL | 24,500 terms (2008) | Research on genes, proteins | www.geneontology.org |
| GALEN | Anatomy, surgical deeds, diseases, health care | Description logic-like language GRAIL | Over 10,000 | Electronic healthcare records, clinical user interfaces, decision support systems, knowledge access systems, natural language processing | www.opengalen.org |
| FMA | Anatomy content | Frames and (partly) OWL | 75,000 classes | Education, biomedical research | http//sig.biostr.washington.edu/projects/fm/AboutFM.html |

| | | | | | |
|---|---|---|---|---|---|
| OBO | Bioinformatics and molecular Biology | OBO/ OWL / OBO_XML / RDF | 60 ontologies | Used as a repository and an unified schema to interoperate biomedical projects | www.obofoundry.org |
| UMLS | Biomedical and health related concepts | Semantic Networks | Over 1 million concepts | scientific literature, guidelines, and public health data, natural language processing | http://www.nlm.nih.gov/research/umls/ |

**Figure 7 -** BMTOs, OBO, UMLS and some of their key features.

## Open issues and challenges

A new era for biomedical informatics is currently unfolding. Besides the algorithms employed in gene research, ontologies are esteemed as an increasingly hot topic. There is already an active community researching and benefitting from semantic interoperability through ontologies, as ontologies are increasingly used for the annotation of research data in Molecular Biology and Genomics. The emerging reusable vocabularies prove useful for describing biomedical data and more and more kinds of applications. The precise capture of biological knowledge in a computational means enables the creation of systems capable of meeting robust requirements as required by biologists, medical researchers and practitioners: easy access to texts and databases containing detailed data, information, and statements; sound and complete reasoning, faster development of decision support systems for a broad range of use cases, etc. However, some hard challenges have to be overcome for the field to become mature.

A first issue resides in modeling. The subtle aspects that have to be described in biomedical ontologies usually requires toplevel ontologies and ontology assessment techniques (Guarino 2000) to come into play, otherwise reasoning resulting from it can fail. An emblematic example can be seen in the relations between the main classes Physical object and Amount of matter. The famous WordNet ontology (Miller 1995), used for informatics researchers particularly from the field of Natural Language Processing, states that *Physical Object is-a Amount of Matter*. On the other hand, *Pangloss*, a large ontology mainly used for translation between languages, describes the two classes in the opposite way, *Amount of Matter* being a superclass of *Physical Object*. Indeed, (Guarino & Welty 2000) state that both interpretations are wrong: Every instance of *Physical Object* is *constituted by* one or more instances of *Amount of Matter*. Yet there is no superclass relation, which can easily be seen by analyzing meta-properties like unity, rigidity or identity). In the biomedical field, such inaccuracies also occur: The Gene Ontology, in an earlier version, included the axiom *Cell has-part Axon*. At a closer investigation, this definition led to ambiguities and underspecifications, since there are cells without axons and axons without cells at least play a role in the lab (Schulz 2004). These two examples stress the need for more formality and semantic richness in biomedical ontologies.

Another key issue, which can also be seen in the first example, is integration. As the number of biomedical ontologies increases, many applications need to employ more than one ontology, which leads to a series of significant consequences. Undeniably, this is not an issue only for biomedicine; the main obstacles for knowledge reuse in the computer science mainstream come from knowledge heterogeneity. Knowledge is naturally diverse in its various features: form, expression, representation formalisms, language, syntax, contents, meaning, modeling principles, practices and standards, points of view, perspectives, uses, granularity, terminology, premises, not to mention that some unions of them can be hard for reasoning, regarding computational resources. Although ontologies (in a stricter sense, viz. statements about what is always true and univocally accepted) only cover a clear-cut segment of what is commonly understood by knowledge representation, these varieties will always have an impact on crucial design decisions and will pose subtle questions for ontology applications. Dealing with heterogeneity has become a recurrent and challenging research issue for ontology employment and, on the other hand, also a good source of ontology usage, e.g. for problems like information integration of heterogeneous ontologies, such as querying for hotels, whose descriptions are distinctly described in each of many systems.

Granularity is a particular issue that has also deep impact on the integration of biomedical ontologies (Schulz 2009). There is a hope to see medical and biological research join ontologies at the level of cell, anatomy, drugs, etc. These communities might need different granularities or even different views of the same ontology. Another challenge related to integration is how to handle existing biomedical ontologies that contain overlapping information, providing different views on a certain subdomain or covering different domains.

To enable ontology integration, plenty of research is taking place. A description of them is summarized in (Freitas et al. 2007) and presented in depth in (Stuckenschmidt et al. 2000).

On the actual application of biomedical ontologies, text processing is surely one of the key ones. A very popular use case is the automatic assignment of MeSH terms to user queries in PubMed. Another one is the automated extraction of information related to individual genes or proteins from scientific texts. The electronic health record and consumer platform also constitute a wide field for text and knowledge processing. To tackle this issue, systems may rely on information extraction and text mining systems (Muslea 1999, Ananiadou 2006). However, many questions remain unanswered, and the combination of high quality text analysis methodologies with high-expressive and well-standardized ontologies constitutes an ongoing research challenge.

## Bibliographic references

Ananiadou S, McNaught J. Text Mining for Biology and Biomedicine, chapter Introduction. Norwood, MA: Artech House Publishers; 2006.

Antoniou G, van Harmelen F. A Semantic Web Primer. MIT Press, Cambridge; 2004.

Bechhofer S, Harmelen F, Hendler J, Horrocks I. OWL Web Ontology Language Reference. W3C Recommendation; 2004 . http: //www.w3.org/TR/2003/PR-owl-ref-20031215/. Last accessed February 3, 2009.

Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology, Oxford University. 2004 January 1; 32(1) Suppl.1: D267-D270.

Chen H, Fuller SS, Friedman C, Hersh W. Knowledge Management and Data Mining in Biomedicine Series: Integrated Series in Information Systems , New York: Springer; 2005. Vol. 8.

Cornet R. and de Keizer N. Forty years of SNOMED: a literature review. BMC Medical Informatics and Decision Making. 2008; 8(Suppl 1): S2.

FMA - Foundational Model of Anatomy sig.biostr.washington.edu/projects/fm Accessed in April 2008. Berners-Lee T, Hendler J, Lassila O, editors. The Semantic Web, Scientific American. 2001; 28-37.

Freitas F, Stuckenschmidt H, Noy N. Ontology Issues and Applications: Guest Editors' Introduction. Journal of the Brazilian Computer Society. 2005; 11(2).

GO - The Gene Ontology http://amigo.geneontology.org/cgi-bin/amigo/go.cgi. Last accessed February 3, 2009.

Gruber T. A translation approach to portable ontologies. Knowledge Acquisition. 1995; 5(2):199-220.

Guarino N. Formal ontology in information systems. Proc FOIS'98. 1998; 3-15.

Guarino N, Welty C. A formal ontology of properties. In: Knowledge Engineering and Knowledge Management - Proceedings of 12th International Conference EKAW 2000. France: Springer; 2000.

IHTSDO - International Healthcare Terminology Standards Development Organisation. http://www.ihtsdo.de. Last accessed February 3, 2009.

Kunierczyk W. Nontological Engineering. Formal Ontology In Information Systems. In: Proceedings of the 4th International Conference FOIS 2006, Amsterdam, The Netherlands: IOS Press; 2006. 39-50.

MESH - Medical Subject Headings, http://www.nlm.nih.gov/mesh/. Last accessed February 3, 2009.

Miller G. WordNet: a lexical database for English. Communications of the ACM; 1995.

Muslea I. Extraction patterns for information extraction tasks: A survey. American Association for Artificial Intelligence (www.aaai.org) he AAAI-99 Workshop on Machine Learning for Information (1999).

Nelson SJ, Powell T, Humphreys LB. The Unified Medical Language System (UMLS) of the National Library of Medicine. Journal of American Medical Record Association. 2006; 61: 40-42.

Nelson SJ, Schulman J. A Multilingual Vocabulary Project - Managing the Maintenance Environment. MeSH Section, National Library of Medicine, Bethesda, Maryland; 2007.

OBO - Open Biomedical Ontologies. http: //www.obofoundry.org. Last accessed February 3, 2009.

OBO-EDIT. An Introduction to OBO Ontologies http: http//oboedit.org/docs/html/An_Introduction_to_OBO_Ontologies.htm. Last accessed February 3, 2009.

OpenGalen Foundation. http: //www.opengalen.org. Last accessed February 3, 2009.

PubMed. http: //www.ncbi.nlm.nih.gov/pubmed/. National Library of Medicine. Last accessed February 3, 2009.

Rector A, Rogers JE, Zanstra PE, Haring E. OpenGALEN: Open Source Medical Terminology and Tools. AMIA Annual Symposium Proceedings. 2003; 982.

Rector A. Clinical Terminology: Why is it so hard? Methods of Information in Medicine. 2000; 38(4): 239-52.

Rubin DL, Shah NH, Noy N. Biomedical Ontologies: a functional perspective. Briefing in Bioinformatics. 2008 Jan; 9(1): 75-90.

Schulz S, Hahn U. Mereotopological Reasoning about Parts and (W) holes in Bio-Ontologies, In: C. Welty and B. Smith, editors, Formal Ontology in Information Systems. Collected Papers from the 2nd International FOIS Conference, New York, NY: ACM Press, 2001; 210-21.

Schulz S, Hahn U. Towards the ontological foundations of symbolic biological theories. Artificial Intelligence in Medicine. 2007 Mar; 39(3): 237-50.

Schulz S, Boeker M, Stenzhorn H, Niggemann J. Granularity Issues in the Alignment of Upper Ontologies. Methods of Information in Medicine. 2009. Accepted for Publication.

Smith B, Ashburner M, Rosse C, Bard C, Bug W, Ceusters W, Goldberg L J, Eilbeck K, Ireland A, Mungall CJ, The OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann R H, Shah N, Whetzel PL and Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, Nature Biotechnology. 2007; 25: 1251-5.

Smith B, Mejino JLV, Schulz S, Rosse C. Anatomical Information Science. In: COSIT 2005: Spatial Information Theory. Foundations of Geographic Information Science, New York: Springer. 2005; 149-64

Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C., Relations in Biomedical Ontologies. Genome Biology. 2005; 6(5).

Spackman KA, Campbell KE, Côté RA. SNOMED RT: A reference terminology for health care. In Masys DR (Ed.) , The Emergence of Internetable Health Care: Systems that Really Work. Proceedings of the 1997 AMIA Annual Symposium, 640-644. Philadelphia: Hanley & Belfus, Inc. 1997 .

Spackman KA. SNOMED CT milestones: endorsements are added to already-impressive standards credentials. Healthcare Informatics. 2004; 21: 54-6.

Stuckenschmidt H, Wache H, Vogele T, Visser U. Enabling technologies for interoperability. In Visser, U. and Pundt, H. editors, Workshop on the 14th International Symposium of Computer Science for Environmental Protection, Bonn, Germany. TZI, University of Bremen. 2000; 35–46.

Trombert-Paviot B, Rodrigues JM, Rogers JE, Baud R, van der Haring E, Rassinoux AM, Abrial V, Clavel L, Idir H. GALEN: a third generation terminology tool to support a multipurpose national coding system for surgical procedures. Intern J Med Informatics. 2000 Sep; 58-59: 71-85.

UMLS - Unified Medical Language System http: //www.nlm.nih.gov/research/umls/. Last accessed February 3, 2009.

WHO - International Classification of Diseases, 10th Edition. World Health Organization. http: //www.who.int/classifications/apps/icd/icd10online/ . Last accessed February 3, 2009.

# About the authors

## Frederico Freitas

Received his PhD at the University of Santa Catarina, Brazil, and is currently affiliated to the Center for Informatics at the Federal University of Pernambuco, Brazil (CIn/UFPE). He researched for almost a year at the Department of Informatics of the University of Karlsruhe as a member of the Brazilian-German project "A semantic approach to data retrieval". He published a number of papers at renowned conferences and workshops such as IJCAI and other sponsored by ACM (Association on Computer Machinery) and IEEE (Institute of Electrical and Electronical Engineering). He has also co-chaired two workshop series: WONTO (Workshop on Ontologies and their Applications) in Brazil and BAOSW (Building Applications with Ontologies for the Semantic Web) in Portugal. He co-edited Special Issues on related themes at JBCS (Journal of Brazilian Computer Society) and JUCS (Journal of Universal Computer Science). He currently has collaborations with the University of Paul Cezanne at Marseille and INRIA, Montbonnot, in France and the Universities of Karlsruhe, Freiburg and Mannheim in Germany. His interest areas comprise ontologies, multi-agent systems, knowledge representation, mediation and text mining.

## Stefan Schulz

Holds a medical degree (Heidelberg University, Germany) and is senior researcher and professor at the Institute for Medical Biometry and Medical Informatics of the University Medical Center Freiburg, where he leads the Medical Informatics Research Group. His work focuses on biomedical terminologies and ontologies, biomedical knowledge representation, cross-language medical document retrieval, text and data mining in clinical document repositories, eLearning in Medicine, and health informatics in developing countries. After clinical work in surgery and internal medicine he obtained his doctoral degree in the field of tropical hygiene where he carried out a parasitological field study on in São Luís, Brazil. After obtaining a technical qualification in medical computing, he moved to the University of Freiburg, where he participated in clinical and educational software development projects and participated in several research projects in the field of information extraction, biomedical terminologies, medical language engineering and semantic technologies. He has played a leading role in several EU-funded research projects, authored more than hundred peer reviewed publications and has received several awards. Since 2001 he has repeatedly contributed to Brazilian health in-formatics research projects as a visiting researcher at the Paraná Catholic University (PUC-PR).