# Vision

## 1 Introduction

The major goal of genetic research is the investigation of what genes do, that is, what is the function of each individual gene within the context of a whole organism. The promise that came with genome sequencing was that we would soon gain a better insight to their function which would allow us to identify genes that are involved in human diseases. [2] Today, the full genomic sequence of human as well as of a wide variety of model organisms such as mouse, fly, worm etc. is available. Whilst the study of the effects of modifying individual genes in those organisms provided a genotype to phenotype mapping for several thousand genes, gaining an understanding of complexity of the gene networks and their function is still a major challenge. However, the correlation of such networks with their phenotypic outcomes can lead to an in depth understanding of the disease pathobiology and the genetic basis of human disease which in turn will ultimately improve of our ability to treat it. Understanding the pathobiology of the disease will also provide the basis for identifying biomarkers serving for novel transcriptomics data.

To achieve such a goal though, it is imperative that we adopt a systemic approach in the examination of the phenotype outcomes at various level of granularity such as the molecular, the cellular, the tissue, the organic and the organismal as well as across different species. Such an approach relies heavily on our ability to track and analyse phenotypic overlaps across multiple species in order to generate good predictors of genetic relationships which in turn require phenotype data of high quality, diversity and granularity.

In the post-genomic era and with the advent of functional genomics there has been a rapid increase of quantity and quality of such biomedical data and information, including data deriving from the various *omics* fields. This information explosion has lead to the development of large number of databases and resources, covering a wide variety of biomedical data, with the intention of storing, retrieving, integrating and analyzing them so as to discover new associations and ultimate new biomedical knowledge. If though, such tasks are to be successful, it is essential that the representation of these data is semantically consistent across the various resources that store them.

# 2 The promise of ontologies

In response to this need, the biomedical research community has invested considerable amount of effort and resources in the development and establishment of ontologies that are increasingly becoming successful as information management and integration tools in a variety of scientific fields allowing interoperability and semantic information processing between diverse resources biomedical resources and domains.

Ontologies define terms in controlled vocabularies in order to establish the meaning of concepts and the relationships between them. The resulting artifact allows logical inference within the ontology and permits its use for knowledge discovery. These properties of ontologies permit investigators to carry out key activities: querying data across heterogeneous data sets, data integration and exchange, natural language processing, and automated reasoning. Since the advent of the Gene Ontology (GO) [ref] in 2000, there has been a proliferation of general and domain specific ontologies for a variety of biomedical ontologies. For example, the BioPortal [ref] currently lists 218 ontologies, including the OBO Foundry ontologies [ref], with more than 1.4 million terms.

## 2.1 Formal ontology

The ontology as an approach to semantic standardisation was proposed more than a decade ago and since then has become the dominant methodology used to semantically categorise phenodeviance. The biomedical research community has invested considerable effort and resources in the development and establishment of ontologies that are becoming increasingly successful as information management and integration tools in many disparate scientific fields allowing interoperability and semantic information processing between diverse biomedical resources and domains.

In computer science, an ontology is a specification of a conceptualization of a domain of knowledge [**?**, **?**]. Ontologies commonly distinguish between *classes* (also called *concepts*, *categories* or *universals*) and *individuals* within a domain of knowledge. A class is an entity that can have *instances*, while individuals are entities that cannot be instantiated [**?**]. Examples of individuals include the Eiffel tower or the 2009 Ironman Triathlon in Hawaii, while examples of classes include *Tower* or *Triathlon*. The Eiffel tower can be an instance of the class *Tower*, and the 2009 Ironman Triathlon an instance of *Triathlon*. The meaning of classes is specified by stating what must be true of their instances.

In addition to classes and individuals, ontologies often include *relations*. Relations hold between entities, they are the "the glue that holds things together, the primary constituents of the facts that go to make up reality" [**?**].

In *formal* ontologies, the specification of classes and relations follows the axiomatic-deductive method. Given a set of terms that are used within a domain and whose meaning we wish to specify, we begin by providing *explicit definitions* for some terms, potentially introducing new terms. An explicit definition of a term $t$ is a statement that can replace every occurrence of $t$ in any sentence.

Eventually, a set of *primitive terms* remains that are not further defined. Following the axiomatic method [**?**], using only the primitive terms, we can construct complex sentences. Based on the intended meaning of the primitive terms, we consider some of these sentences true and some of them false in our domain. We select some of the true sentences as *axioms* which provide the core of our ontology. Ideally, the axioms are chosen so that all true sentences in the domain we intend to represent follow by means of logical deduction from the axioms. More commonly, however, only *some aspects* of the intended meaning are formally represented while other aspects are omitted either due to limitations in language expressivity or due to their irrelevance to the problem for which an ontology is developed.

Based on the axioms and definitions, we can use deduction to infer statements that logically follow from the axioms. The process of automatically deducing sentences from axioms is called *automated reasoning*. Automated reasoning allows users of an ontology to carry out key activities: verifying the ontology's consistency, inferring hidden knowledge and thereby performing powerful queries. An ontology is formally inconsistent if there is a statement $\phi$ such that $\phi$ and its negation $\neg\phi$ can be inferred from the ontology's axioms. If an ontology is formally inconsistent, *every* statement can be inferred from the ontology.

Automated reasoning can further determine whether classes in an ontology are unsatisfiable: a class $C$ is unsatisfiable, if it is impossible for the class to have any instances. Unsatisfiable classes in an ontology are commonly the result of a contradictory class definition.

Automated reasoning in the Web Ontology Language (OWL) can be employed to automatically compute the generalization hierarchy underlying an ontology as well as for verification of data consistency and complex queries [**?**, **?**]. Highly efficient automated reasoners are available to process OWL ontologies [**?**, **?**, **?**]. OWL profiles were developed to support even large ontologies by further reducing the expressivity of OWL in order to enable polynomial-time inferences. In particular the OWL EL profile was found to provide the expressivity required for most biomedical ontologies [**?**, **?**], and highly optimized OWL EL reasoners are available or under development to support reasoning over very large ontologies [**?**, **?**].

A high expressivity is required to accurately specify complex axioms that constrain the domain under investigation, and languages with higher expressivity than OWL are often required in the biomedical domain to achieve this goal [**?**, **?**]. On the other hand, automated reasoning over large ontologies and associated datasets benefits from languages with a low complexity of inferences in which complex axioms cannot be formulated. Therefore, a possible solution is to use a layered approach: to specify the meaning of terms using an expressive language, and derive the axioms that must obtain in a weaker language using deductive inferences.

# 3 Reasoning over ontologies

Up till recently, biomedical ontologies have been developed with the purpose of serving as controlled vocabularies that can be employed by humans with the purpose of exploiting applications for integrating data across desperate resources as well as for knowledge discovery. However, the real power of ontologies lies in their ability to provide a shared understanding of knowledge between humans but more crucially to allow machines to exploit this knowledge. Machine reasoning can then be employed for knowledge acquisition, classification of knowledge, clustering and interpretation of the results, consistency checking, automatic knowledge discovery and hypothesis generation etc.

# 4 Formally representing the knowledge

In order, though, for ontologies to realize their potential, they must provide rich, explicit and consistent descriptions for their terms so that automated systems are able to process and understand their meaning, thereby enabling their use to infer new information. For this purpose, such descriptions are currently being created for numerous ontologies within the biomedical domain expressed, increasingly, in expressive formal languages, such as theWeb Ontology Language (OWL) [**?**]. However, in order to make use of these definitions it is imperative that their semantics are explicit and accessible to automated reasoning. More precisely, it is imperative that their definitions need to include precise descriptions of the relationships that are employed as well as ensure of the consistency of the knowledge represented.

## 4.1 relations

extend of the RO method of defining relations

## 4.2 consistency

By formalizing the relations and classes in an ontology, an automated reasoner could be employed to verify the consistency of the ontology. Moreover, methodologies can be employed that would allow not only the detection of such errors but also the automated repair of them. Such formalization and inconsistency removal enables more expressive queries over ontologies that span levels and domains of granularity.Reasoners can be employed that

## 4.3 Making ontologies interoperable on a large scale

However, due to the issues of tractability arising from the high complexity of reasoning over formal ontologies, their explicit semantics are rarely taken advantage by software systems and analysis methods. As a consequence, current ontology-based resources such as the various model organism databases, search engines, ontology repositories, ontology browsers and interfaces, make little or no use of the semantic power of the ontologies at all, which consequently diminishes their utility towards facilitating data integration and interoperability. Unless an ontologys semantics can be employed by ontology-based applications and methods, the original goal of ontologies to facilitate data integration and interoperability cannot be achieved, thereby diminishing the value of the ontology development and maintenance efforts of the past decade.

The solution seems to arise for a recently proposed EL based common layer of formal interoperability framework for all biomedical ontologies. This framework allows ontologies to be converted and disseminated in the EL subset of OWL, an OWL profile that supports tractable automated reasoning. Furthermore, it ensures that ontologies can now achieve their goal of data integration and interoperability, not only in a static sense that is applied in database annotations, but in the more important dynamic sense that is determined by how these ontologies are used.

### 4.4 Domain coverage

Biomedical ontologies today span a wide range of domains and levels of granularity ranging from the molecular level about to the organismal and environmental one. This is indeed crucial if we are to integrate biomedical data in order to combine analyze ....

## 5 Bridging domains and levels of granularity and going across species

By eliminating inconsistencies and thereby enabling the formalisation of the relation and class definitions in biomedical ontologies, users can employ the resulting ontologies for powerful queries across multiple domains as well as different species.

It paves the way for making biomedical information retrieval a knowledge-driven discipline based on formalized ontologies that make their semantics explicit and accessible to automated reasoning, thus resulting in the capability to answer novel, powerful queries that bridge multiple domains, disciplines, species and levels of granularity, thereby facilitating translational research and knowledge discovery.

## References

[1] Owl web ontology language overview. Technical report, W3C, February 2004.

[2] Philip Benfey and Thomas Mitchell-Olds. From Genotype to Phenotype: Systems Biology Meets Natural Variation. *Science*, 320(5875):495–497, April 2008.