Semantic integration of physiology phenotypes: an application to the cellular phenotype ontology

Robert Hoehndorf[1] and Midori A. Harris[2] and Heinrich Herre[3] and Gabriella Rustici[4] and Georgios V. Gkoutos[1]

[1]Department of Genetics, University of Cambridge, Downing Street, Cambridge, Cambridge CB2 3EH, UK
[2]Department of Biochemistry; University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK
[3]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Haertelstrasse 16-18, 04107 Leipzig, Germany
[4]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, Cambridge CB10 1SD, UK

ABSTRACT

The systematic observation of phenotypes has become a crucial tool of functional genomics, and several large international projects are currently underway to identify and characterize the phenotypes that are associated with genotypes in several species. To integrate phenotype descriptions within and across species, phenotype ontologies have been developed. Applying ontologies to unify phenotype descriptions in the domain of physiology has been a particular challenge due to the high complexity of the underlying domain. Here, we present the outline of a theory for an ontology of physiology-related phenotypes. We provide a formal description of process attributes and relate them to the attributes of their parts and participants. We apply our theory to create the Cellular Phenotype Ontology (CPO). The CPO is an ontology of morphological and physiological phenotypic abnormalities of cells, cell components and cellular processes. Its prime application is the unification of cellular phenotype descriptions across species by providing terms and uniform definition patterns. The CPO can further be used for the annotation of observed abnormalities in domains, such as systems microscopy, in which cellular abnormalities are observed and for which no phenotype ontology has been created. The CPO and the source code we generated to create the CPO are available on http://cell-phenotype.googlecode.com.

Keywords: cell phenotype, process, ontology, Semantic Web

1 INTRODUCTION

Phenotype studies on all scales and levels of granularity are now an invaluable tool for functional genomics research. Phenotypes of targeted mutations in animal models are now systematically recorded to reveal the role of individual genes within a biological system. These phenotype studies now play a key role in translational research and are being used to reveal candidate genes for orphan diseases and to identify chemicals that may have effects on these diseases [46].

The large volume and diversity of phenotypes within different species and across multiple scales and levels of granularity necessitates the application of flexible strategies for managing and integrating data so that it becomes amenable to automated comparative analyses. To integrate biomedical data across heterogeneous information systems, biomedical ontologies are being developed [48]. An ontology is an explicit specification of a conceptualization of a domain and can be used to make the meaning of terms in a vocabulary explicit [19, 20]. They play a crucial role in the annotation of biomedical data and the integration of model organism databases [13, 4, 17].

Ontologies increasingly rely on the use of Semantic Web technologies [7]. The Semantic Web provides a stack of protocols and languages to include explicit semantics in websites. In particular, the Web Ontology Language (OWL) [18] has been designed to express and share ontologies within the Semantic Web. OWL is a language based on description logics (a group of formal languages based on first-order predicate logic). Automated reasoners have been developed within the Semantic Web to perform complex operations on ontologies formulated in OWL. In particular, automated reasoners can verify an ontology's consistency and use deductive inference to perform powerful queries over ontologies. To benefit from automated reasoning and the rapidly increasing number of software tools that are being developed within the Semantic Web, most biomedical ontologies are now available in OWL or can be converted into an OWL-based representation [32, 29].

In the domain of phenotypes, multiple ontologies have been developed. In particular, ontologies to characterize mammalian [49], human [43], yeast [11] and worm [44] phenotypes are now available, while several more phenotype ontologies are under development. To benefit from automated reasoning, integrate phenotypes across species and reuse the content of anatomy and process ontologies, classes in phenotype ontologies were defined using the framework of the Phenotypic Attribute and Trait (PATO) ontology [14]. According to the PATO framework, a phenotype can be decomposed, using an Entity-Quality model, into an affected entity and a quality that characterizes *how* the entity is affected [14]. Such decompositions have been created for several widely used phenotype ontologies [38, 16, 15], and are being applied together with methods for reusing knowledge contained in anatomy ontologies [38, 30].

While the PATO framework is now successfully being applied to semantically integrate phenotypes across species, the diversity and complexity of phenotypes in which biological processes and functions are impaired continues to limit the interoperability

between phenotype ontologies. Major challenges for representing and integrating process-based phenotypes include establishing the link to the components of biological systems that have the capabilities to exhibit such a behaviour, and that attributes of processes are often measured *indirectly* and inferred from other attributes.

To illustrate these challenges, consider physiological processes of the heart. One of the heart's functions is *Heart beating*, i.e., a capability that is realized through processes of the type *Heart beating*. *Blood* is a participant of *Heart beating* processes. An abnormal phenotype of an organism could be that the *rate of heart beating* is increased. The intended meaning of such a description is that the number of *Heart beating* processes in a given time interval is higher than normal. Another important attribute of heart physiology is the fluid flow rate through the heart. For an abnormal phenotype such as *increased rate of fluid flow through the heart*, the intended meaning could either be that the amount of fluid that is moved through the heart within a single heart beating process is increased or that the amount of fluid that is moved through the heart within a period of time interval is increased.

Based on these examples, we can make several observations about process attributes. First, for a process like *Heart beating*, we can distinguish between single occurrences and processes in which *Heart beating* occurs multiple times. Only the latter kind of process may have a *rate of heart beating*, while *Heart beating* processes do not have such an attribute. Second, we can distinguish between abnormal fluid flow rates in *Heart beating* processes and rate of fluid flow through the heart within a given duration. Both may have entirely different underlying causes and it is therefore important to distinguish between them. Finally, we may be able to infer some phenotypes from others, thereby limiting the number of phenotypes that must be experimentally observed. For example, when the fluid flow rate in single heart beating processes is increased and the *rate of heart beating* is increased within a process P, then the rate of fluid flow through the heart will be increased for P.

Here, we present the foundations for an ontology of process phenotypes. We present the outline of a theory in which several kinds of process attributes can be distinguished so that normal and abnormal physiology of biological systems can be formally characterized. We apply this theory to cellular processes and create the Cell Phenotype Ontology (CPO). The CPO is linked to reference ontologies for qualities, biological processes, functions and cell components, and its prime application is the unification of phenotypes on the cellular level across different species as well as for annotation of cellular phenotypes in domains in which no such ontology exists.

2 MATERIALS AND METHODS

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 [19, 20]. 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 [22]. 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" [6].

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 [23], 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 ϕ such that ϕ 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 [26, 25]. Highly efficient automated reasoners are available to process OWL ontologies [47, 50, 37]. 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

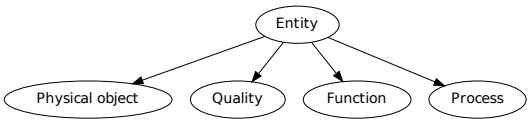

Fig. 1. Overview over basic top-level categories. Physical objects are entities that are wholly present at time points. Qualities are attributes of entities. Functions are capabilities that arise from physical objects with their qualities, and processes are temporally extended entities that may be realizations of a function.

provide the expressivity required for most biomedical ontologies [2, 27], and highly optimized OWL EL reasoners are available or under development to support reasoning over very large ontologies [2, 33].

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 [28, 24]. 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.

2.2 Processes and their participants

Most biomedical ontologies share common distinctions between different kinds of entities: physical objects, qualities, functions and processes. A physical object is an entity that is present as a whole at a time point, i.e., an entity that has no temporal parts. A quality is an attribute or feature of an entity. Physical objects together with their qualities give rise to functions, which are capabilities or potentials of physical objects. These functions can then be realized by processes, which are temporally extended entities [8]. Examples of classes that may have processes as instances include *Drinking*, *Triathlon* or *Apoptosis*. Figure 1 illustrates these basic categories of being.

Processes commonly have physical objects as participants. Physical objects are entities that are present as a whole at time points (i.e., they have no temporal parts) and may persist through time, i.e., they may undergo changes during the process [22, 21]. We can further distinguish different *roles* of participation in a process [36]. For example, a runner may participate in a *Triathlon* process as a *Runner (role)*, while another person can participate in the same process but in a different role (such as *Referee*).

2.3 Gene Ontology

The Gene Ontology provides a set of ontologies for molecular and cellular biology, originally designed to support structured annotations for genes and gene products in all species with respect to molecular function (MF), biological process (BP) and cellular component (CC). MF and BP both describe processes, but at different spatiotemporal scales; in particular, BP includes processes that unfold within cells and within tissues and organs of multicellular organisms. Gene product annotations identify participants in the processes.

Over time, GO development has increasingly emphasized a normalized approach that includes supplementing existing human-readable text description with formally specified explicit definitions for GO classes. The formalization of GO is readily apparent in its representation of biological regulation.

Regulatory processes may regulate other processes, at either the MF or BP scale, or biological qualities. GO accordingly includes three broad categories of regulation terms, regulation of molecular function, regulation of biological process, and regulation of biological quality. The first two are explicitly defined entirely with respect to other GO terms, whereas the third comprises classes in which the regulated qualities are specified by terms from PATO (see below) or anatomy ontologies.

All GO regulation terms use one of three relations, **regulates**, **negatively_regulates** and **positively_regulates**, to link regulation terms to process or quality terms. The **regulates** relations are defined in terms of qualities: a regulatory process causes a change in magnitude to some quality, which in turn has an effect on the frequency, rate or duration of some other type of process. Effects that results in increases and decreases use **positively_regulates** and **negatively_regulates** respectively [39]. The existing ontology structure would also support the addition of subclasses to distinguish, for example, regulation of the rate of a process from regulation of its duration or time of onset.

2.4 PATO and the EQ model

PATO was envisaged and designed to provide a platform for allowing the integration of quantitative and qualitative phenotype related information across different domains, levels of granularity and species [14]. PATO is an ontology of phenotype qualities that form basic entities that we can perceive and/or measure such as colors, sizes, rates etc. One of its classification axes is based on the basic type of entity to which a qualities belongs, and PATO distinguishes between qualities of physical objects and qualities of processes.

PATO allows for the description of affected entities by combining various ontologies that describe the entities that have been affected, such as the various anatomical ontologies, GO [1], the Cell Type Ontology [5], SO [10] etc with the various qualities it provides for defining how these entities were affected. PATO can be used for annotation either directly in a so called post-composed (post-coordinated) manner or for providing formal (logical) definitions (equivalence axioms) to ontologies containing a set of precomposed (pre-coordinated) phenotype terms. For instance, to describe the decrease in the length of the sexual cycle of female animals, we can combine the PATO term *decreased duration* (`PATO:0000499`) with the Gene Ontology term *estrous cycle* (`GO:0042698`), whilst if such a term existed in a pre-composed ontology (for example the `MP:0009007` term from the Mammalian Phenotype) it could be used to provide an equivalence statement between that class and the above PATO-based description.

3 RESULTS

3.1 Attributes of processes

We develop a model of process attributes that is applicable for representations of physiology and related phenotypes. In principle, we distinguish between three different kinds of process attributes: the first are process attributes that arise directly from processes and

include *duration* and *temporal location*; the second are attributes that arise from processes and their temporal parts and include *frequency* and *onset*; and the third are attributes that arise from processes and qualities of their participants, and include *flow rates*.

Attributes that can be directly linked to a process arise from processes' temporal extension. For example, a duration is an attribute that characterizes the temporal extent of a process and is similar to *Length*, *Area* and *Volume* for one-, two- and three-dimensional physical objects. A *Temporal location* positions the time interval at which the process occurs with respect to a reference coordinate system.

However, the majority of attributes that characterize processes are not based on these types of process attributes alone, but rather relate attributes of process participants with the duration of a process. In particular, a *rate* typically refers to an attribute of some entity *with respect to an attribute of another entity*, and in the context of processes, rates often refer to attributes of a process participant with respect to the duration of the process. For example, a *mass flow rate* refers to the *Mass* of a process participant with respect to the duration of the process, i.e., how much matter is moved (from one point to another) through the process. As a more complex example, a *rate of change of position* refers to the *distance* that an object is moved with respect to the duration of the process.

However, not all rates of a process depend on attributes of the process participants. In particular, a *frequency of occurrence* or *event rate* refers to the number of occurrences of a type of process during a reference process. For example, a *rate of heart beating* refers to the number of *Heart beating* processes that occur within a reference process (e.g., a process in which the heart participates with a duration of one minute). Further attributes that depend on types of processes with regard to a reference process are *distribution patterns*, i.e., how the occurrences of processes of a particular type are distributed within a reference process. For example, the heart may beat *rhythmically* or *arrhythmically* within a period of time (see Figure 2).

Related to distribution patterns are *changing qualities* of processes. For example, the rate of heart beating may change (*increase* or *decrease*) throughout the course of a reference process. A simple analysis of *increasing* (*decreasing*) rates would be that the rate of a heart beating within the first half of a process is *lower* (*higher*) than in the second half of the process. To make such an assertion, we divide a process into two temporal parts. Mathematically, this process of sub-division can be iterated until processes occur within infinitesimally small time intervals.

While some processes can be subdivided indefinitely while retaining certain kinds of attributes, others cannot. Examples of processes that can be divided include *continuous movements* or *mass flow* processes, for which all parts have a *speed* or *flow rate* attribute. On the other hand, some processes can be subdivided into stages of activity and stages of inactivity and cannot arbitrarily be divided. For example, a process of *heart beating* has periods of activity (a single heart beat) and inactivity. Consequently, not all parts of the process have a *heart rate* attribute.

We may further attribute a *frequency* or *rate* to an object instead of a process. For example, a heart that beats *now* with a frequency of 80 beats per minute, or a car that is moving at a speed of 180 kilometres per hour *at a particular point in time* (e.g., as observed with a speed camera) can be considered attributes of the objects (the heart or the car), not attributes of the processes in which the objects

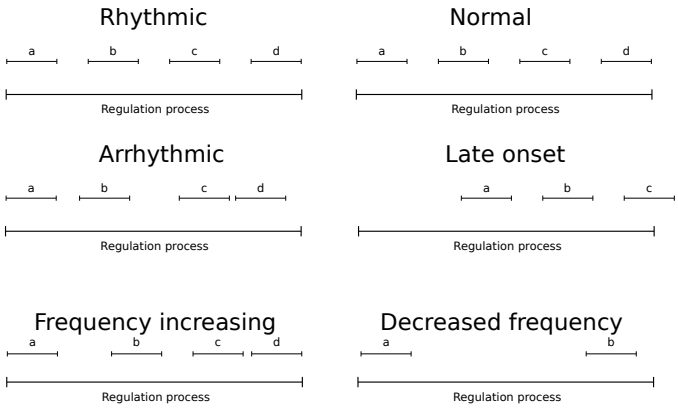

Fig. 2. Six examples of processes with non-comparative and comparative process attributes. We assume that the processes labelled a, b, c and d are all instances of the class of processes P. On the left side, three regulation (of P) processes are illustrated which exhibit non-comparative attributes. The first process has an attribute of *rhythmic* occurrence of P because the instances of P are temporally equidistantly distributed. The second example shows an *arrhythmic* occurrence of P, and the third examples shows an *increasing frequency* (of P). A regulation process with an increasing frequency (of P) is a process in which the frequency of occurrences of P is lower in the first half of the process than in the second half. The right side of the figure illustrates comparative phenotypic descriptions of processes. On the upper right, the *normal* reference is shown. The second example illustrates a *late onset* of P, i.e., the attribute that P processes begin later than *normal*. Finally, the lower right illustrates a *decreased frequency* (of P), since fewer processes of the type P occur within the reference process than *normal*.

participate. However, these are *different* kinds of attributes. Rates, when considered as attributes of objects, may be explicitly defined using rates of processes. For example, the heart beating frequency of a particular heart h at a time point t is the frequency of a reference heart beating process in which h participates. Such a reference process is necessary in order to obtain a value for a frequency even when no *heart beating* process is occurring. However, the frequency is only an attribute of the heart in virtue of such a reference process in which *heart beating* is actually occurring. This reference process can be uniquely determined for processes such as *continuous movement*, where the rate of an object at a time t is the rate of the infinitesimally small process that occurs around t. The reference process is ambiguous for processes such as *heart beating*, and the reference process must be explicitly stated.

3.2 Cell Phenotype Ontology

While our considerations about process attributes are only the beginnings of a full-fledged theory, we have derived several phenotype formalization patterns and a high-level taxonomic structure of process-based phenotypes. To evaluate our approach, we created the Cellular Phenotype Ontology (CPO) by automatically applying our patterns to the GO.

Phenotypes in the CPO are either based on structural abnormalities or abnormal physiology involving cells or cell components. Structural abnormalities in the CPO are based on GO's Cellular Component (GO-CC) hierarchy. GO-CC contains 2,918 classes for cell parts (including *Cell*) and extracellular components of cells. For each cellular component class C in the GO-CC, we

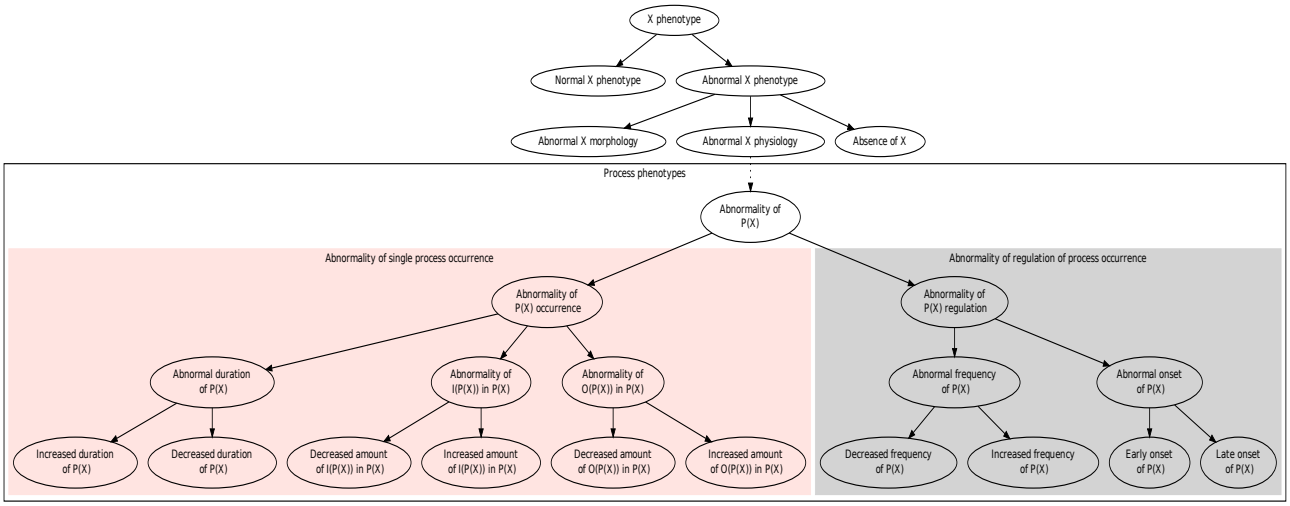

Fig. 3. Overview over the taxonomic structure of CPO. The structure is based on a cellular component class X and the cellular processes $P(X)$ in which X is involved.

create a new class labelled C *phenotype* in the CPO. For example, for the class *Mitochondrion* (GO:0005739) in the GO-CC, we create the class *Mitochondrion phenotype*.

Amongst the structural phenotype classes, we first distinguish between *normal* and *abnormal* phenotypes. An *Abnormal phenotype of* C is a phenotype of an organism that does not have a normal C as part, while a *Normal phenotype of* C represents the state in which an organism has a *normal* C as part.

Amongst the abnormal phenotypes that we include for all cell components listed in GO-CC, we distinguish *abnormal morphology* and *abnormal physiology* phenotypes. An *Abnormal morphology of* C is either the (abnormal) absence of required parts of C, the (abnormal) presence of additional parts, or abnormal qualities of C [30]. For example, an absence of caveolae (MP:0004150) would be a subclass of *Abnormal morphology of plasma membrane* in virtue of caveolae necessarily being part of the *Plasma membrane* (GO:0005886).

Abnormal physiology of a cell component refers to abnormal *functionality* of a cell component. We assume that a functionality of a cell component is (the potential for) a process in which the cell component is (causally) involved. We use the definitions of GO classes that were created based on lexical decompositions of GO class labels [39, 3, 41] to identify the processes in which cell components are involved. For example, the definition of the GO class *Mitochondrial fission* (GO:0000266) is explicitly defined as an *Organelle fission* (GO:0048285) that **results-in-the-division-of** a *Mitochondrion* (GO:0005739). Based on this definition, we make the assumption that *Mitochondrial fission* is one of the functions of a *Mitochondrion* and that an *Abnormality of mitochondrial fission* is a subclass of an *Abnormality of mitochondrion physiology*.

Amongst abnormal physiology, we distinguish between abnormalities in a *single occurrence* of a cell component's functioning and an abnormal *pattern of multiple occurrences* of a cell component's functioning. For example, abnormalities in cell division resulting in *Aneuploidy* refer to abnormalities of *cell division* processes, while an *increased rate of cell division* refers to an abnormality in the pattern of occurrence of multiple cell division processes. In the CPO we follow the GO and represent abnormalities in the pattern of occurrence of X as abnormalities of *regulation of* X processes. In particular, an *increased rate of cell division* is not an attribute of *cell division* processes, but rather an attribute of the *regulation of cell division*.

Single occurrences of processes can be abnormal in multiple ways, depending on the type of process. First, common to all processes is the quality of *duration* and consequently, each process can have an *abnormal* (increased or decreased) duration. For example, a part of an organism may participate in an *Inflammatory response* (GO:0006954) that lasts longer than normal, i.e., the organism has an *Increased duration of inflammatory response* phenotype. We define such a phenotype as a phenotype of an organism which has a part that participates in *Inflammatory response*, and this *Inflammatory response* process has an *Increased duration* (PATO:0000498).

The second common type of abnormality are abnormalities based on process participants in relation to the duration of the process. These include all kinds of *rates* such as *mass flow rate*, *energy flow rate* and *velocity* (the rate of change of position). In each of these cases, an object participates in a process and a quality (or change of quality) of that object throughout the duration of the process is considered to form a new quality. If the process has participants that are distinguished into *inputs* and *outputs*, then a recurring pattern is that the amount of inputs or outputs that participate in the process can be *increased* or *decreased*. For example, an *Increased rate of cytoplasmic streaming* can be defined as an increased amount of inputs or an increased amount of outputs of a *cytoplasmic streaming* process.

Finally, some objects may be divided into stages during which particular states of affairs obtain, and a process may be abnormal in that these states of affairs do not obtain at a particular stage. Notably, at the beginning and the end of a process, pre- and post-conditions may obtain that are abnormally changed in a process. For example, *Aneuploidy* – an abnormality during cell division at which the chromosomes do not separate properly between the two cells – may be considered the result of such an abnormality.

We implement the first two types of abnormality in the CPO. First, as a subclass of each *Abnormality of P* class, we create *Abnormal duration of P*, which in turn has *Increased duration of P* and *Decreased duration of P* as subclasses. Second, if we are able to identify *inputs* $I(P)$ or *outputs* $O(P)$ of the process P in the formal definitions of the GO, we automatically generate *Abnormality of* $I(P)$ *in* P as well as *Abnormality of* $O(P)$ *in* P. The left side of Figure 3 illustrates the schema of classes we generate for single process abnormalities.

The second type of abnormality in the CPO relate to abnormalities of *multiple* occurrences of some process X. According to the GO, *regulation* processes are processes that maintain or modify the occurrence of processes of a particular type. Following this convention, we call an abnormality of multiple occurrences of X *abnormality of the regulation of* X.

A first kind of abnormality of regulatory processes are *abnormal temporal distribution patterns* of a process. In these abnormalities, the *way* in which processes of a particular kind are temporally distributed is abnormal. The most common abnormal distribution pattern is an increased or decreased frequency, and we use PATO's *frequency* class to define *Abnormal frequency of occurrence of* X. For example, an *Abnormal frequency of occurrence of apoptosis* is defined as an abnormality of *Regulation of apoptosis* (`GO:0042981`) with respect to the *frequency* (`PATO:0000044`) of *Apoptosis* (`GO:0006915`) occurrences.

There are further types of deviation from a distribution pattern. For example, a kind of process that is normally *rhythmic* can be abnormal in that it is *arrhythmic*. A typical example of this kind of process is *Heart beating* (`GO:0060047`), in which *Cardiac muscle contraction* (`GO:0060048`) processes occur in a rhythmic pattern. In *Cardiac dysrhythmia*, however, *Cardiac muscle contraction* processes occur arrhythmically, and we consider this to be an abnormality of the regulation of *Cardiac muscle contraction*. While these abnormalities are often highly informative in clinical diagnostics and biological investigations, we usually lack the necessary information that is required to automatically determine meaningful types of abnormal distribution patterns.

A second kind of regulatory abnormalities is related to the *onset* of a process. With respect to a reference process, a particular kind of process may be *delayed* (`PATO:0000502`) or *premature* (`PATO:0000694`). For example, *Delayed apoptosis* refers to an abnormality of the *Regulation of apoptosis* in which apoptosis is induced later than normal. We use the PATO quality *onset* (`PATO:0002325`) and its children *delayed* and *premature* to define these types of regulatory abnormality. Similarly, we use PATO's *offset* (`PATO:0002324`) quality and its children to characterize regulatory abnormalities in which a process ends prematurely or too late.

Finally, a third kind of regulatory abnormality refers to abnormal rates with respect to a participant of the process that is being regulated. For example, a cytoplasmic flow rate can be increased or decreased not within a single *cytoplasmic streaming* process but rather the total cytoplasmic flow rate, as a summation over all cytoplasmic streaming processes that occur within an organism (or a particular anatomical location), is increased or decreased. While a flow rate of a single *cytoplasmic streaming* process is a quality of that process, an increased *total* cytoplasmic flow rate is a quality of the regulation of *cytoplasmic streaming*. In particular, it is possible for an organism to have a normal — or even a decreased — cytoplasmic flow rate in each individual cytoplasmic streaming process while at the same time having an increased total cytoplasmic flow rate due to a large increase in the frequency of occurrence of cytoplasmic streaming processes. Similarly, the frequency of occurrence of cytoplasmic streaming may be normal or decreased while the total cytoplasmic flow rate is increased due to an increased cytoplasmic flow rate in each individual cytoplasmic streaming process. Table 1 illustrates the dependencies between rates of individual processes, their frequency of occurrence and the total rate of these processes. We include total rates as regulatory abnormalities in the CPO since these are the attributes of processes that are often measured or observed, while the rates of individual processes are inferred following a schema such as Table 1.

3.3 Implementation

We were faced with two choices for implementing the CPO: we could either implement a pre-composed ontology in which all classes and their definitions are pre-generated according to the patterns we define, or we could develop an annotation software that enables the selection of our process phenotype patterns based on the current structure of the GO. To maximize the utility and compatibility of the CPO, and to provide stable identifiers for its concepts, we selected the first strategy and developed a software to automatically generate a pre-composed ontology from the GO.

We developed a software that utilizes the OWL API [31] in order to generate an OWL representation of the CPO. The software requires three input files: a version of the GO on which to base the generated CPO, a version of PATO that is used to define abnormal qualities, and a copy of the GO cross-product definitions [39] that is used to relate cell components to the processes in which they participate as well as identify the participants, inputs and outputs of processes.

We automatically generate a unique numerical identifier for each class in the CPO. Since the CPO is based on the GO and need to be updated with subsequent versions of the GO, we must ensure to keep identifiers stable in subsequent versions of CPO. Therefore, we use the identifiers for GO classes to generate CPO class identifiers.

In the CPO, identifiers contain two components and are of the form `CPO:XXGGGGGGG`, where `GGGGGGG` is the seven-digit identifier of the GO class on which the CPO class is based, and `XX` is a prefix that identifies the type of phenotype pattern that is applied to the GO class. For example, based on the class *Apoptosis* (`GO:0006915`), we generate the CPO classes *Abnormality of Apoptosis*, *Abnormality of single occurrence of apoptosis* and *Abnormality of regulation of apoptosis*. We use the prefixes `12`, `14` and `15` for each of the corresponding phenotype patterns, and consequently generate the class identifiers `CPO:120006915`, `CPO:140006915` and `CPO:150006915`. As long as the GO maintains its identifier for the *Apoptosis* class, the identifiers in the CPO will remain stable even when it is regenerated.

	Increased cytoplasmic flow rate	Normal cytoplasmic flow rate	Decreased cytoplasmic flow rate
Increased frequency	increased total flow rate	increased total flow rate	?
Normal frequency	increased total flow rate	normal total flow rate	decreased total flow rate
Decreased frequency	?	decreased total flow rate	decreased total flow rate

Table 1. Interdependency for the attribute *Total cytoplasmic flow rate*. A *Total cytoplasmic flow rate* is an attribute of *Regulation of cytoplasmic streaming* processes, while *Cytoplasmic flow rate* is an attribute of individual *cytoplasmic streaming* processes. Depending both on whether the cytoplasmic flow rate in individual *cytoplasmic streaming* processes is increased or decreased and whether the frequency of occurrence of *cytoplasmic streaming* is increased or decreased, the total cytoplasmic flow rate can be increased or decreased.

Neumann et al.[40]	Schmitz et al.[45]	Fuchs et al.[12]
binuclear	normal mitotic exit	actin fiver cells
cell death	prolonged mitotic exit	big cells
cell migration		bright and large cells phenotype
condensation followed by decondensation without completion of mitosis		bright nuclei
condensation without mitosis/collapse of nucleus		cells with protrusions
dynamic changes		elongated cells
failure in decondensation		elongated cells with protrusions
grape		high actin ratio cells
increased proliferation		lamellipodia + high actin ratio cells
large		lamellipodia cells
large nucleus		large cells
metaphase alignment problems/including no metaphase		large nuclei
metaphase delay/arrest		low eccentricity cells
migration (distance)		metaphase cells
migration (speed)		proliferating cells
mitotic delay/arrest		small cells
nuclei stay close together		small cells with an enrichment of mitotic cells
polylobed		
pulsating nuclei		
segregation problems/chromatin bridges/lagging chromosomes/multiple DNA masses		
small nucleus		
strange nuclear shape		

Table 2. The table summarizes cellular phenotype terms used in three recent systems microscopy studies.

We use the labels of GO classes to automatically generate class labels for phenotype classes as well as textual definitions for classes in the CPO. For example, the label of the class for increased number of occurrences of *Apoptosis* is *Increased frequency of occurrences of Apoptosis*, and its textual definition states that an increased frequency of occurrences of *Apoptosis* is a phenotype of *Regulation of apoptosis* in which the number of occurrences of *Apoptosis* within a given time period is increased in comparison to a reference process that is considered *normal*.

As of November 2011, CPO contains 125,466 classes of which 79,236 are explicitly defined. The ELK reasoner [34] is able to perform a classification of the ontology in under 10 seconds. We make the ontology and the source code that is used to generate it freely available on http://cell-phenotype.googlecode.com.

4 DISCUSSION

4.1 Applications of the CPO

The Fission Yeast Phenotype Ontology (FYPO), a new ontology developed to support annotation of phenotypes in *Schizosaccharomyces pombe*, consists of pre-composed terms describing normal or abnormal cellular phenotypes. Over 80% of FYPO definitions reference descendants of GO-BP's *Cellular process* as the entity; a further 11% reference GO-CC terms. All FYPO explicit definitions reference qualities in PATO, including *normal*, *abnormal*, and several process qualities including *increased duration* and *decreased occurrence*. FYPO will thus fit neatly under the CPO umbrella, and stands to benefit from the automated synchronization between CPO and GO, as well as the integration of cellular phenotypes across species that the CPO can provide. *Schizosaccharomyces pombe* annotations to FYPO terms will provide a rich body of highly specific, well-supported data to be integrated with data from other species.

A further domain that will greatly benefit from the CPO is *systems microscopy*, which aims to understand complex and dynamic cellular systems by combining automated fluorescence microscopy, cell microarray platforms, quantitative image analysis and data mining [35]. If we consider some of the studies, which have been published in this field in the last few years [40, 45, 12], the need for CPO becomes evident. In the three studies illustrated in Table 2, live-cell imaging assays and RNAi knockdown were used to generate phenotypic profiles that quantify the cellular response to a given siRNA thus allowing identification of hundreds of genes involved in diverse biological functions including cell division, migration and survival. In each study, several phenotypes were detected and described by the authors without the use of ontologies (see Table 2), making the integration between datasets extremely difficult. For example, it is evident that cell division phenotypes were observed in all three datasets (e.g. *Mitotic delay/arrest* and *Prolonged mitotic exit* or *Methaphase delay* and *Methaphase cells*) but the overlap between such phenotypes is unclear.

Data integration is also complicated by the lack of standardization at the level of data production and processing; all these issues are currently being address by the different groups involved in the Systems Microscopy Network of Excellence (http://www.systemsmicroscopy.eu/) and the first step towards data integration can be achieved by further developing CPO.

This ontology will be used to integrate phenotypes' definitions across existing datasets and will then become an integrated part of the data processing pipeline and used to annotate the data as it gets generated [9].

4.2 Future research

We implemented the CPO using a pattern-based approach to formulating phenotypes involving processes. The patterns we identify are based on pre-existing ontologies, in particular the PATO ontology and the classification of cellular processes as well as cellular components in the GO. The result of our method is a large ontology in which classes for phenotypes are *pre-composed*: they are named and defined within an OWL ontology. However, the large size of the resulting ontology may impair its utility for data annotation and integration, and software tools may not always support such very large ontologies. The alternative to pre-composing all possible phenotype classes using the patterns we describe is to dynamically generate appropriately defined classes at the time at which they are being used. To achieve this goal, software must be developed to support ontology users in applying these patterns and generate the appropriate class description when required.

A further important task is to develop the theory we outlined and applied for the CPO. In particular, a precise formal characterization of this theory in terms of axioms will further improve the clarity of phenotypic descriptions of processes and enable its integration in well-developed formal ontologies of processes [22, 42].

5 ACKNOWLEDGEMENTS

Funding for RH is provided by the European Commission's 7th Framework Programme, RICORDO project, grant number 248502. MAH is supported by Wellcome Trust, grant WT090548MA. HH is supported by the Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig. Funding for GR is provided by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 258068, EU-FP7-Systems Microscopy NoE.

REFERENCES

[1]M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.

[2]F. Baader, C. Lutz, and B. Suntisrivaraporn. CEL – a polynomial-time reasoner for life science ontologies. In U. Furbach and N. Shankar, editors, *Proceedings of the 3rd International Joint Conference on Automated Reasoning (IJCAR'06)*, volume 4130 of *Lecture Notes in Artificial Intelligence*, pages 287–291. Springer-Verlag, 2006.

[3]M. Bada and L. Hunter. Enrichment of obo ontologies. *Journal of Biomedical Informatics*, 40(3):300–315, June 2007.

[4]Michael Bada, Robert Stevens, Carole Goble, Yolanda Gil, Michael Ashburner, Judith A. Blake, Michael J. Cherry, Midori Harris, and Suzanna Lewis. A short study on the success of the gene ontology. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1(2):235–240, February 2004.

[5]Jonathan Bard, Seung Y. Rhee, and Michael Ashburner. An ontology for cell types. *Genome Biology*, 6(2), 2005.

[6]J. Barwise. *The Situation in Logic*. CSLI, Stanford, CA, 1989.

[7]T. Berners-Lee, J. Hendler, O. Lassila, et al. The Semantic Web. *Scientific American*, 284(5):28–37, 2001.

[8]Patryk Burek. *Ontology of Functions*. PhD thesis, University of Leipzig, Institute of Informatics (IfI), 2006.

[9]Christian Conrad, Annelie Wünsche, Tze Heng H. Tan, Jutta Bulkescher, Frank Sieckmann, Fatima Verissimo, Arthur Edelstein, Thomas Walter, Urban Liebel, Rainer Pepperkok, and Jan Ellenberg. Micropilot: automation of fluorescence microscopy-based imaging for systems biology. *Nature methods*, 8(3):246–249, March 2011.

[10]K. Eilbeck, S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner. The sequence ontology: A tool for the unification of genome annotations. *Genome Biology*, 6(R55), 2005.

[11]Stacia R. Engel, Rama Balakrishnan, Gail Binkley, Karen R. Christie, Maria C. Costanzo, Selina S. Dwight, Dianna G. Fisk, Jodi E. Hirschman, Benjamin C. Hitz, Eurie L. Hong, Cynthia J. Krieger, Michael S. Livstone, Stuart R. Miyasato, Robert Nash, Rose Oughtred, Julie Park, Marek S. Skrzypek, Shuai Weng, Edith D. Wong, Kara Dolinski, David Botstein, and J. Michael Cherry. Saccharomyces Genome Database provides mutant phenotype data. *Nucleic Acids Research*, 38(suppl 1):D433–D436, 2010.

[12]Florian Fuchs, Gregoire Pau, Dominique Kranz, Oleg Sklyar, Christoph Budjan, Sandra Steinbrink, Thomas Horn, Angelika Pedal, Wolfgang Huber, and Michael Boutros. Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. *Molecular Systems Biology*, 6, June 2010.

[13]Gene Ontology Consortium. The gene ontology in 2010: extensions and refinements. *Nucleic acids research*, 38(Database issue):D331–335, January 2010.

[14]Georgios V. Gkoutos, Eain C. Green, Ann-Marie M. Mallon, John M. Hancock, and Duncan Davidson. Using ontologies to describe mouse phenotypes. *Genome biology*, 6(1), 2005.

[15]Georgios V. Gkoutos and Robert Hoehndorf. Ontology-based cross-species integration and analysis of saccharomyces cerevisiae phenotypes. In *Proceedings of the 3rd Workshop for Ontologies in Biomedicine and Life sciences (OBML)*, October 2011.

[16]Georgios V. Gkoutos, Chris Mungall, Sandra Dolken, Michael Ashburner, Suzanna Lewis, John Hancock, Paul Schofield, Sebastian Kohler, and Peter N. Robinson. Entity/quality-based logical definitions for the human skeletal phenome using PATO. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, 1:7069–7072, 2009.

[17]C. Goble and R. Stevens. State of the nation in data integration for bioinformatics. *Journal of Biomedical Informatics*, 41(5):687–693, 10 2008.

[18]B. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patelschneider, and U. Sattler. OWL 2: The next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):309–322, November 2008.

[19]Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5-6), 1995.

[20]Nicola Guarino. Formal ontology and information systems. In Nicola Guarino, editor, *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems*, pages 3–15. IOS Press, 1998.

[21]Heinrich Herre. General Formal Ontology (GFO): A foundational ontology for conceptual modelling. In Roberto Poli, Michael Healy, and Achilles Kameas, editors, *Theory and Applications of Ontology: Computer Applications*, chapter 14, pages 297–345. Springer, Heidelberg, 2010.

[22]Heinrich Herre, Barbara Heller, Patryk Burek, Robert Hoehndorf, Frank Loebe, and Hannes Michalek. General Formal Ontology (GFO) – A foundational ontology integrating objects and processes [Version 1.0]. Onto-Med Report 8, IMISE, University of Leipzig, Leipzig, Germany, 2006.

[23]David Hilbert. Axiomatisches Denken. *Mathematische Annalen*, 78:405–415, 1918.

[24]Robert Hoehndorf, Colin Batchelor, Thomas Bittner, Michel Dumontier, Karen Eilbeck, Rob Knight, Chris J. Mungall, Jane S. Richardson, Jesse Stombaugh, Eric Westhof, Craig L. Zirbel, and Neocles B. Leontis. The RNA ontology (RNAO): An ontology for integrating RNA sequence and structure data. *Applied Ontology*, 6(1):53–89, April 2011.

[25]Robert Hoehndorf, Michel Dumontier, John H. Gennari, Sarala Wimalaratne, Bernard de Bono, Daniel L. Cook, and Georgios V. Gkoutos. Integrating systems biology models and biomedical ontologies. *BMC Systems Biology*, 5(1):124+, August 2011.

[26]Robert Hoehndorf, Michel Dumontier, Anika Oellrich, Dietrich Rebholz-Schuhmann, Paul N. Schofield, and Georgios V. Gkoutos. Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *PLOS ONE*, 6(7):e22006, July 2011.

[27]Robert Hoehndorf, Michel Dumontier, Anika Oellrich, Sarala Wimalaratne, Dietrich Rebholz-Schuhmann, Paul Schofield, and Georgios V. Gkoutos. A common layer of interoperability for biomedical ontologies based on OWL EL. *Bioinformatics*, 27(7):1001–1008, April 2011.

[28]Robert Hoehndorf, Janet Kelso, and Heinrich Herre. The ontology of biological sequences. *BMC bioinformatics*, 10(1):377+, November 2009.

[29]Robert Hoehndorf, Anika Oellrich, Michel Dumontier, Janet Kelso, Dietrich Rebholz-Schuhmann, and Heinrich Herre. Relations as patterns: Bridging the gap between OBO and OWL. *BMC Bioinformatics*, 11(1):441+, 2010.

[30]Robert Hoehndorf, Anika Oellrich, and Dietrich Rebholz-Schuhmann. Interoperability between phenotype and anatomy ontologies. *Bioinformatics*, 26(24):3112 – 3118, 10 2010.

[31]Matthew Horridge, Sean Bechhofer, and Olaf Noppens. Igniting the OWL 1.1 touch paper: The OWL API. In *Proceedings of OWLED 2007: Third International Workshop on OWL Experiences and Directions*, 2007.

[32]Ian Horrocks. OBO flat file format syntax and semantics and mapping to OWL Web Ontology Language. Technical report, University of Manchester, March 2007. http://www.cs.man.ac.uk/~horrocks/obo/.

[33]Yevgeny Kazakov. Consequence-driven reasoning for Horn SHIQ ontologies. In *Proceedings of the 21st International Conference on Artificial Intelligence (IJCAI 2009)*, pages 2040–2045, July 11-17 2009.

[34]Yevgeny Kazakov, Markus Krötzsch, and František Simančík. Unchain my $\mathcal{EL}$ reasoner. In *Proceedings of the 23rd International Workshop on Description Logics (DL'10)*, CEUR Workshop Proceedings. CEUR-WS.org, 2011.

[35]John G Lock and Staffan Strömblad. Systems microscopy: an emerging strategy for the life sciences. *Experimental Cell Research*, 316(8):1438–1444, 2010.

[36]Frank Loebe. Abstract vs. social roles – towards a general theoretical account of roles. *Applied Ontology*, 2(2):127–158, 2007.

[37]Boris Motik, Rob Shearer, and Ian Horrocks. Hypertableau Reasoning for Description Logics. *Journal of Artificial Intelligence Research*, 36:165–228, 2009.

[38]Christopher Mungall, Georgios Gkoutos, Cynthia Smith, Melissa Haendel, Suzanna Lewis, and Michael Ashburner. Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2+, 2010.

[39]Christopher J. Mungall, Michael Bada, Tanya Z. Berardini, Jennifer Deegan, Amelia Ireland, Midori A. Harris, David P. Hill, and Jane Lomax. Cross-product extensions of the gene ontology. *Journal of biomedical informatics*, February 2010. in press.

[40]Beate Neumann, Thomas Walter, Jean-Karim Hériché, Jutta Bulkescher, Holger Erfle, Christian Conrad, Phill Rogers, Ina Poser, Michael Held, Urban Liebel, and et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, 464(7289):721–727, 2010.

[41]P. V. Ogren, K. B. Cohen, G. K. Acquaah-Mensah, J. Eberlein, and L. Hunter. The compositional structure of gene ontology terms. *Pac Symp Biocomput*, pages 214–225, 2004.

[42]Atalay Özgövde and Michael Grüninger. Foundational process relations in bio-ontologies. In *Proceeding of the 2010*

conference on Formal Ontology in Information Systems, pages 243–256, Amsterdam, The Netherlands, The Netherlands, 2010. IOS Press.

[43]P. N. Robinson, S. Koehler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *American journal of human genetics*, 83(5):610–615, 2008.

[44]Gary Schindelman, Jolene Fernandes, Carol Bastiani, Karen Yook, and Paul Sternberg. Worm phenotype ontology: integrating phenotype data within and beyond the c. elegans community. *BMC Bioinformatics*, 12(1):32, 2011.

[45]Michael H. A. Schmitz, Michael Held, Veerle Janssens, James R. A. Hutchins, Otto Hudecz, Elitsa Ivanova, Jozef Goris, Laura Trinkle-Mulcahy, Angus I. Lamond, Ina Poser, Anthony A. Hyman, Karl Mechtler, Jan-Michael Peters, and Daniel W. Gerlich. Live-cell imaging rnai screen identifies pp2a–b55α and importin-β1 as key mitotic exit regulators in human cells. *Nature Cell Biology*, 12:886–893, 2010.

[46]Paul N. Schofield, John P. Sundberg, Robert Hoehndorf, and Georgios V. Gkoutos. New approaches to the representation and analysis of phenotype knowledge in human diseases and their animal models. *Briefings in Functional Genomics*, 10(5):258–265, 2011.

[47]Evren Sirin and Bijan Parsia. Pellet: An OWL DL reasoner. In Volker Haarslev and Ralf Möller, editors, *Proceedings of the 2004 International Workshop on Description Logics, DL2004, Whistler, British Columbia, Canada, Jun 6-8*, volume 104 of *CEUR Workshop Proceedings*, Aachen, Germany, 2004. CEUR-WS.org.

[48]Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, Neocles Leontis, Philippe R. Serra, Alan Ruttenberg, Susanna A. Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech*, 25(11):1251–1255, 2007.

[49]Cynthia L. Smith, Carroll-Ann W. Goldsmith, and Janan T. Eppig. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6(1):R7, 2004.

[50]D. Tsarkov and I. Horrocks. FaCT++ description logic reasoner: System description. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4130 LNAI:292–297, 2006.