The ontology of biological taxa
Stefan Schulz1,*, Holger Stenzhorn1,2 and Martin Boeker1

1Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Stefan-Meier-Str. 26, 79104 Freiburg, Germany and 2Digital Enterprise Research Institute (DERI), National University of Ireland, Galway, University Road, Galway, Ireland

ABSTRACT
Motivation: The classification of biological entities in terms of species and taxa is an important endeavor in biology. Although a large amount of statements encoded in current biomedical ontologies is taxon-dependent there is no obvious or standard way for introducing taxon information into an integrative ontology architecture, supposedly because of ongoing controversies about the ontological nature of species and taxa.

Results: In this article, we discuss different approaches on how to represent biological taxa using existing standards for biomedical ontologies such as the description logic OWL DL and the Open Biomedical Ontologies Relation Ontology. We demonstrate how hidden ambiguities of the species concept can be dealt with and existing controversies can be overcome. A novel approach is to envisage taxon information as qualities that inhere in biological organisms, organism parts and populations.

Availability: The presented methodology has been implemented in the domain top-level ontology BioTop, openly accessible at http://puiri.org/biotop. BioTop may help to improve the logical and ontological rigor of biomedical ontologies and further provides a clear architectural principle to deal with biological taxa information.

Contact: stschulz@uni-freiburg.de

1 INTRODUCTION
The classification of biological entities according to their morphological, genetic, evolutionary and functional characteristics is a fundamental organizing principle since Carolus Linnaeus established conventions for naming living organisms (Ereshefsky, 2001). One century later, the distinction of species received its theoretical underpinning with Charles Darwin’s theory of evolution (1859) and was finally demystified by the spectacular advances of molecular biology in the late 20th century. Although these changes have drastically challenged the basic assumptions of Linnaeus’ biological theory and have given rise to an ongoing debate about the concept of biological species and taxa (Hey, 2006), his main organizing principle remains the same.

All biology is, in some way, related to the concept of biological taxa. Taxa are hierarchically structured labels or categories used for biological classification, such as species, family, class, etc. All organisms, populations, tissues, cells, cell components and biological macromolecules that are under scrutiny of experimental or descriptive biologists are related to some hierarchy of taxa and most biological discoveries have their scope related to one species or taxon. Table 1 gives an exemplary overview of the hierarchical order of taxa. The basic taxon is the species. Several species are grouped together by a genus. Several genera constitute a family, several families an order, several orders a class and then several classes a phylum or division. Finally, the top-most level, the kingdom distinguishes between animals and plants. Similar to the several criteria that are discussed to delineate the concept of species, no clear principles exist that govern the division of superordinate taxa. For instance, orders can be further split into superfamilies and superfamilies. Even more, the number of taxonomic divisions is variable, and there are also divisions without rank name.

The importance of species and biological taxa is evidenced by many sources. Biological taxa constitute 3497 out 24 766 descriptors of MeSH1, the indexing vocabulary of Medline. In the Open Biomedical Ontologies (OBO) collection2 (Smith et al., 2007), 30 out of 66 ontologies are taxon specific, with taxa ranging from species such as Homo Sapiens or Caenorhabditis elegans, genera such as Plasmodium over families such as Poacae to classes such as Mammalae. Due to the sheer number of taxa there is no universal authoritative source, but every important subfield within biology has been independently maintained by curators, so-called systematists, and for a long time the field of biological systematics has been considered an important research discipline. A converging effort in unifying taxon information for whole biology is the Catalogue of Life3 targeted for complete coverage of all 1.75 million known species by 2011. In the mentioned OBO collection, nearly half a million taxon entries of medical interest are available in computer-processable form via the rapidly growing NCBI Taxonomy (Wheeler et al., 2008).

To sum up, biological taxa constitute an overarching and systematic ordering principle that is relevant in practically all biological subject areas.

In this article, we will show how the realm of biological systematics can be embedded into an ontological framework. It is structured as follows: We start with a summary introduction of domain ontologies in general, as well as in the context of the biology, addressing the OBO ontologies and the BioTop biomedical top-domain ontology. Then we provide a formal account of different aspects of the conceptualization of biological taxa and demonstrate how this is implemented in BioTop. Finally, we briefly describe our tentative implementation supporting our claim that an overarching ontological framework for biology must have a conclusive and practical account of biological taxa.

1Medical Subject Headings, http://www.nlm.nih.gov/mesh
2Open Biomedical Ontologies, http://www.obofoundry.org
3Catalogue of Life, http://www.catalogueoflife.org

*To whom correspondence should be addressed.
Table 1. Biological taxa with examples

| Taxon (rank) | Asian elephant | Chimpanzee | Drosophila |
|-------------|----------------|------------|------------|
| Species     | *Elephas maximus* | *Pan troglodytes* | *Drosophila melanogaster* |
| Genus       | *Elephas* | *Pan* | *Drosophila* |
| Subfamily   | Elephantidae | Hominidae | Drosophilidae |
| Family      | Elephantidae | Hominidae | Drosophilidae |
| Superfamily | Proboscidea | Primates | Digitera |
| Class       | Mammalia | Mammalia | Insecta |
| Subphylum   | Vertebrata | Vertebrata | |
| Phylum      | Chordata | Chordata | Arthropoda |
| Kingdom     | Animalia | Animalia | Animalia |

2 BIOMEDICAL ONTOLOGIES

2.1 The foundations of biomedical ontology

It is mainly the information explosion in biology and the necessity to process huge amounts of research data that have stimulated the proliferation of biomedical ontologies. Rubin et al. (2008) give an overview of the broad range of biomedical information services that can be supported by domain ontologies, with the Gene Ontology Ashburner et al. (2000) and the OBO collection as the most prominent examples. Whereas this tenet used to be addressed in the tenet of terminology from the one of ontology subscribing to the following general rejection of this term. In this article, we detach the concept of terminology from the one of ontology subscribing to the following definitions:

According to ISO (2000), a terminology is defined as a set of terms representing the system of concepts of a particular field. Terminologies relate the senses or meanings of linguistic entities. In contrast, according to Quine (1948), Ontology (in singular and upper case) is the study of what there is. In our understanding, ontologies (plural and lowercase) are formal theories that attempt to give precise formulations of the types of entities in reality, of their properties, and of the relations between them (Guarino, 1998). In contradistinction to terminology, formal ontologies strive for describing (as much as possible) what the consensus in a given scientific domain is, independently of human language. Their constituent nodes are referred to as types, kinds or universals. As they are well suited to hierarchically order and classify particular entities (e.g. a given piece of tissue, a cell under a microscope, an amount of biological substance, an animal, a particular population of bacteria, etc.), they are also referred to as classes, a paradigm we will use in the following, in accordance with the more recent language use in current biomedical ontology engineering and research. 3

Although the question whether certain entities really exist are subject to major philosophical disputes, we contend that at any given stage in the development of science, there is a consensus core of scientific understanding of reality, and in our view, it is this which should serve as starting point for developing science-based ontologies. Examples of statements belonging to this consensus core are that: primates are vertebrates, cells contain cytoplasm, aspirin tablets contain a derivative of salicylic acid, ADP is phosphorylated in mitochondria or that certain biochemical compounds have a clearly delineated composition.

2.2 Top-level ontologies

It is widely recognized that the construction of formal ontologies should obey principled criteria. To this end, several top-level ontologies have been devised, such as DOLCE (Gangemi et al., 2002), BFO (Smith et al., 2005), or GOL (Heller and Herre, 2004). These ontologies mainly coincide in their fundamental division between continuants (also called endurants, e.g. material objects) and occurrents (also called perdurants, e.g. events, processes). Orthogonal to this distinction, there is also a coincidence in clearly separating concrete entities or particulars (e.g. ‘the chimpanzee named Washoe’, ‘the elephant named Clyde’, or ‘the 3rd author of this article’) from the classes they instantiate (e.g. *Chimpanzee*, *Asian Elephant*, *Human*). To this end, we introduce the irreflexive, anti-transitive and asymmetric instantiation relation instance_of which relates particulars to classes. In addition, we need a formal relation for subsumption between classes. Here we follow the OBO standard and introduce, for this purpose, the taxonomic subsumption relation is_a by means of instance_of as just as proposed by Smith et al. (2005):

\[ \text{is}_a (A, B) \equiv_{def} \forall x : (\text{instance}_o f(x, A) \rightarrow \text{instance}_o f(x, B)) \]

In the following discussion, we are proposing several possible alternative solutions for an ontological account of species.

2.3 Domain top-level ontologies

Whereas top-level ontologies contain only a restricted set of highly general classes, such as the aforementioned Continuant, Occurrent, Function or Object, which are not tied to any particular domain of interest, a domain top-level ontology contains all the classes that are essentially needed to describe a certain domain, like Organism, Tissue, Cell and also Species in the case of biology. Those more specific classes are in turn a specialization of the top-level classes as expressed in the formula is_a (Cell, Object).

2.4 BioTop—a domain top-level ontology

Recently, two separate implementations to encode the top-level of the biomedical domain into ontologies have been created, namely, BioTop (Stenzhorn et al., 2007) and the Simple Top Bio (Rector et al., 2007). At the moment, efforts set forth by the authors are ongoing to converge these two implementations. The goal of BioTop is to provide classes and classificatory criteria to categorize the foundational kinds of biology, without any restriction to granularity, species, developmental stages or states.
Although there are rarely any two individuals with exactly identical genetic and phenotypic modifications can be assumed to lie on a clear, well-drawn, as there are no organisms existing in the middle. species but they are clearly discrete. For instance, the boundary between the view, the distinguishing features do not lie on a continuous scale with clear criteria of species identity. In contrast to the diachronic viewpoint, this manifests itself as groups of organisms that can interbreed and produce fertile offspring with elephants from Cambodia, for instance. Abbreviating the ability of producing fertile offspring by ψ, according to the biological species concept, the pertinence of biological organisms to the same species is expressed by the predicate \( \sigma(\omega_1, \omega_2) = \text{def}(\exists t: \psi(\omega_1, \omega_2, t) \vee (\exists o_1, t_1, t_2: (\psi(o_1, o_2, t_1) \vee (o_2, o_2, t_2))) \)

The shortcomings of Mayr’s definition are well known (Greene and Depew, 2004, ch. 10): first, it only allows the comparison of organisms living at the same time. Second, the definition depends on the dispositional criterion ψ, the verification of which remains speculative in many cases. Third, the definition fails with infertile individuals, as well as with species in extinction of which only female or male individuals remain. Fourth, it fails in the numerous cases of asexual reproduction such as bacteria. It is therefore neither easily applicable, nor generally valid, in spite of its theoretical soundness (Hull, 1997). So it is not surprising that other species concepts compete with Mayr’s one. The 22 different conceptualizations of species identified and discussed by Mayden (1997) bear witness on the intensive discussions and disagreements among theoretical biologists and philosophers.

For our practical purpose of biomedical ontologies the formalization of species or—more generally—of biological taxa that we propose, is intended to be neutral to the different and conflicting species conceptualizations. It departs from the principle that biological taxa are something that regardless of its existence in nature or its (fiat) attribution by biologists has a highly ranked importance in biology and therefore requires to be accounted for in biomedical ontologies.⁶

In the following, we will analyze the ontological status of biological taxa and propose and critically assess alternative solutions.

3 CONCURRENT ACCOUNTS OF BIOLOGICAL TAXA

3.1 Biological taxa as meta-properties

The above restriction to a two-leveled ontological framework (i.e. dividing the world exhaustively into particulars and universals) has often been challenged. (Gangemi et al., 2001) contend that there is a fundamental difference between instances in an ontology on the one hand and domain entities (particulars, cf. Section 2.1) on the other hand. The authors are aware of the fact that this may challenge some of the philosophical foundations underlying Basic Formal Ontology (BFO).

⁶Web Ontology Language (OWL): http://www.w3.org/2004/02/owl-faqs
the other hand. They argue that we can extend a Theory A (which follows the two-level assumption) by a meta-Theory B. Whereas Theory A describes domain entities (particulars) that instantiate universals (classes), B takes A’s universals as instances of so-called meta-properties. Indexing the instantiation relation by theory level (using subscripts in the formulae) we may state in Theory A that

\[ \text{instance of}_A(x, \gamma) \]

and then place this in the context of Theory B with

\[ \text{instance of}_B(y, \gamma) \]

To give a concrete example:

\[ \text{instance of}_A(\text{'Clyde'}, \text{Elephas maximus}) \]
\[ \text{instance of}_B(\text{'Elephas maximus'}, \text{Genus Elephas}) \]

Due to the algebraic property of antitransitivity (as claimed by (Gangemi et al., 2001), we can then coherently reject the hypothesis that our elephant ‘Clyde’ is an instance of Species. There are several arguments against this solution. Let us consider the second-level predications \[ \text{instance of}_B(\text{Elephas maximus}, \text{Genus Elephas}) \] on the one hand and \[ \text{instance of}_A(\text{'Clyde'}, \text{Genus Elephas}) \] on the other hand. Whereas the first one asserts that the class \text{Elephas maximus} is an instance of a Species, the second one states that the species \text{Class Elephas maximus} as a member of the genus \text{Elephas}. In the same right as we have stated

\[ \text{instance of}_B(\text{Elephas maximus}, \text{Species}) \]

we could then assert in a third-level predication \[ \text{instance of}_C(\text{Clyde}) \]

‘Clyde’ would then be a second-level instance of Species and a second-level instance of Genus Elephas, as well as, in virtue of the latter, a third-level instance of Genus.

Given \[ \text{instance of}_C(\text{Species}, \text{Genus}) \] and \[ \text{instance of}_C(\text{Genus}, \text{Taxon}) \], ‘Clyde’ would finally act simultaneously as third and fourth-level instance of Subtaxon. Together with the argument that ‘Clyde’ might also directly instantiate Genus \text{Elephas} and the fact that some taxonomic levels (such as subfamilies) are sometimes skipped, it is very obvious that this solution leads to an obscure and inconsistent picture.

Another shortcoming of this approach lies in the fact that it lacks a transitive hierarchical relation between taxa of different levels that would be able to express in simple terms (e.g. that all Indian elephants are vertebrates). From a computational point of view, there is also an important performance argument. For example, efficient reasoning algorithms which have been developed for description logics (Baader et al., 2003) and are coherent with the Semantic Web standard OWL DL do not provide support for reasoning capabilities about instances of instances.

3.2 Biological taxa as hierarchies of classes

We finally end up with all taxa in a specialization hierarchy, having individual organisms as instances. This neither captures the nature of a biological organism, nor the intended meaning of \text{Species} or \text{Genus}, since neither Clyde nor Washoe or any other individual animal is an instance of the class \text{Species}.

Nevertheless, we could consistently do this excluding the terms species, genus, etc. This would reduce the instances of taxa (\text{Elephant, Elephantidae, Vertebrates}) to classes of organisms and we would no longer be able to account for the meaning of terms like \text{Genus or Species} in a description logic-based framework. However, the resulting assertions such as ‘Clyde is an instance of Mammalia’ (on par with ‘Clyde is an instance of Elephant’) would collide with the plural meaning of the taxon terms.

3.3 Biological taxa as populations

Several authors have argued in favor of the inclusion of collectives into an ontological framework (Bittner et al., 2004; Rector et al., 2006; Schulz et al., 2006). BioTop has embraced these aspects by introducing the relation has\textunderscore granular\textunderscore part, an irreflexive and intransitive subrelation of the OBO Relation Ontology relation has\textunderscore part (Schulz et al., 2006b).

This allows us to relate a collective entity to each of its constituent elements, without, however, resorting to set theory. For instance, has\textunderscore granular\textunderscore part (‘PopulationofThaiElephants’, ‘Clyde’) asserts that there is a collective entity ‘Population of Thai Elephants’ that is constituted by ‘granular parts’ like our elephant ‘Clyde’ and a number of other individuals similar to ‘Clyde’.

\[ \text{Is}_\alpha(\text{Simia troglodytes, Species}) \]
\[ \text{Is}_\alpha(\text{Genus Elephas, Genus}) \]
\[ \text{Is}_\alpha(\text{Genus Pan, Genus}), \]

just as

\[ \text{Is}_\alpha(\text{Elephas maximus, Genus Elephas}) \]
\[ \text{Is}_\alpha(\text{Simia troglodytes, Genus Pan}) \]

The weakness of this solution, however, immediately derives from the above definition of the \text{Is}_\alpha relation. So given that

\[ \text{instance of}_A(\text{'Clyde'}, \text{Elephas maximus}) \]
\[ \text{instance of}_B(\text{'Washoe'}, \text{Simia troglodytes}) \]

we can infer that

\[ \text{instance of}_C(\text{'Clyde'}, \text{Genus Elephas}) \]
\[ \text{instance of}_C(\text{'Washoe'}, \text{Genus Pan}) \]

as well as that

\[ \text{instance of}_C(\text{'Clyde'}, \text{Species}) \]
\[ \text{instance of}_C(\text{'Washoe'}, \text{Species}) \]
\[ \text{instance of}_C(\text{'Clyde'}, \text{Genus}) \]
\[ \text{instance of}_C(\text{'Washoe'}, \text{Genus}) \]

We finally end up with all taxa in a specialization hierarchy, having individual organisms as instances. This neither captures the nature of a biological organism, nor the intended meaning of \text{Species} or \text{Genus}, since neither Clyde nor Washoe or any other individual animal is an instance of the class \text{Species}.

Nevertheless, we could consistently do this excluding the terms species, genus, etc. This would reduce the instances of taxa (\text{Elephant, Elephantidae, Vertebrates}) to classes of organisms and we would no longer be able to account for the meaning of terms like \text{Genus or Species} in a description logic-based framework. However, the resulting assertions such as ‘Clyde is an instance of Mammalia’ (on par with ‘Clyde is an instance of Elephant’) would collide with the plural meaning of the taxon terms.

3.3 Biological taxa as populations

Several authors have argued in favor of the inclusion of collectives into an ontological framework (Bittner et al., 2004; Rector et al., 2006; Schulz et al., 2006). BioTop has embraced these aspects by introducing the relation has\textunderscore granular\textunderscore part, an irreflexive and intransitive subrelation of the OBO Relation Ontology relation has\textunderscore part (Schulz et al., 2006b).

This allows us to relate a collective entity to each of its constituent elements, without, however, resorting to set theory. For instance, has\textunderscore granular\textunderscore part (‘PopulationofThaiElephants’, ‘Clyde’) asserts that there is a collective entity ‘Population of Thai Elephants’ that is constituted by ‘granular parts’ like our elephant ‘Clyde’ and a number of other individuals similar to ‘Clyde’. It permits to define...
ElephantPopulation
The union of all possible instances of ElephantPopulation and an instance of the universal collective 'Total ElephantPopulation' would then be the maximal population collectives in terms of granular parts such as

\[ \forall x \text{ instance_of} (x, \text{ElephantPopulation}) \Rightarrow \exists y_1, y_2, \ldots, y_n \text{ instance_of} (y_1, y_2, \ldots, y_n, \text{Elephant}) \cap \\
\text{has_granular_part} (x, y_1, y_2, \ldots, y_n) \land \\
\text{inheres_in} (z, \text{Elephant}) \Rightarrow \\
\text{has_granular_part} (x, z) \]

Note that 'Population of Thai Elephants' is a particular collective and an instance of the universal collective ElephantPopulation. The union of all possible instances of ElephantPopulation, namely, 'Total ElephantPopulation' would then be the maximal population of elephants every individual elephant is a granular part of.

\[ \forall x \text{ instance_of} (x, \text{Elephant}) \Rightarrow \\
\text{has_granular_part} (\text{Total ElephantPopulation}, x) \]

Yet, 'TotalElephantPopulation' is a particular entity. Our proposal here is to consider it as an instance of Species. In the same way, we could introduce other populations in different degrees of abstraction such as 'TotalVertebratePopulation' which would then be an instance of Phylum.

It may be practical for many purposes to equate biological taxa with biological populations although the meaning of Elephantidae or Vertebratia, in practice, goes further. Especially in molecular biology, species information is not only attributed to whole organisms, but also to organism parts, their constituting cells and derived cell lines. As an example, individual cells from the HELA cell line are considered human cells, but their existence is not dependent on any human population. The interpretation of biological taxa as populations is therefore not adequate for such cases. We can use the OBO relation derives_from in order to express that a HELA cell is a human cell:

\[ \forall x \text{ instance_of} (x, \text{HELA Cell}) \Rightarrow \\
\text{derives_from} (x, y) \land \\
\text{has_granular_part} (\text{Total Human Population}, y) \]

### 3.4 Biological taxa as qualities

Most top-level ontologies coincide in granting qualities a prominent status. For instance, BFO describes the class Quality as ‘A dependent continuant that is exhibited if it inheres in an entity or categorical property. Examples: the color of a tomato, the ambient temperature of air, the circumference shape of a nose, the mass of a piece of gold, the weight of a chimpanzee’. DOLCE introduces qualities as ‘...the basic entities we can perceive or measure: shapes, colors, sizes, sounds, smelles, as well as weights, lengths, electric charges’ (Masolo, 2003) and also makes reference to the relationship of inherence. The position of the class Quality in BFO makes clear that qualities are dependent entities, i.e., they can only exist in dependence on the entities they inheren in.

Our proposal here is to interpret the relation of a biological object to a given taxon as the ascription of a quality. For example, the quality of belonging to the species Homo sapiens is a quality that inheres in any human organism, tissue or cell. The quality of belonging to the phylum Chordata is a quality that inheres in any biological object that is part of or derived from an organism the species of which belongs to the phylum Chordata.

Figure 1 depicts a segment of our proposed subclass hierarchy of taxon qualities. The hierarchy exhibits two organizational principles: generalization versus specialization on one side, and the relevance to an organizational level on the other. Every instance of a material biological object has one inherent taxon quality.

Since, e.g. every human is a hominid, every inhering instance of the class Homo sapiens Quality is also an instance of Family Hominoides Quality, etc. The introduction of qualities is helpful for

---

10 SNAP Continuant Definitions: http://www.fon.is/organ/(bfo/bfo manual/snap.pdf)
S. Schulz et al.

Fig. 2. Taxon qualities inhering in individual organism and their location in Taxon Regions consistent with the DOLCE upper level ontology.

ontological definitions such as

\[ \forall x : \text{instance of}(x, \text{Human}) \leftrightarrow \text{instance of}(x, \text{Organism}) \wedge \forall y : \text{instance of}(y, \text{HomosapiensQuality}) \wedge \text{inheres in}(y, x) \]

Based on a hierarchy of qualities, such definitions permit inferences such as that every human is a vertebrate or that every human population is part of some vertebrate population. In addition, it allows for linking organism parts with qualities such as

\[ \forall x : \text{instance of}(x, \text{VertebrateHeart}) \leftrightarrow \text{instance of}(x, \text{Heart}) \wedge \forall y : \text{instance of}(y, \text{VertebrateQuality}) \wedge \text{inheres in}(y, x) \]

If the import of the taxon concept should be extended from biological organisms to their parts, as argued in Section 3.3 (e.g. human leukocyte), the attribution of qualities to organism parts or derivatives can easily be axiomatized by the so-called right identity rules (with \( \circ \) being the relation concatenation symbol):

\[ \text{part of}(x, y) \circ \text{inheres in}(z, y) \rightarrow \text{inheres in}(z, x) \]

\[ \text{derives from}(x, y) \circ \text{inheres in}(z, y) \rightarrow \text{inheres in}(z, x) \]

3.5 Biological taxa as Qualia

An alternative approach to a subclass hierarchy based on the DOLCE upper ontology (Masolo, 2003) is represented in Figure 2. Since DOLCE is inspired by trope theory (Goodman, 1951), which distinguishes between qualities and their values (i.e. Qualia) this proposal introduces another layer of abstraction. Each quality type has an associated quality space (i.e. Region) in which it is located.

As in BFO, qualities are dependent entities which are inherent in their respective particulars. Compared to the representation depicted in Figure 1 only few taxon qualities—one for every taxon—are organized in a flat hierarchy and are related to corresponding value regions. The subsumption hierarchy of taxon qualities of the former approach is represented as a partonomic hierarchy of the Taxon Regions in the latter, e.g. the Species Region is part of the Kingdom Region. The variety of features is represented as subclasses of the basic Taxon Regions, e.g. Mammalia Class Region Is_a Class Region.

The main advantages of this approach are a clearer separation of hierarchies and the possibility to make explicit assertions on the specialized Taxon Regions without uncontrolled inheritance of restrictions. Its disadvantage lies in a higher complexity.

3.6 Synthesizing different taxon accounts

We have proposed four mutually dependent kind of ontologically relevant entities that describe different aspects of what is meant
The class `biotop:TaxonQuality`

- **Taxon classes**, the instances of which are defined as parts of some instance of `Maximal Biological Population`. For instance, 'Elephant Population in Thailand' is an instance of the class `Elephas Maxima Population`, the latter being a subclass of `Elephas Population` and so on. For each taxon there is one such population class.
- **Taxon quality classes** that are instantiated by each and every particular object to which a taxon can be ascribed. There is one such taxon quality class for each taxon. Because taxon classes are arranged in an hierarchy, the quality of a subordinate taxon is also the quality of a superordinate taxon. For example, an instance `tqClyde` of `Elephas Maxima Quality` can be ascribed to the elephant 'Clyde'. `tqClyde` is equally an instance of `Genus Elephas Quality`, of `Family Elephantidae Quality`, and so on.
- **Taxon quality regions** that are represented by a mereological inclusion hierarchy. In contrast to the third approach, every taxon-relevant entity has an inherent quality instance from each taxonomic level.

4 IMPLEMENTATION

We extended BioTop by the notion of biological taxa following the quality approach discussed in Section 3.4.

The class `biotop:TaxonQuality` has the following restrictions:\[11]

- `\exists\text{inheres}_\text{in}(\exists\text{has}_\text{part}.\text{biotop:TaxonQuality})` AND `\forall\text{inheres}_\text{in}(\exists\text{has}_\text{part}.\text{biotop:NucleicAcid})` AND `\forall\text{inheres}_\text{in}(\exists\text{has}_\text{part}.\text{biotop:ContinuantQuality})`

So we claim the existence of genetic information as a limiting and necessary condition for those entities biological taxa can be ascribed to.

In the inverse direction, we claim the inherence of taxon qualities to the classes `biotop:Cell`, `biotop:Organism`, `biotop:Tissue`, `biotop:OrganismPart`, `biotop:NucleicAcid`, e.g.

- `\text{biotop:Cell} \implies \exists\text{inheres}_\text{in}.\text{biotop:TaxonQuality}`

The class `biotop:TaxonQuality` is then the interface to a specialized ontology such as the NCBI taxonomy. For demonstration purposes we created `taxdemo`, a small example ontology.\[12]

\[11\] For the Description Logics notation cf. (Baader et al., 2003), or http://en.wikipedia.org/wiki/Description_logic

\[12\] Available at http://purl.org/biotop

In parallel, the taxonomic ranks (`TaxonQuality`, `KingdomQuality`, etc.) are indirectly represented as a second hierarchy.

This allows us to define population as a plurality of organism of the same species as follows:

- `\exists\text{has}_\text{granular}_\text{part}.\text{biotop:Organism} \implies \{\forall\text{inheres}_\text{in}.\text{biotop:TaxonQuality} \}\equiv \{\exists\text{has}_\text{granular}_\text{part}.\text{biotop:SpeciesQuality}\}

These criteria are not met by mixed groups of individuals, e.g. a group of different primates which coincide only at the level of `taxdemo:OrderQuality`.

The flexibility of our approach becomes obvious when we use taxon information for parts of the organisms. For instance, the class `HumanLeukocyte` can be defined as:

- `\text{taxdemo:HumanLeukocyte} \equiv \text{taxdemo:Leukocyte} \land \exists\text{inheres}_\text{in}.\text{taxdemo:TaxonQuality} \land \exists\text{inheres}_\text{in}.\text{taxdemo:KingdomAnimaliaQuality}

If we define:

- `\text{taxdemo:AnimalCell} \equiv \text{taxdemo:Cell} \land \exists\text{inheres}_\text{in}.\text{taxdemo:KingdomAnimaliaQuality}`
then taxdemo:HumanLeukocyte can be classified as taxdemo:AnimalCell, provided that the ontology supports:

taxdemo:HomoSapiensQuality ls_a

taxdemo:KingdomAnimaliaQuality

together with

taxdemo:Leukocyte ls_abiotop:Cell

It is obvious that this kind of reasoning can be of great advantage for biological fact retrieval from databases or for semantically enriched information extraction from texts.

From a computational perspective, however, we acknowledge that there still is a bottleneck with regard to the use of inverses (such as inheres_in versus inv_inheres_in) and qualified number restrictions (such as =1) in description logics reasoners. 13

We admit that the meaning of the taxonomic rank classes SpeciesQuality, GenusQuality, KingdomQuality, etc. is somewhat counterintuitive, since every instance of SpeciesQuality is also an instance of GenusQuality and so on. 14 They are, therefore, not suited to comprehensively represent the meaning of Species as disjoint from Genus, Kingdom, etc. Such a reading would require the meta-class representation as discussed in Section 3.1, discarded due to computational reasons. In our framework, the only way to have an instantiable Species (Genus, Kingdom) class would be to collect all maximal populations (cf. Section 3.3) with identical species-(genus-, kingdom-) level qualities as instances of Species (Genus, Kingdom) which, again, would only partially match the meaning of Species (Genus, Kingdom). We refrained from implementing the solution discussed in Section 3.5, because its more differentiated approach to the representation of qualities is not supported by the BFO upper ontology, currently in use for BioTop.

5 RELATED WORK

Literature on the ontology of taxa roughly falls into two categories: the conceptualization of the nature of species on the one hand, and the ontological status of taxa on the other. In both cases, the focus lies mainly on species whereas higher taxa are seldom addressed.

The first line of scientific discussion is characterized by numerous publications that started with the seminal book of Mayr (1942), who compared several approaches to delineate the nature of species 13 and propagated the popular concept of species as a group of organisms that interbreed and produce fertile offspring. Hull (1997) casts doubt on the monistic assumption that there is one single and ideal way to define species and hypothesizes a trade-off between theoretical significance and practical applicability of species concepts. He classifies the existing species concepts into three categories, namely, (i) similarity-based (which, of course, hinges on some unambiguous notion of phenic or genetic resemblance), (ii) biological and evolutionary (which includes Mayr’s and other proposals such as Hennig, 1966) centering around the behavior (i.e. mating, reproduction) of biological organisms and (iii) phylogenetic, focusing the historic development of species. Mayden (1997)

performed an extensive literature review and identified 22 distinct species concepts. In contradistinction to Hull, he propagates the cladistics-based evolutionary significant unit (‘Evolutionary Species Concept’, Simpson, 1961), rooted in the philosophical principle of identity: ‘An evolutionary species is an entity composed of organisms that maintains its identity from other such entities through time and over space and that has its own independent evolutionary fate and historical tendencies’. According to (Goodman, 1951) this concept of species is the most acceptable and most compatible with other species concepts that are rather criterion-based detection protocols than theoretically underpinned concepts. He argues that no criterion that presumes to delineate natural boundaries can overcome the generic vagueness (Hull, 1965) of species concepts. Our approach advocates neutrality towards the conceptualization of species and is apt to coexist with both monistic and pluralistic approaches. We are aware of the fact that in the latter case species qualities with multiple parents may be taken into account, due to different categorizations according to conflicting species concepts.

The second line of discussion is on more abstract grounds, and scrutinizes the ontological nature of species, regardless of the species concepts subtleties as exposed above. A fundamental question in here is whether species—seen as single evolving lineage that act as units of evolution—are classes or individuals, the latter being advocated by Ghiselin (1974) and Hull (1978), with the consequence that every single organism is a spatiotemporal part of its species. This theory comes close to our view of species as the totality of organisms belonging to one specific species, which can be generalized from species to taxa. We prefer this mereological approach over the set-theoretical one (also pointed out by Ereshefsky, 2007), because the view of a group of organisms as mathematical sets (that are not localized in space and time) is rather counterintuitive. The conceptualization of species as universals or natural kind conflicts with the fact that there are relatively few ‘essential’ properties that are shared by all individuals of a species (including developmental stages and malformations). Boyd’s (1999) Homeostatic Property Cluster Theory tries to overcome this, but is still too much committed to similarity-based criteria according to Ereshefsky, 2007. The approach pursued in this article, namely, introducing theory-neutral species qualities— that are extensible to general taxon qualities—seems to be rather novel.

6 CONCLUSION

We have proposed an ontological approach to biological taxa in the context of the domain top-level ontology BioTop. 16 It is essentially based upon the assumption that every biological organism, population or biological matter has some inherent taxon quality. Since it does not raise further reaching ontological claims, our approach largely bypasses the ongoing dispute on species concepts. This enables us to delineate biological populations in terms of shared taxon qualities and to formulate taxon-specific axioms in the framework of description logics. Our proposal is fully embedded into the standards of Open Biological Ontology and is in line with a major top-level ontology, BFO. Our account of taxon qualities (i.e. the preference of the

13See frequently updated list at http://www.cs.man.ac.uk/~sartle/reasons.html
14An instance of HomoSapiensQuality would be an instance of KingdomQuality, too.
15For an overview of earlier approaches see Hey (2006).
16BioTop, together with a tentative taxon-specific extension is available at http://purl.org/biotop.
simpler approach described in Section 3.4 over the more complex solution found in Section 3.5) also demonstrates how fundamental ontology design decisions depend on the choice of the underlying top-level model.

As our approach represents taxon qualities as a simple is_a hierarchy, the import of subsets of existing taxonomy databases such as the NCBI taxonomy is straightforward and scalable. These data can automatically be transformed into an OWL subtype hierarchy and linked to the BioTop node TaxonQuality.

ACKNOWLEDGEMENTS

The authors would like to thank Alan Rector (Manchester), Elena Beilwanger (Jena), Udo Hahn (Jena), Eric van Mulligen (Rotterdam) and László van den Hoek (Rotterdam), as well as Olivier Bodenreider (Bethesda), for fruitful discussions.

Funding: This work was supported by the EC STREP project ‘BOOTStrep’ (FP6 – 028099).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. et al. (2000) Gene ontology: tool for the unification of biology. Gene ontology consortium. Nat. Genet., 28, 25–29.

Baader,F. et al. (eds) (2003) The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, Cambridge, UK.

Bittner,T. et al. (2004) Individuals, universals, collections - on the foundational relations of ontology. In Formal Ontology in Information Systems: Proceedings of the 3rd International Conference – FOIS. 2004. IOS Press, Amsterdam, The Netherlands, pp. 37–48.

Boy golden, (1999) Homostasis, species, and higher taxa. In Wilson,R. (ed.), Species: The Units of Biodiversity. Special Vol. 54. Systems Association, Chapman and Hall Ltd, London, pp. 357–380.

International Organization for Standardization (ISO) (2000) ISO 1087 – 1: Terminology Work – Vocabulary – Part 1: Theory and Applications. Geneva, Switzerland.

Kuleszewska,W. (2006) Nontological Engineering. Formal Ontology In Information Systems. Proceedings of the 4th International Conference FOIS 2006, IOS Press, Amsterdam, The Netherlands, pp. 59–50.

Maushof,C. et al. (2003) WonderWeb Deliverable D4.3 Ontology Library. Infrastructure for the Semantic Web. http: //wonderweb.semanticweb.org, last accessed date January 11, 2008.

Mayden,R.L. (1997) A hierarchy of species concepts: the denotement of the species problem. In Clancy,M.F. et al. (eds). Species: The Units of Biodiversity. Special Vol. 54. Systems Association, Chapman and Hall Ltd, London, pp. 381–424.

Mary,E. (1942) Systematics and the Origin of Species. Columbia University Press, Irvington/NY, USA.

Mary,E. (1999) Principles of Systematic Zoology. McGraw-Hill, New York.

Olta,T. et al. (2002) GENIA Corpus: an annotated research abstract corpus in molecular biology domain. Proceedings of the Human Language Technology Conference (HLT 2002): pp. 73–77.

Queiroz,O. (1948) On what there is: review of metaphysics. In From a Logical Point of View. Harper & Row, New York.

Rector,A. et al. (2006) Granularity, scale and collectivity: when size does and does not matter. J. Biomed. Informatics, 39, 333–340.

Rector,A. et al. (2007) Simple Bus Upper Ontology. [http: //www.cs.man .ac.uk/~reector/ontologies/simple-top-bus] – last accessed date January 15, 2008.

Rubin,D. et al. (2008) Biomedical ontologies: a functional perspective. Brief. Bioinform., 9, 75–90.

Schulz,S. et al. (2006a). Biomedical ontologies: what part-of is and isn’t. J. Biomed. Informatics, 39, 350–361.

Schulz,S. et al. (2006b). From GENIA to BioTop – Towards a Top – domain Ontology for Biology. The 4th International Conference on Formal Ontology in Information Systems (FOIS 2006), Baltimore, USA, pp. 103–114.

Schulz,S. and Hahn,U. (2007). Towards the ontological foundations of symbolic biological theories. Artif. Intell. Med., 39, 257–250.

Schulz,S. and Johnson,J. (2007). Continues in biological systems. The Monist, 90, 4 October 2007.

Simpson,G.G. (1961) Principles of Animal Taxonomy. Columbia University Press, Irvington/NY, USA.

Smith,B. et al. (2005) Relations in biomedical ontologies. Genome Biol., 6, R46.

Smith,B. et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol., 25, 1251–1255.

Steenhoorn,H. et al. (2007) Towards a top – domain ontology for linking biomedical ontologies. In: Kuhn,K. A. et al. (eds). MEDINFO 2007 – Proceedings of the 12th World Congress on Health (Medical) Informatics – Building Sustainable Health Systems. IOS Press, Amsterdam, pp. 1225–1229.

Wheeler,D.L. et al. (2008) Database resources of the national center for biotechnology information. Nucleic Acids Res., 36, D13–D21.

Guarino,N. and Giaretta,P. (1998) Ontologies and knowledge bases: towards a terminological clarification. In Towards Very Large Knowledge Bases. Knowledge Building and Knowledge Sharing. IOS Press, Amsterdam, The Netherlands, pp. 25–32.

Heller,R. and Here,H. (2004) Ontological Categories in GDL. Artificial, Vol. 14, Springer, Heidelberg, Germany, pp. 57–76.

Hennig,W. (1966) Phylogenetic Systematics. University of Illinois Press, Urbana.

Hey,J. (2006) On the failure of modern species concepts. Trends Ecol. Evol., 21, 447–450.