

# Applying ontology design patterns to the implementation of relations in GENIA

Robert Hoehndorf<sup>\*,1,4</sup> , Axel-Cyrille Ngonga Ngomo<sup>\*,2</sup> , Sampo Pyysalo<sup>3</sup> , Tomoko Ohta<sup>3</sup> , Anika Oellrich<sup>1</sup> and Dietrich Rebholz-Schuhmann<sup>1</sup>

<sup>1</sup>European Bioinformatics Institute, Hinxton, Cambridge, UK

<sup>2</sup>Department of Computer Science, University of Leipzig, Leipzig, Germany

<sup>3</sup>Department of Computer Science, University of Tokyo, Tokyo, Japan

<sup>4</sup>Department of Genetics, University of Cambridge, UK

Email: Robert Hoehndorf\* – rh497@cam.ac.uk; Axel Ngonga\* – ngonga@informatik.uni-leipzig.de; Sampo Pyysalo – smp@is.s.u-tokyo.ac.jp; Tomoko Ohta – okap@is.s.u-tokyo.ac.jp; Anika Oellrich – anika@ebi.ac.uk; Dietrich Rebholz-Schuhmann – rebholz@ebi.ac.uk;

\*Corresponding author

## Abstract

---

**Motivation:** Annotated reference corpora such as the GENIA corpus play an important role in biomedical information extraction. A semantic annotation of the natural language texts in these reference corpora using formal ontologies and logic is challenging due to the ambiguous use of natural language and natural language semantics. Providing formal definitions and axioms for these relations would offer the means for developing consistent and verifiable annotation guidelines and allow for the automatic verification of annotations as well as enabling the discovery of new information through deductive inferences.

**Results:** We developed a formal ontology of relations based on the relations used in the recent GENIA corpus annotations. For this purpose, we selected existing axiom systems based on the desired properties of the relations within the domain and provided new axioms for several relations. To apply this ontology of relations to the semantic annotation of natural language texts, we developed and implemented two ontology design patterns. We provide an implementation of the ontology of relations in the Web Ontology Language (OWL). By combining the implementation of the design patterns and that of the relation ontology, we also provide a software application to convert annotated GENIA abstracts into OWL ontologies. In this way, we make these ontologies amenable for automated verification, deductive inferences and other knowledge-based applications.

**Availability:** Documentation, implementation and examples are available from <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/>.

**Contact:** rh497@cam.ac.uk

---

## 1 Background

The goal of Information Extraction (IE) is to recognize specific pieces of information in natural language texts and to represent them in a structured form that comprises meaningful associations of relevant entities. For this reason, IE approaches typically involve Named Entity Recognition (NER) where *mentions* of specific types of “real-world” entities, such as people or places, are detected in text. To facilitate reliable biomedical IE, considerable efforts have been made with regard to the development of specialized NER methods for key domain entities, focusing in particular on the recognition of gene and gene product (GGP) mentions (1; 2; 3). As GGP mentions can further be normalized to identify specific entries in databases such as UniProt, they provide a connection to entities relevant to biomolecular research and thus a solid basis for domain IE. However, in contrast to the well-defined meaning of the basic entities, the semantics of their associations are often only informally defined.

In biomedical IE, extracted information is frequently represented simply as untyped pairs of entities representing, for instance, protein-protein or gene-disease associations (4). However, even resources identifying protein-protein interactions as entity pairs diverge considerably in their actual annotations (5), leading to restrictions ranging from usability to interpretability of both the annotations and IE results. In response to the limitations of such representations, there has recently been increased interest in richer representations of extracted information (6) and a number of corpora have been published that annotate associations between entities by using fine-grained types drawn from ontologies (7; 8). Yet, no definition or axiomatization of these relations has been proposed so far. Definitions and axioms are necessary to make the *meaning* of the relations explicit, and to provide the means for developing consistent and verifiable annotation guidelines allowing for the automatic detection of inconsistent annotations, and enabling the discovery of new information through deductive inferences. Here, our aim is to define such relations and axioms for fundamental relations such as *part-of* connecting GGPs to referents of non-specific domain terms such as *promoter region*. Annotations to these fundamental relations to have been introduced recently (9; 10) to the widely used GENIA corpus (11).

<sup>1</sup><http://code.google.com/p/information-artifact-ontology/>

Providing formal definitions and axioms for these relations is challenging because the annotations are based on the use of the relations in text, where it is generally not possible to enforce a common understanding of terms. We present a formal characterization of the relations used in GENIA annotation based on two ontology design patterns. These patterns are not restricted to GENIA and can be applied in a wide number of domains, in particular in ontology- and knowledge-based applications using the categories of biological sequences, DNA, RNA or proteins. We implement the developed formalisms in OWL and provide a conversion software to represent GENIA annotations in OWL.

### 1.1 The GENIA corpus

The GENIA corpus consists of 2,000 PubMed abstracts annotated manually by biomedical domain experts as a resource for the development and evaluation of domain information extraction (IE) methods. GENIA is one of the most widely used corpora for biomedical IE and has served as the basis for two shared tasks on named entity recognition (1) and event extraction (6). The corpus annotation includes markup that identifies occurrences of domain terms and named entities as well as statements of events and relations involving them (8; 9; 11; 12).

### 1.2 Formal ontology

An ontology is the formal specification of a conceptualization of a domain (13). A conceptualization is a system of categories accounting for a particular view on the world (14). Ontologies are used to specify the meaning of terms within a vocabulary. A basic ontological distinction is made between classes and individuals (or particulars). A *class* is an entity that can be predicated of other entities and that can have instances. The **instance-of** relation links *instances* to the *class* of which they are an instance. Some instances may be classes themselves and have further instances. An *individual* is an entity that cannot be further instantiated (15).

For the purpose of formalizing the relations used in the GENIA corpus, we make use of several biomedical domain ontologies: the Information Artifact Ontology<sup>1</sup> (IAO), the Sequence Ontology (SO) (16), the Ontology of Biomedical Investigations (OBI) (17), the Gene Ontology (GO) (18) and the GENIA term ontology (11).

### 1.3 Preliminaries of GENIA corpus annotation

The first question we have to answer before we can formalize relations used in corpus annotation is what kind of entities are connected through these relations. Our first observation is that relations in corpus annotations are usually asserted between names and other biomedical domain terms, i.e., between strings that are identified as referring to some kind of entity. For the purpose of this work, we assume that these names denote one entity that can be either a class or an individual.

In some cases, there is ambiguity in determining the referent of a name or domain term, i.e., certain terms may not refer to identical entities, yet their referents are regarded as *indistinguishable* within the context of a task such as the annotation or recognition of named entities. Regarding certain referents as indistinguishable can improve the automatic extraction of relations and entities. The indistinguishability assumption also allows the definition of generic relations that hold between disjoint categories. Through these means, the effort to create annotation can be reduced, while the applicability of the relations in different tasks and the feasibility of automatic extraction can be maximized. Within GENIA annotations (12), and the NER systems based on it, genes and gene products are not distinguished. Therefore, a basic precursor for our work is an equivalence relation which states that, within the context of a named entity annotation task, two classes are considered to be *indistinguishable*.

## 2 Results

### 2.1 Equivalence

Names or terms referring to either a class of genes, DNA, proteins, RNAs and their splice variants, gene products, arbitrary transcripts or similar are considered to be equivalent within the context of the GENIA relation annotations. These classes are called *genes/gene products* (GGPs). For example, *CD19*, *CD19 protein* and *CD19 gene* may be considered to be equivalent and represent a single GGP.

To provide a decidable implementation of our formalization, and to facilitate automated queries, verification and inferences, we provide a definition of GGP-equivalence in OWL. We define a class  $G_C$  based on a class  $C$ , which is assumed to be a subclass of  $DNA$ , and entities derived from  $C$  through chains of transcription and translation relations *between in-*

*dividuals*. The classes *Protein*, *DNA* and *RNA* are those used in the GENIA term ontology.

$$\begin{aligned} C \sqcup (RNA \sqcap \exists \text{transcribedFrom}.C) \sqcup \\ (Protein \sqcap \exists (\text{translatedFrom} \circ \\ \text{transcribedFrom}).C) \sqsubseteq G_C \end{aligned} \quad (1)$$

Such a formalization has the benefit of connecting the different kinds of GGPs through formal relations that can be exploited by an automated reasoner.

For example, the name “CD19 protein” refers to a class of proteins, and instances of this class stand in a *translated-from* relation to instances of a class of RNA which may be referred to as “CD19 RNA”. Instances of this class of RNA stand in a *transcribed-from* relation to instances of a class of DNA which may be referred to as “CD19 gene”. Thus, according to our definition, all three classes are subclasses of the GGP class  $G_{CD19}$ .

### 2.2 Subclass

The *class-subclass* relation is used to annotate the relation between terms or names in the GENIA corpus where one term refers to a more general class than the other term. For example, this relation holds between the names “CD19 human” (denoting the class *CD19 human*) and “CD19” (denoting a class that is indistinguishable from the class *CD19 (GGP)*). We base the definition of the *class-subclass* relation upon the ontological *is-a* relation (19): the classes  $C$  and  $D$  stand in the *is-a* relation, if and only if, every instance of  $C$  is also an instance of  $D$ .

For example, the referent of the name “human CD19 gene” (the class *CD19 human gene*) stands in the *is-a* relation to the referent of the name “CD19” (the GGP class *CD19 (GGP)*), because all instances of *CD19 human gene* are also instances of *CD19 (GGP)*.

### 2.3 Mereological relations

The largest group of relations in the relationship annotations of the GENIA corpus refers to mereological relations, i.e., relations between parts and their wholes. Three kinds of parthood relations are distinguished within GENIA:

- relations between a whole and its components, for example between the classes *CD19 promoter* and *CD19*,
- relations between a collection and its members, as between *Hox gene family* and *HOXA1*,

- the relation between an entity and the location at which this entity exists, such as *CD19* which is located at *CD19 locus*.

Substantial work has already been undertaken with regard to mereological relations and their representation in OWL and biomedical ontologies (20; 21; 22). In particular, the relation *CC-part-of*<sup>2</sup>, as a relation between classes, must be defined in terms of another relation *II-part-of*, which is a relation between individuals (20; 23). For example, **CC-part-of** can be defined as

$$C \sqsubseteq \exists \text{partOf}.D \quad (2)$$

Although such a definition is valid for many of the parthood relations asserted between classes in biological ontologies, it is an inadequate schema for parthood relations which have a GGP class as argument, because the GGP class is “too general”.

However, as a GGP class has several *GGP-equivalent* subclasses, the *CC-has-part* and *CC-part-of* relations may be valid for one of these classes but not for the others. For example, assuming the definition of *CC-has-part* above, asserting a *CC-has-part* relation between the GGP class *CD19 (GGP)* and *CD19 promoter* would be incorrect, because the GGP class will also include the *CD19 protein* class, which has no promoter as part (in virtue of being a class of proteins). Similarly, although it would be correct to assert that *CD19 promoter CC-part-of CD19*, it would be incorrect to say that *CD19 CC-part-of CD19/CD21/CD81/Leu-13 complex*. If the two statements above would hold, we could infer that *CD19 promoter* is *CC-part-of* the *CD19/CD21/CD81/Leu-13 complex*, which is incorrect because protein complexes have no promoters as part.

Consequently, we use the following alternative definition for the *GGP-subclass-has-part* relation (where the argument  $G_C$  refers to a GGP class, and  $X$  to an arbitrary class):

$$\begin{aligned} \text{GGP-subclass-has-part}(G_C, X) \iff \\ (G_C \sqcap \text{DNA} \sqsubseteq \exists \text{II-hasPart}.X) \text{ or} \\ (G_C \sqcap \text{RNA} \sqsubseteq \exists \text{II-hasPart}.X) \text{ or} \\ (G_C \sqcap \text{Protein} \sqsubseteq \exists \text{II-hasPart}.X) \end{aligned} \quad (3)$$

In the OWL syntax, a disjunction of subclass axioms is not permitted. Consequently, we have to

<sup>2</sup>We generally prefix relations between two classes with *CC-*, and relations that hold between two individuals with *II-*. The *CC-* type relations are not available in OWL but are defined using complex description logics statements and converted to OWL using the software tool we provide.

reformulate the right side of the definition by using a single subclass axiom (where  $\perp$  refers to the OWL class `owl:Nothing`) and derive the equivalent definition:

$$\begin{aligned} \text{GGP-subclass-has-part}(G_C, X) \iff \\ (G_C \sqcap \text{DNA} \sqcap \neg \exists \text{II-has-part}.X) \sqcup \\ (G_C \sqcap \text{RNA} \sqcap \neg \exists \text{II-has-part}.X) \sqcup \\ (G_C \sqcap \text{Protein} \sqcap \neg \exists \text{II-has-part}.X) \sqsubseteq \perp \end{aligned} \quad (4)$$

Intuitively, this definition states that if the GGP class  $G_C$  stands in the *GGP-subclass-has-part* relation to the class  $X$ , then either the DNA, RNA or Protein subclass of  $G_C$  must stand in a *CC-has-part* relation to  $X$ . Using this pattern, we are further able to define the relation *GGP-subclass-part-of* by replacing *II-has-part* with *II-part-of* in definition 4.

*II-part-of* is a primitive relation and we assert axioms that hold for it. *II-part-of* is reflexive, transitive and antisymmetric. We define *II-proper-part-of*:

$$\text{II-proper-part-of}(x, y) \iff \text{II-part-of}(x, y) \wedge \neg x = y \quad (5)$$

It is the *II-proper-part-of* relation which will provide the basis for the mereological relations within the GENIA, because identical (or co-extensional) classes are not annotated as standing in a parthood relation.

Parthood relations that are not based upon location are further distinguished into two kinds in the GENIA relation annotation: a relation between components and the objects of which they are components, and membership in collections. We assume that the component-object relation (between individuals) *II-oc-part-of* is similar to the relation of determinate parthood (21) in that it is reflexive, transitive, antisymmetric and satisfies the strong supplementation principle (22). Assuming these axioms for *II-oc-part-of* provides compatibility with the SO, which also assumes the axioms of extensional mereology for the entities classified by it (16).

The member-component relation, on the other hand, is a relation between entities of different kinds and is neither reflexive nor antisymmetric (21; 24). The *II-member-of* relation is a sub-relation of the *II-proper-part-of* relation and is non-reflexive, asymmetric and non-transitive (24). *II-member-of* is not the same relation as the *member-of* relation in the SO; in the SO, *member-of* is transitive, while *II-*

*member-of* is non-transitive. The relation *GGP-subclass-member-of* holds between a GGP class and a collection, such that for one of the subclasses of the GGP class, all instances are a member of some instance of the collection. Therefore, the same pattern as in definition 4 applies for the definition of *GGP-member-of*.

For example, the *Lck* (*GGP*) class stands in the *GGP-member-of* relation to the protein family *Src family*, because there is a subclass of *Lck* (*GGP*), i.e., *Lck protein*, such that all instances of this subclass (*Lck protein*) stand in an *II-member-of* relation to some instances of *Src family*<sup>3</sup>.

The third parthood relation used in the GENIA corpus annotations is *GGP-subclass-region-of*, which we define by using the primitive *II-region-of* relation. In the GENIA relation annotations, *GGP-subclass-region-of* is used to relate a GGP class to a genomic location. We introduce *GGP-subclass-region-of* to relate the GGP class to the class of loci. The region is a place where all instances of one subclass of the GGP class are located. As for the definition of *GGP-subclass-has-part*, *GGP-subclass-part-of* and *GGP-subclass-member-of*, we assume that there is a subclass of the GGP class for which all instances are located in some instance of the locus, and we use the same pattern as in formula 4.

Next we define the interactions of *II-region-of* with *II-part-of*. We want to be able to infer that if the individual  $x$  is part of  $y$ , and  $y$  is located at  $z$ , then  $x$  is located at  $z$ . Furthermore, if the individual  $x$  is located at  $y$  and  $y$  is a part of  $z$ , then we infer that  $x$  is located at  $z$ .

$$II\text{-part-of} \circ II\text{-region-of} \subseteq II\text{-region-of} \quad (6)$$

$$II\text{-region-of} \circ II\text{-part-of} \subseteq II\text{-region-of} \quad (7)$$

## 2.4 Objects and their variants

The second major group of GENIA corpus relations connects names of GGP classes to names of classes of their variants. Again, we formalize the relations that hold between the classes that are *denoted* by these names.

The GENIA annotations for GGP classes and their variants use six different relations to express the following relationships:

- GGPs to modified proteins, e.g., *TR alpha 1* (*GGP*) to *35S-TR alpha 1* (*Protein*),
- GGPs to protein isoforms, e.g., *ACTA1* (*Protein*) to *G-Actin* (*GGP*),
- GGPs to mutants, e.g., *TNFR1* (*GGP*) to *dominant-negative mutant TNFR1* (*Protein*),
- GGPs to recombinants, e.g., *Oct-2* (*GGP*) to *Oct-2 expression vector* (*DNA*),
- GGPs to precursors, e.g., *IL-16* (*GGP*) to *pro-IL-16* (*Protein*),
- GGPs to experimental material, in particular to antisense elements, e.g., *GATA-3* (*GGP*) to *antisense GATA-3 RNA* (*RNA*).

We will call the basic relation between a GGP and its variant *GGP-has-variant*. There is a general schema involved in the sub-relations of *GGP-has-variant* that we exploit in its definition: whenever *GGP-has-variant*( $G_C, D$ ), then every instance of  $D$  is a variation of some instance of  $G_C$ . Although it is possible to identify a more specific subclass of  $G_C$  in some cases, this is not true for all sub-relations of *GGP-has-variant*. We define the relation  $G_C$  *GGP-has-variant*  $D$  by using the relation *II-has-variant*, which is a relation between individuals:

$$D \sqsubseteq \exists II\text{-has-variant}.G_C \quad (8)$$

Again, we provide basic axioms for the *II-has-variant* relation. Our first observation is that variance is reflexive, i.e., everything (every molecule) is a variant of itself. Furthermore, variance is symmetric, i.e., if  $x$  is a variant of  $y$ , then  $y$  is a variant of  $x$ . Whether *II-has-variant* is transitive is more difficult to ascertain. While it seems to be the case that, if  $x$  is a variant of  $y$  and  $y$  a variant of  $z$ , then  $x$  is a variant of  $z$ , this principle may fail if the distance between  $x$  and  $z$  increases, i.e., more intermediate variants are introduced. Consequently, we do not assume that *II-has-variant* is transitive.

To formalize a sub-relation of *II-has-variant*, e.g., *II-has-isoform*, we note domain and range of the relation as well as basic axioms. In the definition of the GGP relation, we must carefully consider whether the relation holds between all instances of the GGP class, or only one of its subclasses. For example, the

<sup>3</sup>We do not provide a formal characterization of *protein family* here, but re-use the class from the GENIA term ontology. Arguably, protein families should not be classes but rather individual collections. If this approach is taken, protein family classes can be defined using nominals as having exactly one instance: the individual collection that constitutes the protein family.

definition of *GGP-has-isoform* between  $G_C$  and  $D$  is:

$$\begin{aligned}
 G_C \sqsubseteq & \exists II\text{-has-isoform}.D \sqcup \\
 & \exists \text{translates-into}.\exists II\text{-has-isoform}.D \sqcup \\
 & \exists \text{transcribes-into}.\exists \text{translates-into}. \\
 & \exists II\text{-has-isoform}.D
 \end{aligned} \quad (9)$$

The relations **GGP-has-recombinant**, **GGP-has-precursor** and **GGP-has-modified-protein** follow the same pattern.

*II-has-mutant* is a relation between an instance of a GGP class and a mutant of this instance. The relation *II-has-mutant* is irreflexive and symmetric, and consequently not transitive. The definition of  $G_C$  *GGP-has-mutant*  $D$  is as follows:

$$G_C \sqsubseteq II\text{-has-mutant}.D \quad (10)$$

*II-has-experimental-material* relates an instance of a GGP class to experimental material such as an antisense element. The formal characterization is subject to future work and requires integration with ontologies of experiments such as the Ontology of Biomedical Investigations (OBI) (17).

### 3 Implementation

We provide an implementation which consists of two parts. The first part covers the integration of the basic axioms of relations between individuals into an OWL ontology. It formalizes GENIA's relation ontology and provides the taxonomy of relations as illustrated in figure 1. To be applicable for automated inferences, we had to omit axioms pertaining to reflexivity or symmetry from the OWL ontology, as those are not permitted for non-primitive properties (25). The OWL ontology contains the hierarchy of relations and a single new OWL class, the class *GGP*. Furthermore, to provide the definitions of the relations, we also import the OWL versions of the Sequence Ontology (SO) (16) and the GENIA term ontology (26) so that we can refer to relations such as *transcribes-into* from the SO, and to classes such as *DNA* or *Protein* from the GENIA term ontology. The second part provides a conversion from the relations between names and terms that refer to classes in OWL. It is a prototypical conversion tool that translates annotated GENIA abstracts into an OWL file based on the definitions we provide for GENIA's relationship annotations.

II-relation	Axioms
II-has-region	IR, AS
II-has-part	R, T, Anti
II-has-oc-part	R, T, Anti
II-has-proper-oc-part	IR, T, Anti
II-has-proper-part	IR, AS
II-has-member	IR, T, Anti
II-has-proper-oc-part	IR, T, Anti
II-has-variant	R, S
II-has-isoform	T, S
II-has-modified-protein	IR, AS
II-has-mutant	IR, T, S
II-has-precursor	IR, T, AS
II-has-recombinant	IR, T, AS
II-part-of	R, T, Anti
II-oc-part-of	R, T, Anti
II-oc-proper-part-of	IR, T, Anti
II-proper-part-of	IR, AS
II-member-of	IR, T, Anti
II-oc-proper-part-of	IR, T, Anti

Figure 1: Axioms for the relations in the GENIA relation ontology. R stands for reflexivity, IR for irreflexivity, S for symmetry, T for transitivity, Anti for antisymmetry and AS for asymmetry.

The resulting OWL file is based on GENIA's relation ontology. The conversion tool implements the ontology design patterns we have developed to define relations that take a GGP class as an argument. The conversion tool and examples of converted abstracts can be found on the project website<sup>4</sup>.

## 4 Discussion

### 4.1 Related work

The BioTop Ontology (27) is derived from the GENIA term ontology and provides definitions and axioms for the classes in the GENIA ontology. Additionally, this ontology includes several relations. Some of these relations overlap with those used in the GENIA relation annotation and in the relation ontology, in particular the mereological relations. Yet, BioTop includes mostly the generic definitions of mereological relations. Thus, BioTop's formalization of mereological relations cannot be used with respect to GGP, as their axioms do not always hold

<sup>4</sup><http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=Relation+annotation>

for GGP as shown earlier. Furthermore, the BioTop ontology does not include any of the variance relations. As BioTop provides a rich axiom system for the classes of the GENIA term ontology, we will aim at integrating the BioTop ontology with the relation ontology and the design patterns we provide in future work.

Another relevant ontology is the Gene Regulation Ontology (GRO) (28), which is an ontology for the domain of gene regulation. It provides axioms and definitions for the classes DNA, RNA and protein. Furthermore, it establishes relations between these classes. Therefore, it provides a means for a more detailed specification of GGP classes. GRO does not cover the relations formalized in this work. Rather, it could be allow to provide a more fine-grained definition of GGP classes if necessary.

## 4.2 Applications in GENIA

There are several applications of formalized relations within the GENIA corpus:

- development of unambiguous annotator guidelines,
- verification of annotations,
- inference of hidden knowledge and
- abductive reasoning, inductive logic programming, rule learning.

Firstly, the development of clear annotator guidelines can be facilitated to increase inter-annotator consistency through the provision of less ambiguity. For this purpose, high expressivity is necessary to specify the meanings of relationship terms or other terms as precisely as possible. To proceed towards the goal of unambiguous, formal guidelines for corpus annotation, we used predicate logic for the formalization, and additionally associated our definitions and axioms with explanations in natural language.

Secondly, the axioms provide a means to *verify* annotations. Such a verification is made possible because axioms restrict the combinations of relations and may lead to contradictions which are sometimes automatically detectable. In particular, the OWL implementation of both the axioms and the ontology design patterns is amenable to automated reasoning and can be used to detect inconsistencies. Additionally, it is possible to draw inferences from the asserted knowledge automatically. These inferences

can be used to verify whether or not erroneous annotations have been asserted by identifying undesired or false inferences. Moreover, automatic inferences can be used to infer hidden or new knowledge.

The conversion tool we provide converts annotated GENIA abstracts into an OWL ontology. This conversion is a form of ontology induction or ontology generation. The resulting ontologies – each covering a domain described within one abstract – can be used for abductive or inductive logic programming, rule learning or other knowledge-based machine learning techniques.

## 4.3 Ontology design patterns

To provide definitions for the relations between classes that are used in the GENIA corpus, we developed two closely related ontology design patterns (29). They are particularly suited for applications in text mining where the exact referent of a term cannot always be reliably determined. However, the patterns could be useful in other domains and applications as well.

The first ontology design pattern is applicable when a class  $C$  with the subclasses  $D_1, \dots, D_n$  stands in a relation  $CC-R$  to a class  $E$  such that every instance of at least one subclass of  $C$  stands in a relation  $II-R$  to some instance of  $E$ . This pattern is useful when one class cannot be entirely disambiguated, and a superclass is used in a relation statement instead. For example, GGP classes in GENIA are primarily introduced because it is not always possible – or reasonable – to disambiguate entirely whether a term refers to *DNA*, *RNA* or *Protein* classes. Instead, the GGP class is used in relation statements, and the GGP class unifies the classes of *DNA*, *RNA* and *Protein*. In many cases, the relation is only relevant for the instances of one of the subclasses, e.g. only the *Proteins*, such that some property or relation applies to every instance of this subclass but not to all the instances of the other subclasses.

The specialized pattern for a relation  $GGP-subclass-R$  is as follows:

$$\begin{aligned}
 CC-R(G_C, X) \iff & (G_C \sqcap DNA \sqcap \neg \exists II-R.X) \sqcup \\
 & (G_C \sqcap RNA \sqcap \neg \exists II-R.X) \sqcup \\
 & (G_C \sqcap Protein \sqcap \neg \exists II-R.X) \sqsubseteq \perp
 \end{aligned}
 \tag{11}$$

The pattern in formula 11 can be further generalized, as it still uses the classes *DNA*, *RNA* and *Protein*. In

terms of a class  $C$  with subclasses  $D_1, \dots, D_n$  whose instances are standing in a relation  $II-R$  to some instance of  $E$ , the pattern is formulated as follows (where  $R$  is the relation between the two classes):

$$R(C, E) \iff (C \sqcap D_1 \sqcap \neg \exists II-R.E) \sqcup \dots \sqcup (C \sqcap D_n \sqcap \neg \exists II-R.E) \sqsubseteq \perp \quad (12)$$

The second ontology design pattern is derived from the definitions of the *has-variant* relations. It is applicable when every instance of a GGP class is related by the relation  $II-S$  either to some instance  $x$  of a class  $D$ , or to some individual which stands in a combination of the relations  $\mathbf{T}_1, \dots, \mathbf{T}_m$  to  $x$ . The general pattern is as follows:

$$S(G_C, D) \iff G_C \sqsubseteq \exists II-S.D \sqcup \exists II-S \circ T_1.D \sqcup \dots \sqcup \exists II-S \circ T_1 \circ \dots \circ T_m.D \quad (13)$$

In general, it is possible to consider either an order defined on the relations  $T_1, \dots, T_m$  or arbitrary permutations. Intuitively, the pattern is used to state that all instances of one general class (the GGP class in the case of GGP annotations) stand in a relation  $II-S$  to some instance of a class  $D$  or to any entity reachable by a chain (or permutation) of the relations  $T_1, \dots, T_m$  from any instance of this class.

#### 4.4 Future research

Although the formalization of relationships used in the GENIA annotation is itself valuable to provide a means for automated inferences, verification and the development of annotation guidelines, formalized relations will be much more useful in combination with a formal characterization of *events* (8). Events include more dynamic entities such as the *binding* of a molecule to a binding site. In conjunction with the formalization of the relations, useful inferences can be drawn. For example, from the assertion that a class  $X$  *binds*  $Y$  which is a *GGP-part-of*  $Z$ , we would be able to infer that  $X$  *GGP-binds*  $Y$ . However, a formalization of the GENIA event annotations and its interrelation are subject to future work. Furthermore, an extensive evaluation of the utility of the axioms and definitions for the verification of annotations and the inference of hidden knowledge is subject to future work. To support the detection of complex annotation inconsistencies in the evaluation, a formalization of the event annotations is required in addition to the relation annotations.

## Conclusions

We present and discuss a formal ontology-based characterization of the relations used for annotating the GENIA corpus. The main challenge is the ambiguity of the terms upon which the relations we are interested in are based. These terms refer to one of several ontological classes, and the definitions of the relationships between two terms must reflect the fact that only one of these classes may stand in some relation to another class. To characterize these phenomena formally, we introduce the notion of a *GGP class*, which is an ontological class with subclasses whose names are not distinguished within a certain annotation task. In particular, the class is a common superclass for classes of DNA, RNA and proteins, and is intended to unify classes of genes and their products (GGP stands for “gene/gene product”).

To define relations that hold between a GGP class and another class formally, we introduce two ontology design patterns. The ontology design patterns are general enough to be useful for other domains and applications besides text mining, although they are especially useful whenever it is not possible – or not feasible – to determine the exact class that stands in some relation to another class, and a more general class is chosen in a relation statement instead.

We implement the axioms and definitions as well as the ontology design patterns in a tool that converts GENIA abstracts into OWL ontologies. These OWL ontologies can be used subsequently to answer queries, verify annotations or provide a basis for knowledge-based machine learning techniques.

Formalizing the relations used in the relationship annotations of the GENIA corpus provides a powerful means to verify the annotations, to use them for knowledge-based machine learning techniques and inferences, and to establish and communicate unambiguous and precise annotation guidelines. However, the relations that are used in the GENIA annotations, and the axioms and definitions we provide for them, are applicable and useful beyond GENIA, and can be integrated in other ontology- or knowledge-based resources such as ontologies of biological sequences, RNA or proteins. Similarly, the ontology design patterns we developed are useful not only in defining the relations used in the GENIA corpus annotations, but have an impact on other efforts to annotate text corpora semantically, and can additionally be used for defining relations between ontological classes in general.



## References

1. Kim JD, Ohta T, Tsuruoka Y, Tateisi Y, Collier N: **Introduction to the bio-entity recognition task at JNLPBA**. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA)* 2004:70–75.
2. Yeh A, Morgan A, Colosimo M, Hirschman L: **BioCreAtIvE Task 1A: gene mention finding evaluation**. *BMC Bioinformatics* 2005, **6**(Suppl 1):S2.
3. Wilbur J, Smith L, Tanabe L: **BioCreative 2. Gene Mention Task**. In *Proceedings of Second BioCreative Challenge Evaluation Workshop*. Edited by Hirschman L, Krallinger M, Valencia A 2007:7–16.
4. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB: **Frontiers of biomedical text mining: current progress**. *Brief Bioinform* 2007, **8**(5):358–375, [http://bib.oxfordjournals.org/cgi/content/abstract/8/5/358].
5. Pyysalo S, Airola A, Heimonen J, Björne J, Ginter F, Salakoski T: **Comparative analysis of five protein-protein interaction corpora**. *BMC Bioinformatics* 2008, **9**(Suppl 3):S6, [http://www.biomedcentral.com/1471-2105/9/S3/S6].
6. Kim JD, Ohta T, Pyysalo S, Kano Y, Tsujii J: **Overview of BioNLP'09 Shared Task on Event Extraction**. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task* 2009:1–9.
7. Pyysalo S, Ginter F, Heimonen J, Björne J, Boberg J, Järvinen J, Salakoski T: **BioInfer: a corpus for information extraction in the biomedical domain**. *BMC Bioinformatics* 2007, **8**:50.
8. Kim JD, Ohta T, Tsujii J: **Corpus annotation for mining biomedical events from literature**. *BMC Bioinformatics* 2008, **9**(10).
9. Pyysalo S, Ohta T, Kim JD, Tsujii J: **Static Relations: a Piece in the Biomedical Information Extraction Puzzle**. In *Proceedings of the BioNLP 2009 Workshop*, Association for Computational Linguistics 2009:1–9.
10. Ohta T, Sampo P, Kim JD, Tsujii J: **A Re-evaluation of Biomedical Named Entity - Term Relations**. In *Proceedings of the 3rd International Symposium on Languages in Biology and Medicine (LBM2009)* 2009:97–102.
11. Kim JD, Ohta T, Tateisi Y, Tsujii J: **GENIA corpus – a semantically annotated corpus for bio-textmining**. *Bioinformatics* 2003, **19**(suppl.1):i180–182.
12. Ohta T, Kim JD, Pyysalo S, Wang Y, Tsujii J: **Incorporating GENETAG-style annotation to GENIA corpus**. In *Proceedings of the BioNLP 2009 Workshop* 2009:106–107.
13. Gruber TR: **Toward principles for the design of ontologies used for knowledge sharing**. *International Journal of Human-Computer Studies* 1995, **43**(5-6), [http://dx.doi.org/10.1006/ijhc.1995.1081].
14. Guarino N: **Formal Ontology and Information Systems**. In *Formal Ontology in Information Systems: Proceedings of the First International Conference (FOIS'98), Trento, Italy, 6-8 June 1998, Volume 46 of Frontiers in Artificial Intelligence and Applications*. Edited by Guarino N, Amsterdam: IOS Press 1998:3–15.
15. Herre H, Heller B, Burek P, Hoehndorf R, Loebe F, Michalek H: **General Formal Ontology (GFO) – A Foundational Ontology Integrating Objects and Processes [Version 1.0]**. Onto-med report, Research Group Ontologies in Medicine, Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig 2006.
16. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M: **The Sequence Ontology: a tool for the unification of genome annotations**. *Genome Biol* 2005, **6**(5), [http://dx.doi.org/10.1186/gb-2005-6-5-r44].
17. Courtot M, et al.: **The OWL of Biomedical Investigations**. In *OWLED, Volume 432 of CEUR Workshop Proceedings*. Edited by Dolbear C, Ruttenberg A, Sattler U, CEUR-WS.org 2008[http://www.bibsonomy.org/bibtex/27efede95690cf226301af45%f878b03b0/dblp].
18. Ashburner M, et al.: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**:25–29, [http://dx.doi.org/10.1038/75556].
19. Brachmann RJ: **What IS-A Is and Isn't: An Analysis of Taxonomic Links in Semantic Networks**. *IEEE Computer* 1983, **16**(10):30–36.
20. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C: **Relations in biomedical ontologies**. *Genome Biol* 2005, **6**(5), [http://dx.doi.org/10.1186/gb-2005-6-5-r46].
21. Rector A, Rogers J, Bittner T: **Granularity, scale and collectivity: When size does and does not matter**. *Journal of Biomedical Informatics* 2006, **39**(3):333–349, [http://dx.doi.org/10.1016/j.jbi.2005.08.010].
22. Simons PM: *Parts: a study in ontology*. Oxford University Press 1987.
23. Hoehndorf R, Oellrich A, Dumontier M, Kelso J, Rebholz-Schuhmann D, Herre H: **Relations as patterns: Bridging the gap between OBO and OWL**. *BMC Bioinformatics* 2010, **11**:441+.
24. Wood Z, Galton A: **A taxonomy of collective phenomena**. *Applied Ontology* 2009, **4**(3–4):267–292.
25. OWL Working Group W: *OWL 2 Web Ontology Language: Document Overview*. W3C Recommendation 27 October 2009. [Available at http://www.w3.org/TR/owl2-overview/].
26. Kim JD, Ohta T, Tateisi Y, Tsujii J: **GENIA corpus - a semantically annotated corpus for bio-textmining**. In *ISMB (Supplement of Bioinformatics)* 2003:180–182, [http://bioinformatics.oupjournals.org/cgi/content/abstract/1%9/suppl\1/i180?etoc].
27. Schulz S, Beisswanger E, Wermter J, Hahn U: **Towards an Upper-Level Ontology for Molecular Biology**. *AMIA Annu Symp Proc* 2006, **2006**.
28. Beisswanger E, Lee V, Kim J, Rebholz-Schuhmann D, Splendiani A, Dameron O, Schulz S, Hahn U: **Gene Regulation Ontology (GRO): Design Principles and Use Cases**. In *MIE* 2008:9–14.
29. Aranguren ME, Antezana E, Kuiper M, Stevens R: **Ontology Design Patterns for bio-ontologies: a case study on the Cell Cycle Ontology**. *BMC Bioinformatics* 2008, **9**(Suppl 5):S1+.