Published in final edited form as:
*Bioinformatics*. 2006 October 1; 22(19): 2421–2429. doi:10.1093/bioinformatics/btl405.

# Bio-Ontologies and Text: Bridging the Modeling Gap Between

**Carol Friedman**[1,*,§], **Tara Borlawsky**[1], **Lyudmila Shagina**[1], **Hongmei R Xing**[2], and **Yves A. Lussier**[1,3,§]

[1]Department of Biomedical Informatics, Columbia University, New York, USA, 10032

[2]Depts. Of Pathology and of Radiation Oncology, The University of Chicago .The University of Chicago, 5841 South Maryland Ave, Chicago, IL

[3]Department of Medicine Center for Biomedical Informatics, The University of Chicago .The University of Chicago, 5841 South Maryland Ave, Chicago, IL

## Abstract

**Motivation—**Natural language processing (NLP) techniques are increasingly being used in biology to automate the capture of new biological discoveries in text, which are being reported at a rapid rate. To facilitate the computational reuse and integration of information buried in unstructured text, we propose a schema that represents a comprehensive set of biological entities and relations as expressed in natural language. In addition, the schema connects different scales of biological information, and provides links from the textual information to existing ontologies, which are essential in biology for integration, organization, dissemination, and knowledge management of heterogeneous information. A comprehensive representation for otherwise heterogeneous datasets, such as the one proposed, are critical for advancing systems biology because they allow for acquisition and reuse of unprecedented volumes of diverse types of knowledge and information from text.

**Results—**A novel representational schema, PGschema, was developed that enables translation of information in textual narratives to a well-defined data structure comprising genotypic and phenotypic concepts from established ontologies along with modifiers and relationships. Initial evaluation for coverage of a selected set of entities showed that 85% of the information could be represented. Moreover, PGschema can be realized automatically in an XML format by using natural language techniques to process the text.

## 1 INTRODUCTION

New biological discoveries are being reported at an extremely rapid rate. This new information is found in diverse resources that encompass a broad array of journal articles and public databases associated with different sub-disciplines within biology and medicine. The integration of biological knowledge and information is recognized as a critical knowledge gap in science (Pennisi 2005), and as essential for the future of the field because dissemination and subsequent deployment of the knowledge by automated applications and by researchers who need to access and connect the diverse information is also recognized as critical (Gardner 2005;Gopalacharyulu.et. al. 2005). Additionally, a large quantity of biological information resides in unstructured or semi-structured textual databases, thus posing a frequent, yet special, category of integration problem that we address in this paper. While linguistic knowledge is computable using natural language processing (**NLP**), it does not allow for the same quality of inference as declarative knowledge. Thus, it is essential to translate linguistic data structures



*To whom correspondence should be addressed..
§These authors have contributed equally to the work

generated by NLP into ontology-anchored declarative datasets to acquire otherwise unattainable large-scale or cross-disciplinary inferences. There are several requirements for high throughput large-scale integration of textual information with biological knowledge: 1) natural language processing (NLP) methods that automatically acquire biomedical information occurring in unstructured text (Cohen and Hersh 2005; Hirschman et. al. 2005), 2) the existence of a comprehensive information model specifying the biological entities and relations as described in text (Gkoutos et. al. 2004), 3) the existence of ontologies or terminologies (Ashburner M et. al. 2000; Blake JA 2004) that specify and describe biological concepts, 4) methods, likely based on a biological information schema, allowing for translation of the data structures produced by NLP into those of structured and ontology-anchored databases, and 5) integration and knowledge management tools that are based on coded data associated with established databases (Cantor et. al. 2005). Therefore, to achieve reusability, it is critical that NLP systems that structure textual information also map the information to a representation that provides codes linking the information in text to established ontologies. Additionally, the representation must be rich enough to model the complex relationships that are typically described in text. Such a representation entails at least two levels of specification: 1) representation of the biomedical concepts via identifiers that correspond to existing ontologies or controlled terminologies, and 2) representation of salient contextual information and relations, such as information that modifies and connects the coded concepts because these are critical for accurate and fine grained representation of biological information. A substantial amount of work by numerous researchers has been devoted to the first level as it involves formal knowledge and reasoning within traditional ontologies. The second level, which is the focus of this paper, provides for fine-grained representation of information and relations, which is necessary for enabling expressiveness, such as that found in natural language, and also for enabling subsequent fine-grained retrieval of structured information that was extracted from text.

In this paper, we propose an ontology-anchored representational schema for biological information called Phenotype-Genotype Schema (PGschema), which is based on information found in the language of biological text. It represents individual concepts, modifiers of the concepts, and identifiers associated with external ontologies. Most importantly, it incorporates external ontological identifiers as building blocks in order to represent more complex and expressive relations. Thus, this schema is intended to utilize existing ontologies while serving as a bridge between natural language and the more formal bio-ontologies. The schema is designed so that it can be directly realized automatically using a NLP technology that generates a compatible form of XML output.

### 1.1 Natural Language Processing Systems

A number of NLP and text mining systems have been described that extract limited information from biological text. For example, there are many systems that recognize or identify the names of biomolecular entities (BNER) (Krauthammer M and Nenadic G 2004); Hirschman.et. al. 2005), while other systems extract interactions between biomolecular entities (Rzhetsky et. al. 2004; Hirschman.et. al. 2005), capture subcellular locations of proteins from text (Craven and Kumlien 1999), or capture the kinase, substrate, and residue associated with phosphorylation (Narayanaswamy et. al. 2005). These systems require a relatively straightforward representational model. For example, a BNER system may insert tags around the entities in text where the tags specify the corresponding semantic classes and possibly unique identifiers. Similarly, a system that captures interactions can represent an interaction as a triplet *interaction* $entity_1$ $entity_2$, where the entities are or are not necessarily coded. However, systems that capture more comprehensive informational relations generally do require representational schemas, particularly if convergence, completeness, and integration with many different systems are objectives. Since we are currently developing an NLP system called BioMedLEE,

which aims to capture a broad range of genotypic-phenotypic entities and relations, we require a schema to represent the extracted information. Furthermore, for interoperability purposes, the schema should be well-defined and use an established specification language so that other applications can access the information appropriately. The BioMedLEE NLP system that uses PGschema has been implemented and evaluated for use with an application called PhenoGO (Lussier et. al. 2006), providing a proof of concept that PGschema can be realized automatically. PhenoGO uses BioMedLEE to obtain information that augments gene-GO relations in the GO annotations database with additional context, such as cellular and other anatomical information. However, the focus of the work reported here is the representational schema and not the NLP component.

Two other efforts involving integration of NLP and ontologies are the Obol effort (Mungall 2004), and the GENIA effort (Kim 2003). Obol has a different focus from our work because the aim is to assist in ontological development. More specifically, because ontological terms are expressed using natural language, Obol uses NLP technology to process the ontological terms in order to discover unique computable definitions for them, to elicit relations between the elements composing the ontological terms, and to facilitate reasoning over the ontology. GENIA maintains an annotated corpus of biological entities, which substantially furthers the development of NLP systems. The entity types conform to a model consisting of substances and biological locations involved in protein interactions. The model has a semantic category 'Other' for entities such as disease, process, and phenotypic descriptions, which is our primary focus.

### 1.2 Ontologies

There are substantial efforts in the biological community for organizing biological concepts as controlled terminologies or ontologies (Ashburner M et. al. 2000; Blake JA 2004; Schulze-Kremer 1998; Stevens et. al. 2000; Stevens et. al. 2002), and for developing tools that provide interoperability among different ontologies (Bodenreider 2004; Cantor et. al. 2005) in order to support intra- and inter-operability among the different research communities. This is critical for the field because there are so many different groups working on the same model organism, different model organisms, or different scales of biology. Some integrative ontologies concerned with biomolecular entities are UniProt (Bairoch et. al. 2005), and UniGene (Wheeler et. al. 2005), while Gene Ontology (Ashburner M et. al. 2000) is concerned with biomolecular functions, processes, and subcellular components. Other ontologies are associated with phenotypic traits, such as mouse anatomy (MA) (Evsikov et. al. 2004), mammalian phenotype ontology (MP) (Smith et. al. 2005b), cell ontology (CL) (Bard et. al. 2005), the Unified Medical Language System (UMLS) (Lindberg et. al. 1993), and SNOMED (Spackman 2004). In general, these efforts involve specification of the individual concepts so that they have well-defined definitions, are associated with non-ambiguous unique identifiers, and are appropriately situated within a classification or part-whole hierarchy. The Open Biological Ontologies (OBO) (http://obo.sourceforge.net/) consortium hosts over 50 open source ontologies associated with phenotypic and biomolecular information. One of the OBO ontologies, called Phenotype, Attribute and Trait Ontology (PAtO) (Gkoutos et. al. 2004), is a general ontology for describing phenotypes that can be measured either quantitatively or qualitatively. What is significant about PAtO is that it is species-independent. PAtO actually consists of two components, where one is the model, and the other is the attribute ontology. It contains an Entity-Attribute-Value (EAV) representation where three ontological terms are linked together to form a description. The Entity component is the phenotype being described, and, most importantly, it can be associated with an ontology that is external to PAtO. In contrast to the entity component, the Attribute and Value components generally correspond to concepts internal to PAtO. There also has been work concerning the ontology of relations in biomedical ontologies (Smith et. al. 2005a). This work differs from the treatment of relations in PGschema

because in PGschema the relations are linguistically based and represent terms, such as *cause*, and *play a role in*, that connect different observations or events whereas the relations specified by Smith and colleagues provide consistent and formal ontological definitions. A more complete discussion of general issues concerning ontologies for biological concepts is found in (Schulze-Kremer 2002); Baker et. al. 1999; Stevens et. al. 2000), and a fuller discussion of issues associated with requirements for clinical terminologies can be found in (Chute et. al. 1999; Cimino et. al. 1994).

### 1.3 Representation Schemas for Biomedical Language

In addition to development of ontologies for individual concepts, there have been efforts in the clinical domain to model the complex clinical information associated with the language of patient documents. Models have been developed to represent information in specific medical domains, such as radiology (Bell DS et. al. 1994; Evans et. al. 1994; Friedman et. al. 1994b; Friedman et. al. 1994a; Rector et. al. 1995; Rocha and Huff 2001), nursing (Button et. al. 2001; Matney et. al. 2004), anatomy (Rosse et. al. 1998), and surgical procedures (Rodrigues et. al. 1997; Rossi et. al. 1996), as well as for the broad medical domains (Campbell K et. al. 1994; do Amaral et. al. 2000; Friedman et. al. 1999). The different models or schemas are represented using a variety of formalisms, such as frames (Minsky 1975), description logics (Cornet and Abu-Hanna 2005; Hartel et. al. 2005), conceptual graphs (Sowa 1984), and XML (Friedman et. al. 1999). Generally, these represent specific relations among concepts so that a clinical event or observation may be associated with multiple attributes and values denoting different types of informational qualifiers, such as negation, time, severity, frequency, body location, and descriptive information. These different types of modifiers are critical for automated applications that use structured information because they significantly affect accuracy during retrieval, and are needed to achieve highly precise retrieval results. Negation, uncertainty, and previous events occur frequently in the clinical documents, and therefore, an application that seeks to detect a current clinical condition must retrieve reports containing that condition and filter out ones that are negated, that have occurred in the part, or that have not been asserted. For example, in *rule out pneumonia*, the condition *pneumonia* is not being asserted, and therefore should not be retrieved. Similarly, anatomical and other qualifiers are also critical to retrieve when high accuracy and fine granularity is needed. For example, in *worsening left lower lobe pneumonia*, the lack of improvement of *pneumonia* may be important to capture along with the specific lobular location.

A clinical informational schema was developed for the MedLEE NLP system (Friedman et. al. 1994a), a natural language extraction and encoding system, which covers a broad range of clinical information (Friedman et. al. 1994a; Friedman et. al. 1999). Numerous evaluations have demonstrated that MedLEE performs similarly to medical experts (Friedman et. al. 1999; Hripcsak et. al. 1995; Knirsch et. al. 1998). A critical factor for achieving high performance was that retrieval of the information encoded by MedLEE was fine-grained due to the way the extracted information was modeled. The evaluation studies that were performed were designed for realistic clinical applications associated with decision support tasks, and they relied on queries to retrieve the structured output generated by MedLEE. What is significant about these queries is that they required complex medical logic, which included selecting and then filtering out cases based on clinical conditions along with various modifier combinations, such as certainty, time, anatomical locations, change, and other contextual modifiers. Most importantly, the ability to include modifiers was critical to achieving high performance.

Our schema, PGschema, is framed on that of MedLEE (Friedman et. al. 1999), but differs significantly from it in that PGschema is specifically designed to represent genotypic and phenotypic information, as well as compound relations and functions instead of clinical events.

There are many similarities between phenotypic information and clinical events, and therefore representational schemas for clinical information is highly relevant. For example, they each include anatomical, disease, morphological, and functional entities, many of which are associated with similar modifier types, such as degree, change, and certainty. PGschema is not an ontology, but a schema that represents compositional and contextual aspects of terms where the terms may be associated with ontological concepts in external ontologies. It is most similar to PAtO in that it represents observations and attributes that have values. PGschema differs from PAtO in several ways: 1) an observation may have many qualifiers that represent different types of information, 2) whenever possible, an attribute may be associated with codes from external ontologies, 3) an attribute may have nested attributes, providing a mechanism for representation of very complex information, 4) an observation may be a phenotype, biomolecular entity, relation, or function, 5) complex entities, such as functions and relations, are represented as having arguments that are also associated with directionality, 6) functions and relations may be nested, 7) an observation can be but does not have to be associated with an external code, and 8) the schema is based on information and relationships that occur in the language of biological text.

## 2 METHODS

Since the ability to formally represent ***all*** information that occurs in text is not currently possible, we modeled a broad but selective set of genotypic and phenotypic types of entities and relations as expressed in the literature. The model was developed iteratively using a sample of 50 abstracts selected from a corpus of 3,705 MEDLINE abstracts, where the corpus consisted of articles annotated for functional information by the Mouse Genomics Informatics group (MGI) (Blake JA.et. al. 2003). First, an initial schema was established using the MedLEE schema as a foundation because clinical information has many similarities to phenotypic information. However, certain entities, such as **diagnostic procedure**, **recommendation**, **laboratory test**, and **demographic information**, were removed because they were not applicable. Similarly, certain modifiers were also removed, such as **date** and **family history**. We then performed a manual analysis of the information in the sample corpus. Based on the sample and knowledge of biology, we revised the initial schema accordingly by determining the basic types of entities in the language of the biological text that were important to represent (e.g. gene, gene product, anatomy, process, disease, cell), and then the types of information that modify the basic entities. For example, in text, a cell may occur in the context of a mutated gene (e.g. *p53 −/− T cell*), a gene may occur with organism information (e.g. *mouse Ror2*), and a phenotypic trait may occur with severity, negation, and/or anatomical information (e.g. moderate memory deficit, absence of limbs, stiffness in joints).

After a revised design was established, the sample articles were analyzed again in order to see if it was possible to manually map relevant information in the text into the model. Several rounds of refinements were made based on results of the above manual mapping activity. Whenever relevant information could not be represented in the schema, it was revised accordingly if possible. Once the modeling of the basic entities and their modifiers was deemed satisfactory, modeling of the relations and functions was performed in a similar manner. However, in addition to modifiers, a mechanism for representing arguments of the relations and functions as well as their directionality was specified. For example, in *Dexamethasone induced cell death of T-cells,* the **function** *induce* is represented so that it has an agent argument (e.g. the **substance** *dexamethoasone*), and a target artument (e.g. the **process** *cell death of T-cells*). After several more rounds of analysis and refinement, we determined that the model was adequate for representing information captured automatically by an NLP system. We then modified the BioMedLEE NLP system so that it would automatically structure biological information in text in accordance with PGschema. BioMedLEE generates output in XML form that is compatible with the representational schema. A document type definition (DTD) was

created that specifies the entity types, their modifiers, the relations and functions in conformance with PGschema.

We performed an initial evaluation of PGschema for coverage. This consisted of assessing the completeness of the modifier relations associated with the various types of entities. For this initial effort, we choose the entities that were most important to represent for the NLP applications we were currently working on. These included the following types: **problem** (diseases, morphologies, symptoms, phenotypic descriptions), **process**, **body location**, **cell**, **organism,** and **biomolecular entity.** A set of sentences corresponding to each type was randomly selected for manual analysis. The set was obtained by first collecting a different set of abstracts than the ones used for establishing the schema. We used the set of gene-GO annotations recorded in the GO database for the human. In order to facilitate the process of collecting sets of sentences for each type of entity, we used BioMedLEE to process the abstracts and to obtain structured output. For each type of entity, a program was then used to select sentences containing a tag in the XML output that corresponded to the type of entity. For example, to select sentences containing a cell entity, sentences containing the tag **cell** in the XML output were chosen. Thus, the same sentence may be put into different sets depending on the tags it contained. Once the sentences associated with the specific entity types were collected into sets, all tags were removed so that the sets consisted of the original sentences. Thus, BioMedLEE was used only as a tool to identify sentences containing the type of entity to be analyzed. From each set, 100 sentences were randomly selected, and then the first 50 sentences were chosen for manual analysis. The remaining sentences in each set formed a reserve set. The expert performed the manual analysis by reading each sentence in each set, identifying term(s) associated with the corresponding entity type, finding the modifiers of those terms, determining their semantic types if possible, and finally determining whether they were included in PGschema. For training, the expert was given guidelines and a set of 25 sentences for each type. This helped us identify problems concerning the evaluation and identify areas where the guidelines needed to be revised. After the training session the guidelines that were established were used by the expert performing the analysis to help further consistency for the study.

## 3 RESULTS AND DISCUSSION

PGschema was developed representing a variety of information associated with biomolecular and phenotypic entities, modifiers, and relationships that are found in biological text. Simplified overviews of the entity and modifier types are shown in Tables 1 and 2, but the actual representation is an XML form, which can be generated automatically as a result of processing text. An example of a few specifications of the XML representation is shown in Figure 1 in the form of a document type definition (DTD). There are currently 27 types of entities or information that are represented. The complete DTD for PGschema is located in the Website: http://zellig.cpmc.columbia.edu/PGschema. Table 1 lists the entity types, provides examples of each type, and shows what types of modifiers each entity type can have. For example, the entity **ORG**(anism**)** may have an **AL** (allelic) modifier (*homogygous mice*), a temporal modifier (e.g. *newborn mice)*, a **STR**(***ain***) modifier corresponding to an organism strain (e.g. C57BL/6J *mice*), and a **MUTG** (mutated gene) modifier as in *p53 −/− mice*. Another entity type is **process**, which can have several modifiers including **temporal** (e.g. *embryonic stage development*), **AN** (e.g. *liver development, hepatocyte proliferation*), **change** (e.g. *increased proliferation*), **certainty** (*failure to develop*). The entity **cell** may have different modifiers than the other types of anatomical entities because it can have an allelic modifier (e.g. *wild-type fibroblast* cells), or specify a gene that has been modified (e.g. *Traf5 −/− cells*). The entity type **GGP** (gene_gproduct) is an artifact useful when it is not possible to determine whether an occurrence of an entity is a gene or gene product. This type of situation typically occurs when using NLP techniques to extract information.

Column 1 of Table 1 is used to group types for the convenience of specifying ones with similar modifiers, as shown in column 4. In addition, the full term for the abbreviation is specified in parentheses in column 1. Table 2 lists common types of modifiers **MD1** that were also grouped for convenience. Some types of entities can only occur as modifiers of other types because they do not correspond to independent observations or entities (**quantity**, **degree**, **certainty**). These are noted in Tables 1 and 2 by adding a single '*' following the name. The types without an '*' can occur in text as an observation or modifier (**disease, anatomy, gene)**. One type of modifier, named **code,** is different than the others. It does not occur in the text of the documents, but is used as metadata to associate identifiers of an entity with an external ontology. For phenotypes, the identifier may consist of 3 fields (e.g. **MP:0000351^increased cell proliferation**) where the first field specifies the applicable ontology, the second the identifier of the concept within the ontology, and the third the preferred or official name of the corresponding concept according to the ontology. In the above example, **MP** is an abbreviation representing the mammalian phenotype ontology. For genes, the identifier may have an additional fourth field, which specifies the taxonomic code of the organism.

More complex information is represented by the entities **FUN**(ction) (*inhibit, bind*) and **REL** (ation**)** (*correlate with, play role in*). These entities may be qualified by degree, change, certainty, temporal, and anatomical modifiers (e.g. high level of activation, decreased activation, not activated, expression in liver), but they also specify arguments with directionality or order. Therefore PGschema has a mechanism for specifying these phenomena.

An argument is different from a modifier because the meaning of the function or relation substantially depends on the arguments and their roles. An example is the sentence *Tenascin-C regulates cell proliferation*, where the function *regulate* is represented so that it has an argument *Tenascin-C* belonging to the class **GGP** which is the agent of *regulate***,** and an argument *cell proliferation,* which is a **process** that is the target. *Tenascin-C* is specified as an argument by adding a metadata tag **arg**, which has the value **agent** to the **GGP** element, and a metadata tag **arg** with the value **target** to the **process** element. Similarly, in *Tenascin-C plays a role in cell proliferation* the relation *play role in* would have two arguments, where the first argument would be the **GGP** element and the second argument would be the **process** element. This would be accomplished by adding a metadata tag **arg** to each element and assigning it a value **1** or **2** specifying the order of the argument in the text. A specific role, such as agent or target, is not assigned to the arguments of **relation** at this point because the role would depend on knowledge of the particular relation and post-processing or additional knowledge would be necessary to determine it. An example of the representation of relations, functions and arguments will be described below.

Although, Tables 1 and 2 show the entities in the schema in tabular form, the actual representation is an XML form that can be generated automatically by BioMedLEE when processing articles. Figure 1 illustrates examples of the DTD for several elements of PGschema.

The element **structured** is a child of the root element called **pgschema** (not shown in Figure 1). Note that the elements of **structured** are the entities in Table 1 that are not followed by a single '*'. These correspond to the primary types of entities or observations. Also note that the elements of all the entities except for **structured** are optional and may occur zero or more times. For example, **problem** has elements **arg**, **bodyloc**, **code**, and **process**, etc. These are nested structures and are considered modifiers or qualifiers of **problem**. The **v** attribute of **problem** is a string corresponding to a textual term that denotes that type of information. Note that the **change** entity type can modify the type **problem**, but that **change** itself can only be modified by **degree**, **certainty**, and **temporal** types of information.

Figure 2 illustrates a simplified form of the XML output obtained as a result of processing Hepatocytic proliferation was increased in livers of newborn C/EBPalpha knockout mice. Note that the XML output is consistent with Tables 1 and 2. In the XML form, the types are specified as tags and the instances as values of the tags. The primary observation for the information in the sample sentence is a **process** whose value is **proliferation**. In addition, it has several modifiers that are represented as nested elements. One modifier is an entity **cell** whose value is **hepatocyte**, which is linked to a code CL:0000182 from cell ontology, and a UMLS code UMLS:C0227525. Other modifiers of **proliferation** are a **change** entity **increase**, which is not linked to any code, a **body location** entity **liver**, which is linked to a code **MA:0000358** corresponding to mouse anatomy ontology, and to another code **MP:0000598** corresponding to the mammalian phenotype anatomy. The string following the code identifier is shown for readability. In addition, a code **MP:0000351** corresponding to **increased cell proliferation** has been specified, which is associated with the **proliferation** structure. This code is the most specific code found by BioMedLEE for the **process** structure. Note that the **bodyloc** tag includes nested modifiers. Thus, **organism** whose value is **mouse** modifies **liver;** similarly **newborn**, which corresponds to **temporal** information modifies **mouse** as does the gene_product **C/EBPalpha**, which itself has a modifier **knockout.**

Figure 3 illustrates a simplified XML output form generated as a result of processing the sentence *Tenascin-C regulates cell proliferation*, which contains a gene function *regulate*. Note that the highest level in the output is **function** with the value **regulate.** It has nested under it an argument **gproduct** with value **tenascin-C,** which also has a UMLS code. It also has another nested element called **arg** with the value **agent,** signifying that **gproduct** is the agent argument of the higher level function **regulate**. Similarly, the **process** with the value **proliferation** has a qualifier **arg** with value **target**, signifying it is the target argument of **regulate**.

There are several issues concerning PGschema that are important to note. The schema allows for some redundancies in representations in order to accommodate natural language. The issue of focus or different viewpoints arises frequently in natural language because the expressiveness of language incorporates such flexibility. Since our schema is based on relations as expressed in natural language, it has entity-modifier combinations where the entity and modifier types can be reversed. For example, according to Table 1, a **process**, which is a **PO** (phenotypic observation) can be qualified by another **PO**. Thus, when representing the information *abnormal development, development*, which is a process, could be considered the primary entity, and *abnormal*, which is a problem, its qualifier. However, in *abnormal in development*, the primary entity could be considered the problem *abnormal* and the qualifier *development*. There are two ways that this redundancy could be handled subsequent to natural language processing. One way would be to allow the redundancy to remain as is, and to formulate queries that retrieve the structured output so that the queries account for the different possible combinations. Another way would be to write transformations that map the NLP output to a uniform representation or to one that conforms to another ontology. For example, another representational system, such as PAtO, may view the appropriate representation of *hepatocyte proliferation* differently than the one shown in Figure 2. In the PAto representation, the primary observation would be *hepatocyte* and the qualifier *proliferation*. Additionally, the view where **cell** is modified by **liver,** which in turn is modified by **mouse** may also be considered incorrectly represented based on world-knowledge. By transforming the XML structure appropriately, the correct view could be obtained. A second issue concerning the schema is that it is permissive and allows combinations that are not likely, such as *embryonic diabetes mellitus*. The purpose of PGschema is to represent the general compositional and relational aspects of the various types of information in text and not to represent specific knowledge concerning individual concepts. Thus, while an ontology may not permit a concept such as *embryonic diabetes mellitus*, it would be permissible in PGschema for disease type information in general to be qualified by temporal type information. A third issue concerning

PGschema is that external ontologies are currently represented as identifiers, but are not integral to the model. It may be advantageous in the future to link them directly to the source ontologies using URLs, a method that would be more in keeping with the semantic Web.

Table 3 shows the results of the manual analysis for coverage of selected entity types in PGschema. Column 1 shows the type of entity; column 2 shows the number of modifiers that were judged to be covered in PGschema followed by the total number of modifiers found to correspond to that type; column 3 shows frequent types of modifiers, and column 4 includes examples modifiers that were not covered. The average coverage for all the types combined was 85%. Note that the lowest coverage, which was 67%, is associated with biomolecular entities. This reflects our focus, which is the phenotype, and that we have not concentrated on the representation of components of genes and gene products. Results for anatomical body locations (**BPT**) and **problem** indicate certain difficulties that occurred during the manual analysis. For example, *normal*, which occurred 5 times in the test set, was considered by the expert as 'not covered' but it actually was classified as a **descriptor** in the BioMedLEE lexicon, and therefore should have been considered as covered. The coverage for BPT would be 94% if the manual analysis were corrected.

A number of other issues arose in the manual analysis, reflecting the complexity of the task. Our experience demonstrated that the task required expertise in four disciplines: linguistics, biology, medicine, and ontology, and thus was very difficult. One type of problem associated with the analysis occurred because BioMedLEE output was used to collect sentences containing an occurrence of a particular type of entity. However, occasionally the BioMedLEE system tagged an entity as denoting an incorrect type due to ambiguity. For example, *backbone* was tagged as an anatomical entity but it actually occurred in the text as *peptide backbone of three proteins*. Since the study involved evaluating the coverage of the schema and not the performance of the NLP system, such a sentence was removed from the manual analysis set by the expert, and another sentence from the reserve set containing that entity type was selected. Thus, we ensured that the sets each contained exactly 50 sentences associated with the particular entity type being evaluated.

A second type of problem occurred because the expert had to determine the semantic categories of the terms that modified the entity being evaluated. The semantic classes of many of the modifier terms were clear cut, such as *limb*, *mouse*, *hepatocyte*, and *tumor*. But the semantic types of certain terms (e.g. *direct*, *specific*) were vague and difficult to ascertain. This difficulty was discovered during the training session for the manual analysis. This difficulty was partially addressed by using a pre-existing source of semantic knowledge as the reference standard. Thus, if the expert could not semantically classify a term, the BioMedLEE lexicon was used to obtain its semantic category. If the term was in the lexicon, the semantic class that was specified in the lexical entry was the one that the expert used as a basis for the analysis. If the term was not in the lexicon, it was considered "not covered" by the schema. We felt this approach was reasonable because the existence of a lexical entry for a term meant that it was already semantically categorized independently based on another knowledge source. In the future, it may be more appropriate to search for the term in other ontologies. However, if the term is not found in any biological knowledge source, the problem will still exist. Determining the semantic category occasionally led to errors in the analysis. For example, the expert did not correctly classify *normal*. In the future, if multiple experts perform the analysis, and resolve their differences, this type of error is likely to be reduced.

A third type of problem occurred because the expert performing the manual analysis had to determine whether a multi-word term was compositional in meaning and thus consisted of a entity and modifiers or if it was an atomic unit. This required knowledge of both biology and medicine. For example, *essential hypertension* would be considered an atomic unit in medicine

and would not be considered as denoting *hypertension* with a modifier *essential*. In contrast *moderate hypertension* should be considered to be compositional, and therefore as an entity *hypertension* with a modifier.

The fourth problem was that it was not always straightforward to determine which entity was the modifier and which entity was being modified because different interpretations or viewpoints were possible. According to linguistics, the focus of a noun phrase is typically the head noun, as discussed above. Thus, when analyzing *abnormal in development, abnormal* would be considered the observed entity, and the modifier would be the head noun *development* of the prepositional phrase. However there were exceptions, which seemed to occur whenever the head noun could not be a primary entity (e.g. it could only occur as a modifier). For example, in *increase in proliferation*, the head noun *increase* modifies *proliferation.*

Other limitations of our study were: only one expert performed the manual analysis, an evaluation of certain entity types only was performed, and PGschema was not complete. In future work we plan on expanding and refining the schema. For example, experimental methods are not represented, and temporal information could be represented using a finer granularity because there are many different aspects associated with temporal information, such as duration, frequency, developmental stages, and disease stages. Additionally, in future work we also plan on performing further evaluations.

## CONCLUSIONS

We have proposed a novel informational schema called PGschema, which represents phenotypic and genotypic entities, modifiers, and relationships. PGschema is significant in several ways: 1) it can be realized automatically using NLP techniques, 2) it bridges the gap between language and ontologies by providing compositional expressiveness similar to that found in natural language, while also linking to formal ontologies, which are needed for reasoning and specification of external declarative and world knowledge, 3) it is in the form of XML, which is textual and easy to read, and 4) it connects diverse biological scales of information. Evaluation for coverage of selected entities demonstrated that those entities were appropriately covered, but more work was needed. Moreover, we found that guidelines were critical for the manual analysis study, although the task still was very complex because it required expertise in biology, medicine, linguistics, and knowledge representation.

Rapid technological improvements of biomedical ontologies and natural language processing should lead to a profound transformation in the reuse of heterogeneous narrative information when it occurs in the form of curated and highly computational knowledge stored in specialized biomedical databases. Thus, the proposed schema should result in accelerated reuse of biomedical knowledge. Moreover, technological standardization of declarative knowledge and the semantic Web have profoundly accelerated the development cycles in computational semantics, resulting in ontology-anchored databases that could be automatically transformed with a common expressive information schema, such as the one proposed. As the gap between linguistic and declarative knowledge is bridged with highly expressive and computable information schemas, such schemas are poised to produce a paradigm shift. Indeed, comprehensive information models are likely to enable rapid large-scale computational analyses of unprecedented volumes of information and knowledge.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This work was supported in part by grants R01 LM07659 (CF), R01 LM08635 (CF), and1K22 LM008308-01(YL).

## REFERENCES

Ashburner M, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25–9. [PubMed: 10802651]

Bairoch A, et al. The Universal Protein Resource (UniProt). Nucleic Acids Res 2005;33:D154–D159. Database issue. [PubMed: 15608167]

Baker PG, et al. An ontology for bioinformatics applications. Bioinformatics 1999;15(6):510–520. [PubMed: 10383475]

Bard J, Rhee SY, Ashburner M. An ontology for cell types. Genome Biol 2005;6(2):R21. [PubMed: 15693950]

Bell DS, Pattison-Gordon E, Greenes RA. Experiments in concept modeling for radiographic image reports. J Am Med Inf Assoc 1994;1(2):249–259.

Blake JA. Bio-ontologies-fast and furious. Nature Biotechnology 2004;22(6):773–774.

Blake JA, et al. MGD: The Mouse Genome Database. Nucleic Acids Res 2003;31:193–195. [PubMed: 12519980]

Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32:D267–D270. Database issue. [PubMed: 14681409]

Button PS, et al. Challenges in the development and testing of a reference terminology model for nursing interventions. Medinfo 2001;10(Pt 1):176–180.

Campbell K, Das AK, Musen MA. A logical foundation for representation of clinical data. J Am Med Inf Assoc 1994;1(3):218–232.

Cantor, MN.; Sarkar, IN.; Bodenreider, O.; Lussier, YA. Genestrace: phenomic knowledge discovery via structured terminology; Pac.Symp.Biocomput.; 2005; p. 103-114.

Chute, CG.; Elkin, PL.; Sherertz, DD.; Tuttle, MS. Desiderata for a clinical terminology server; Proc.AMIA.Symp.; 1999; p. 42-46.

Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge Based approaches to the maintenance of a large controlled medical terminology. J Am Med Inf Assoc 1994;1:35–40.

Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief.Bioinform 2005;6(1): 57–71. [PubMed: 15826357]

Cornet R, Abu-Hanna A. Description logic-based methods for auditing frame-based medical terminological systems. Artif.Intell.Med 2005;34(3):201–217. [PubMed: 15994071]

Craven, M.; Kumlien, J. Constructing biological knowledge bases by extracting information from text sources; Proc.Int.Conf.Intell.Syst.Mol.Biol.; 1999; p. 77-86.

do Amaral, MB.; Roberts, A.; Rector, AL. NLP techniques associated with the OpenGALEN ontology for semi-automatic textual extraction of medical knowledge: abstracting and mapping equivalent linguistic and logical constructs; Proc.AMIA.Symp.; 2000; p. 76-80.

Evans DA, et al. Toward a medical concept representation language. J Am Med Inf Assoc 1994;1(3): 207–217.

Evsikov AV, et al. Systems biology of the 2-cell mouse embryo. Cytogenet.Genome Res 2004;105(2-4): 240–250. [PubMed: 15237213]

Friedman C, et al. A general natural language text processor for clinical radiology. JAMIA 1994a;1(2): 161–174. [PubMed: 7719797]

Friedman, C.; Cimino, JJ.; Johnson, SB. A conceptual model for clinical radiology reports; Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care; Safran, C. NY, McGraw Hill. 1994b. p. 829-833.

Friedman C, Hripcsak G, Shagina L, Hongfang Liu. Representing information in patient reports using natural language processing and the extensible markup language. J Am Med Inf Assoc 1999;6:76–87.

Gardner SP. Ontologies and semantic data integration. Drug Discov.Today 2005;10(14):1001–1007. [PubMed: 16023059]

Gkoutos, GV., et al. Building mouse phenotype ontologies; Pac.Symp.Biocomput; 2004; p. 178-189.

Gopalacharyulu PV, et al. Data integration and visualization system for enabling conceptual biology. Bioinformatics 2005;21(Suppl 1):i177–i185. [PubMed: 15961455]

Hartel FW, et al. Modeling a description logic vocabulary for cancer research. J.Biomed.Inform 2005;38 (2):114–129. [PubMed: 15797001]

Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of Bio-CreAtIvE: critical assessment of information extraction for biology. BMC.Bioinformatics 2005;6(Suppl 1):S1. [PubMed: 15960821]

Hripcsak G, et al. Unlocking clinical data from narrative reports. Ann.of Int.Med 1995;122(9):681–688. [PubMed: 7702231]

Knirsch CA, et al. Respiratory isolation of tuberculosis patients using clinical guidelines and an automated decision support system. Infection Control and Hospital Epidemiology 1998;19(2):94–100. [PubMed: 9510106]

Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus - semantically annotated corpus for bio-textmining. Bioinformatics 2003;19(Suppl 1):il80–82.

Krauthammer M, Nenadic G. Term identification in the biomedical literature. J Biomed.Inform. 2004 in press.

Lindberg DAB, Humphreys B, AT McCray. The Unified Medical Language System. Meth Inform Med 1993;32:281–291. [PubMed: 8412823]

Lussier, YA.; Borlawsky, T.; Rappaport, D.; Liu, Y.; Friedman, C. PhenoGO: Assigning phenotypic context to gene ontolog annotations with natural language processing; Pac Sym Biocomput; 2006; in press

Matney S, Dent C, Rocha RA. Development of a compositional terminology model for nursing orders. Int.J.Med.Inform 2004;73(7-8):625–630. [PubMed: 15246043]

Minsky, M. A framework for representing knowledge. In: P, Winston, editor. The Psychology of Computer Vision. McGraw-Hill; New York: 1975. p. 211-277.

Mungall CJ. Obol: integrating language and meaning in bio-ontologies. Comp Funct Genom 2004;5:509–520.

Narayanaswamy M, Ravikumar KE, Vijay-Shanker K. Beyond the clause: extraction of phosphorylation information from medline abstracts. Bioinformatics 2005;21(Suppl 1):i319–i327. [PubMed: 15961474]

Pennisi E. How will big pictures emerge from a sea of biological data? Science 2005;309(5731):94. [PubMed: 15994540]

Rector AL, Glowinski AJ, Nowlan WA, Rossi-Mori A. Medical-concept models and medical records: an approach based on GALEN and PEN&PAD. J.Am.Med.Inform.Assoc 1995;2(1):19–35. [PubMed: 7895133]

Rzhetsky A, et al. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J Biomed Inf 2004;37(1):43–53.

Rocha RA, Huff SM. Development of a template model to represent the information content of chest radiology reports. Medinfo 2001;10(Pt 1):251–255.

Rodrigues JM, et al. Galen-In-Use: an EU Project applied to the development of a new national coding system for surgical procedures: NCAM. Stud.Health Technol.Inform 1997;43(Pt B):897–901. [PubMed: 10179798]

Rosse, C.; Shapiro, LG.; Brinkley, JF. The digital anatomist foundational model: principles for defining and structuring its concept domain; Proc.AMIA.Symp.; 1998; p. 820-824.

Rossi; Mori, A.; Galeazzi, E.; Consorti, F. An ontological perspective on surgical procedures; Proc.AMIA.Annu.Fall.Symp.; 1996; p. 115-119.

Schulze-Kremer, S. Ontologies for molecular biology; Pac.Symp.Biocomput.; 1998; p. 695-706.

Schulze-Kremer S. Ontologies for molecular biology and bioinformatics. In Silico.Biol 2002;2(3):179–193. [PubMed: 12542404]

Smith B, et al. Relations in biomedical ontologies. Genome Biol 2005a;6(5):R46.1–R46.15. [PubMed: 15892874]

Smith CL, Goldsmith CA, Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol 2005b;6(1):R7. [PubMed: 15642099]

Sowa, J. Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley; Reading: 1984.

Spackman KA. SNOMED CT milestones: endorsements are added to already-impressive standards credentials. Healthc.Inform 2004;21(9):54, 56. [PubMed: 15457880]

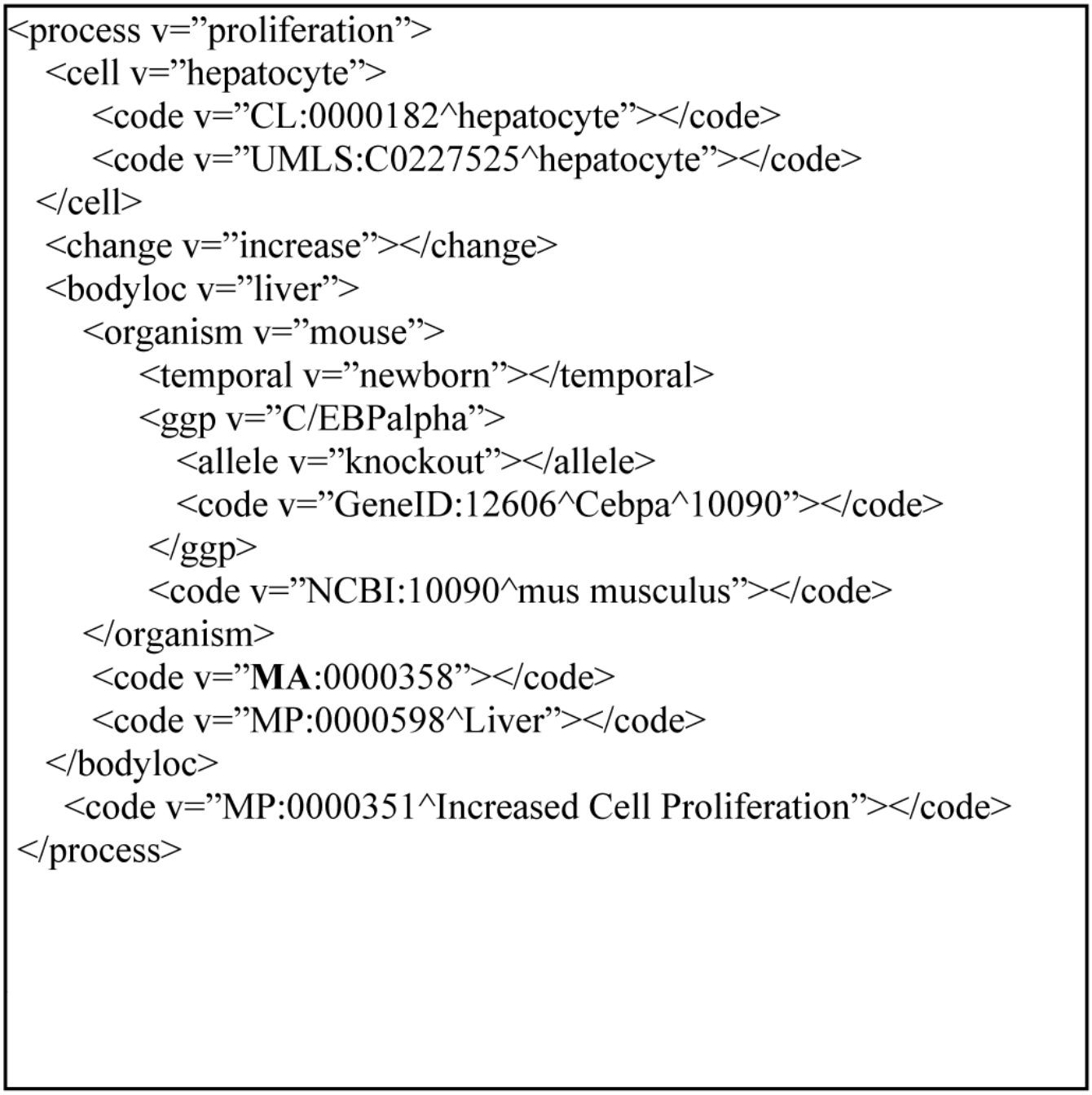
Stevens R, Goble C, Horrocks I, Bechhofer S. Building a bioinformatics ontology using OIL. IEEE Trans.Inf Technol.Biomed 2002;6(2):135–141. [PubMed: 12075668]

Stevens R, Goble CA, Bechhofer S. Ontology-based knowledge representation for bioinformatics. Brief.Bioinform 2000;1(4):398–414. [PubMed: 11465057]

Wheeler DL, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2005;33:D39–D45. Database issue. [PubMed: 15608222]

```
<!-- Denotes structured output-->
<!ELEMENT structured (process | bodyloc | cell | cellcomp|
 relation | gene_gproduct | gene| gene_product | function |
 descriptor | problem | substance | organism | bodymeas)+>
<!-- Information associated with change of state of observation-->
<!-- Ex: 'increased', 'no change', 'greatly increased'-->
<!ELEMENT change (degree|  certainty|  temporal)*>
<!ATTLIST change
   v CDATA #REQUIRED>
<!-- Information associated with degree-->
<!-- Examples: 'complete', 'mild', 'moderate', 'very severe' -->
<!ELEMENT degree (degree)*>
<!ATTLIST degree
   v CDATA #REQUIRED>
<!-- Information associated with disease, sign, symptom, functional
 or morphological abnormality-->
<!-- Ex: 'diabetes mellitus', 'fever', 'enlarged', 'dysfunctional'-->
<!ELEMENT problem (arg | bodyloc | code | process | cell |
 certainty | organism | change | descriptor | substance | temporal |
quantity | problem| measure | region | cellcomp |  gene_gproduct |
gene_product | degree |bodymeas | gene)*>
<!ATTLIST problem
   v CDATA #REQUIRED>
```

**Figure 1.**
Examples of document type description (dtd) for elements of PGschema



```
<process v="proliferation">
  <cell v="hepatocyte">
     <code v="CL:0000182^hepatocyte"></code>
     <code v="UMLS:C0227525^hepatocyte"></code>
  </cell>
  <change v="increase"></change>
  <bodyloc v="liver">
     <organism v="mouse">
        <temporal v="newborn"></temporal>
        <ggp v="C/EBPalpha">
           <allele v="knockout"></allele>
           <code v="GeneID:12606^Cebpa^10090"></code>
        </ggp>
        <code v="NCBI:10090^mus musculus"></code>
     </organism>
     <code v="MA:0000358"></code>
     <code v="MP:0000598^Liver"></code>
  </bodyloc>
  <code v="MP:0000351^Increased Cell Proliferation"></code>
</process>
```

**Figure 2.**
XML representation obtained as a result of processing the sentence Hepatocytic proliferation was increased in livers of newborn C/EBPalpha knockout mice.

```
<function v="regulate">
  <gproduct v="tenascin-C">
    <code v="UMLS:C0076088^tenascin-c"></code>
     <arg v="agent"></arg>
  </gene>
  <process v="proliferation">
    <cell v="cell"></cell>
     <code v="GO:0008283^cell proliferation"></code>
      <code v="UMLS:C0596290^cell proliferation"></code>
      <arg v="target"></arg>
   </process>
</function>
```

**Figure 3.**
XML representation obtained as a result of processing the sentence *Tenascin-C regulates cell proliferation*, which contains a function and arguments.

## Table 1

### Entities

| Type Group | Type | Examples | Modifiers |
|---|---|---|---|
| PO (phenotype observation) | Problem | *diabetes, defect Tourettes,fever,* | MD1, AN, PO, Q, REG, BE |
| | Process | *angiogenesis, walking* | MD1, PO, AN, BE |
| | DESC | *large, red* | MD1, AN, DESC, BE |
| | BMS | *weight, creatinine clearance* | MD1, AN |
| AN (anatomy) | BPT | *liver, epithelium, chest, arm* | ORG, AL, Arg, MUT, quantity, REG, BPT, Cell, Code, Ccomp, PO, temporal |
| | Cell | *Leukocyte, stem cell* | ORG, AL, Arg, MUT, quantity, BPT, Code, Ccomp, PO, temporal |
| | Ccomp | *Nucleus, cytoplasm* | Cell, ORG, MD1, PO, Code |
| (allele) | AL* | *Wild-type, heterozygote* | |
| BE (biomolecular entity) | Gene | *P53, cdk1* | ORG, AN,Arg, MUT, AL, Code |
| | GP | *Estrogen receptor* | ORG, AN, Arg, MUT, Code |
| | GGP | *Ambiguous term (may be gene or GP)* | ORG, AN, Arg, MUT, AL, Code |
| | MUTG | *P53 −/−* | AL, MUT (required) |
| (link to ontology) | Code* | *CL:0000182^hepatocyte* | |
| (function) | FUN** | *Bind, activate* | MD1 , Code |
| (mutation) | MUT* | *Mutant, deletion* | |
| (organism) | ORG | *Mouse, human* | STR, AL,Arg, Code, temporal, MUT, PO, MUTG |
| (region) | REG* | *Upper, right* | REG |
| (relation) | REL** | *Depend on, play role in,* | MD1 |
| (strain) | STR* | *C57BL/6J* | |
| (substance) | SUB | *Acetol* | Arg |

For convenience some entities appear as abbreviations and some were collected into groups. In column 1, full forms are noted in parentheses. Entities that have '*' following the name are modifier types only, and entities having '**' have arguments as well as modifiers. When the **Modifier** column is blank, it signifies that the corresponding entity has no modifiers. Legend for abbreviations not specified in table: **DESC**-descriptor, **BMS**-quantifiable measurement concerning an aspect of organism, **BPT**-an anatomical body part or location excluding cell, **Ccomp**-cell component; **GP**-gene product, **GGP**-gene or gene product, **MUTG** – a gene or gene product with a required mutation or allelic modifier.

**Table 2**

**Modifier Group 1 (MD1)**

| Type | Examples | Mods |
|---|---|---|
| Arg* | Agent, target, 1, 2 | |
| Certainty* | Possible, definite, no | Degree |
| Change* | Increased, unchanged | Certainty, degree, temporal |
| Code* | GO: 0008283 | |
| Degree* | Moderate, substantial | Degree |
| Measure* | 5% | |
| Quantity* | 1, many | |
| Temporal* | Adult, embryonic | |

**Table 3**

**Results of manual analysis for coverage**

| Entity Type | Coverage | Frequent Mods | Not covered |
|---|---|---|---|
| BE | 34/51 (67%) | Org | *Sequence, 2* |
| BPT | 44/52 (85%) | Org, AN, BE | *Normal** |
| Process | 64/69 (93%) | BE, disease | *Rat, intron* |
| Cell | 48/50 (96%) | Org, temporal | *Female* |
| Problem | 42/51 (83%) | AN | *t(15;15)* |
| Combined | 232/273 (85%) | | |