Literature-Based Knowledge Discovery from Relationship Associations Based on a DL Ontology Created from MeSH

Steven B. Kraines[1], Weisen Guo[1], Daisuke Hoshiyama[2], Takaki Makino[3], Haruo Mizutani[4], Yoshihiro Okuda[5], Yo Shidahara[5], and Toshihisa Takagi[6]

[1] Future Center Initiative, The University of Tokyo
5-1-5 Kashiwa-no-ha, Kashiwa, Chiba, 277-8563, Japan
sk@fc.u-tokyo.ac.jp
weisen.guo@gmail.com
[2] Springer Japan KK
3-8-1 Nishi-Kanda, Chiyoda-ku, Tokyo 101-0065, Japan
daisuke.hoshiyama@springer.com
[3] Institute of Industrial Sciences, University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8505 Japan
mak@sat.t.u-tokyo.ac.jp
[4] Department of Molecular and Cellular Biology,
Harvard University 52 Oxford St, Cambridge MA, 02138, USA
mizutani@fas.harvard.edu
[5] NalaPro Technologies, Inc.
4-12-16 Hongo, Bunkyo-ku, Tokyo 113-0033 Japan
{okuda,shidahara}@nalapro.com
[6] Department of Bioinformatics, School of Frontier Science, The University of Tokyo
5-1-5 Kashiwa-no-ha, Kashiwa, Chiba, 277-8568, Japan
tt@k.u-tokyo.ac.jp

Abstract. Literature-based knowledge discovery generates potential discoveries from associations between specific concepts that have been previously reported in the literature. However, because the associations are generally between individual concepts, the knowledge of specific relationships between those concepts is lost. A description logic (DL) ontology adds a set of logically defined relationship types, called properties, to a classification of concepts for a particular knowledge domain. Properties can represent specific relationships between instances of concepts used to describe the things studied by a particular researcher. These relationships form a "triple" consisting of a domain instance, a range instance, and the property specifying the way those instances are related. A "relationship association" is a pair of relationship triples where one of the instances from each relationship can be determined to be semantically equivalent. In this paper, we report our work to structure a subset of more than 1300 terms from the Medical Subject Headings (MeSH) controlled vocabulary into a DL ontology, and to use that DL ontology to create a corpus of A-Boxes, which we call "semantic statements", each of which describes one of 392 research articles that we selected from MEDLINE. Relationship associations were extracted from the corpus of semantic statements using a previously reported technique. Then, by making the assumption of the transitivity of association used in literature-based knowledge discovery, we generate hypothetical relationship associations by combining pairs of relationship

A. Fred et al. (Eds.): IC3K 2010, CCIS 272, pp. 87–106, 2013.

associations. We then evaluate the “interestingness” of those candidate knowledge discoveries from a life science perspective.

Keywords: Relationship associations, Association rules, Semantic relationships, Semantic matching, Semantic web, Ontology, Logical inference, Life sciences, Literature-based knowledge discovery.

1 Introduction

Potentially interesting and valuable scientific discoveries can be made simply by following associations, such as co-occurrence, between terms describing particular concepts or entities that have been previously reported in the literature. For example, in the 1980’s Don Swanson noted that many research articles mentioned “Raynaud's syndrome”, which results in discoloration of extremities, together with medical terms such as “blood viscosity”. Other articles mentioned the same medical terms together with “fish oil”. However, no articles mentioned “fish oil” and “Raynaud’s syndrome” together. He therefore proposed the new hypothesis that fish oil is effective for treating Raynaud’s syndrome [1]. That hypothesis was later confirmed experimentally.

Following this pioneering discovery, just by examining the current literature, of the relationship between Raynaud’s syndrome and fish oil, Swanson and other investigators made a few more interesting scientific discoveries by finding evidence in the existing literature for hitherto unreported associations between specific terms in the target domain [2], [3] ,[4], [5]. However, due to problems of polysemy and synonymy in natural language, the discovery process often produced a large number of false positives that had to be manually filtered out to find useful relationships.

In order to address the issues of term ambiguity in natural language, several research communities have established controlled vocabularies (CVs) that provide a one-to-one mapping between terms and concepts. One of the most well known CV is the Medical Subject Headings or MeSH terms. Currently, curators at the National Library of Medicine assign specific MeSH terms to research articles in life sciences that are stored in the MEDLINE repository. Because there is a controlled one-to-one matching between MeSH terms and the corresponding concepts from life sciences, term association can be replaced with actual concept association which should generate more semantically accurate discovery candidates. However, attempts to improve the accuracy of literature-based scientific discovery in life science by using the MeSH terms have been less successful than one might have hoped [6], [7], [8]. Some problems that have been noted include 1) the limited expressiveness of the MeSH vocabulary and 2) the inevitable mistakes in interpretation that are made by even the most careful curators.

Ontologies based on Description Logics (DL) extend the expressiveness of CVs in at least two important ways. First, in DL knowledge bases a distinction is made between ontology classes, which describe sets of semantically similar things, and instances of those classes, which represent actual things that are being described e.g. in a particular research article. Because ontology instances can be given arbitrary labels, this separation makes it possible to combine the precision of a CV with the flexibility of free text. For example, to represent a newly discovered protein “XYZ”, an instance of the class **Protein** could be created and labeled “XYZ” (throughout this paper, we show class labels in bold, property labels in italics, and instance labels in quotes). Also, the DL instantiation mechanism makes it possible to describe multiple instances of a particular class, each having different attributes, and how they interact in the particular study being described. For example, one could describe the interactions between an adolescent and adult mouse each having different attributes.

The second important contribution of DL ontologies is a means for expressing unambiguously the specific relationships between the instances that have been created to represent the key concepts and entities in the resource described. These relationships are expressed by using special terms, called properties, that connect a domain instance to a range instance, forming a semantic triple that consists of a domain instance, a range instance, and a connecting property expressing a specific directed relationship between the two instances. The properties can be assigned logical characteristics, such as transitivity. Then DL reasoners can be used to infer additional relationships between instances that are implied by the stated properties [9].

Unfortunately, current text mining techniques cannot accurately extract semantic relationships between concepts from natural language text due to the complexity and ambiguity of natural language [10], [11]. Furthermore, annotations by third party curators suffer both from mistakes in interpretation and also the limited scalability of a small group of curators to the rate of research article publication [12], [13].

A third alternative that is receiving interest recently is to get the original authors of research articles to create computer-readable descriptors of the objects of their research [14], [15]. Several initiatives have been made to get the scientific community to create wiki entries for biological entities such as proteins or to create structured digital abstracts for research articles [12], [16], [17]. The descriptors are made "computer-readable" by using specific templates to mitigate the problem of natural language ambiguity. This enables search engines, text mining systems and perhaps even human readers to more accurately establish the relationships between the entities that are described [18]. Furthermore, this approach has the additional benefit of putting the responsibility of correctly describing a research article in the hands of the author, who is clearly the person who best knows the main points of the article. However, even in structured digital abstracts or wiki entries, the granularity of expression for most of the descriptive information is still at the sentence or paragraph level [12]. Consequently, computers still need to make sense of the sentences in the delimited entries in the digital abstracts [19], [20], which is notoriously difficult due to the complexity and ambiguity of natural language [21], [22].

We suggest that by drawing on new techniques and standards for semantic representation of knowledge in a computer-interpretable form, such as description logics, it should be possible to enable human researchers to author descriptions of their shared knowledge that are not just "computer-readable", but actually "computer-understandable". By "computer-understandable", we mean that computers can reason with the semantics of the descriptors in reference to shared mental models or conceptualizations of the knowledge domain, e.g. the ontologies, and that they can infer new "facts" or "assertions" in the form of relationships between concepts and/or entities that are implied but not explicitly stated. In order to test this idea, we have developed a system, called EKOSS for Expert Knowledge Ontology-based Semantic Search, that enables researchers to author computer understandable descriptors in the form of "semantic statements", which define the specific relationships between entities and concepts described by a research article [23]. The system provides a set of intuitive authoring tools that guide researchers who may not be experts in formal knowledge representation through the process of creating a semantic statement to represent their research work based on a shared DL ontology.

Here, we describe the process in which we developed a DL ontology from a subset of the MeSH vocabulary, and we present some statistics of the use of the ontology to create a corpus of semantic statements for 392 research articles that were chosen to represent the researchers in life sciences at the University of Tokyo. We then present an algorithm for discovering hypotheses based on associations between specific relationships, called “relationship associations”. The relationship associations are mined from the semantic statements using a previously reported technique. We attempt to demonstrate the effectiveness of this approach by applying the algorithm to the corpus of semantic statements, and we discuss some of the hypothetical relationship associations that are discovered.

This paper is organized as follows. In Section 2, we describe the process of creating the UoT ontology by adding DL structure to a set of MeSH terms. In section 3, we describe the process of building the corpus of semantic statements based on the UoT ontology. In Section 4, we describe our algorithm for generating hypothetical relationship associations that represent new and potentially meaningful associations of specific relationships. In Section 5, we report the results of an experiment applying this algorithm to the corpus of semantic statements created previously. In Section 6, we review related work. In section 7, we finish with a discussion of the effectiveness of our approach and suggestions for future research.

2 Creating the UoT Ontology

In previous work to link a textbook used by undergraduate students at the University of Tokyo to research articles written by researchers at the same university, we have developed a DL ontology, called UoT for “University on Textbooks” [24]. The purpose of the DL ontology is to provide a formal knowledge representation language for positioning a research article in the “knowledge space” of the specific knowledge domain in a form that a computer can “understand” well enough to accurately determine the semantic similarity of different descriptors, e.g. in order to link the textbook and the articles. The UoT ontology was constructed by disambiguating the relationships between a subset of MeSH terms that were selected to cover the range both of the topics of the textbook and of 392 research articles selected from PubMed to represent the researchers in life sciences at the University of Tokyo.

Soualmia et al. reported initial efforts to add logical structure the entire MeSH CV [25]. However, they were only able to use a few heuristic methods to structure the terminology. Here, we focus our efforts on using various techniques to add logical structure to a relatively small subset of MeSH terms. This helps us to explore more thoroughly the possibilities for reframing the MeSH CV into a DL ontology that can function as a richly descriptive knowledge representation language.

We have implemented the ontology in OWL-DL [26]. The textbook is in Japanese, so we developed links from concepts in the UoT ontology to both English and Japanese terms. Thus the ontology also functions to link the natural languages of English and Japanese.

In the following subsections, we describe the details for the two main steps of the process of creating the UoT ontology: selecting the subset of MeSH terms to structure, and adding the logical structure to those MeSH terms using upper level classes and properties from other ontologies.

2.1 MeSH Term Selection Process

For the work reported here, we have focused on a small subset of the MeSH CV. Our hope is that the methods described in the next subsection could later be applied to the entire MeSH CV.

First, we used the 1997 version of the Japanese-English Life Science Dictionary (LSD) [27] and the UMLS (Unified Medical Language System) thesaurus to identify MeSH terms that match the 1078 Japanese terms in the index of the textbook. We identified a total of 883 MeSH terms (793 MeSH headings, 90 other MeSH terms) as possible matches with the terms in the index. Of those, 346 could be identified just by using the Japanese version of MeSH, 285 could be identified using the Japanese version of UMLS, and 252 were identified using the LSD. After manually filtering out false matches and choosing among candidate MeSH terms for each index term, we arrived at a list of 469 MeSH terms matching terms from the textbook index.

Simultaneously, we identified a subset of the 2024 MeSH terms (both major and minor) that had been assigned by PubMed to the 392 research articles selected for linking to the textbook. The subset to be added to the ontology was determined using the following conditions. First, only MeSH terms that were subsumed by one or more of the MeSH terms that had been mapped to the textbook were used – this eliminated about 700 terms. Second, the MeSH term must appear in at least 2 articles, which eliminated about 900 terms. We checked that at least one MeSH term from each article was included in the list of remaining terms.

The 297 MeSH terms that remained were added to the 469 MeSH terms from the textbook index, for a total of 766 MeSH terms mapped to either the textbook or the research articles. Next, we added all of the parent terms in the MeSH classification hierarchy, resulting in a grand total of 1360 MeSH terms. The total number of MeSH descriptors in 2011 was 26,142 [28], so our subset represents less than 10% of the entire MeSH CV. However, because our MeSH terms were selected based on the coverage of a general undergraduate textbook for life sciences and a set of research articles covering a wide range of topics, we believe that they represent much of the topic breadth of the MeSH CV and therefore the types of issues that would be involved in structuring the entire MeSH CV in a similar manner.

2.2 Adding Logical Structure to the MeSH Terms

In order to provide a higher-level logical structure for supporting logical reasoning to the set of 1360 MeSH terms, we added 45 upper level classes drawn from several popular upper-level ontologies. The most important upper level classes added to the basic MeSH categories are **Physical Objects**, **Phenomena** and **Abstract Classes** (we show classes in bold, properties in italics, and instances in quotes) – these correspond roughly to the ontological classifications used in the UMLS Semantic Network [29], ISO 15926 [30], SUMO [31] and GALEN [32].

We also added 51 abstract classes for defining roles of **Physical Objects** and **Phenomena**, 32 abstract classes for defining types of things, 189 classes for chemical elements, and 85 classes for general biological concepts such as chemical compounds and biological features. Most of these classes were reused from previous ontologies we had developed in other work [33]. The abstract classes were added to support faceted concept specification [34]. For example, the abstract class **Abnormal** can be used to specify the way in which a particular protein is reported to be abnormal in a research article. The other classes, such as the chemical elements and compounds, were added to increase the scope of the ontology. Thus the total number of classes in the ontology is 1762.

We next added a set of ontology properties (the OWL-DL objectProperty) for relating the concepts represented by the ontology classes. Based on reference to the upper-level ontologies described above, we compiled a list of 151 properties. These include domain specific properties such as *has homology* and *activates*, drawn from the UMLS semantic network [29] and GALEN [32], as well as more generic properties such as *has location* and *has part*, drawn mainly from SUMO [31] and the upper ontology based on ISO 15926 [30]. For ontologies that are not based on a description logic, such as the UMLS semantic network, some interpretation is required in order to determine how to represent the properties logically [35]. We have drawn on recent work in making these interpretations [36], [37], [38].

As discussed in the introduction, these properties play two important roles in making computer-aided knowledge sharing more intelligent. First, we can use them to define the specific types of relationships between the concepts expressed by classes in the UoT ontology. In particular, as described in the next paragraph, properties can be used to disambiguate "thesaurus type" subsumption relationships such as "related to", "narrower" and "broader". Second, the properties provide a means for connecting the specific entities in a semantic description of a research article or search query. In other words, they give us the "verbs" for making simple grammatical statements expressing knowledge in a computer understandable form.

The MeSH hierarchy is based on "thesaurus-type" subsumption relationships: the positioning of a term as a "child" of another term simply means that the child term somehow narrows the concept described by the parent term. The type of "narrowing" might be a set-theoretical *is a* relationship, but it can also be some other relationship such as composition, participation or location. For example, **Binding Sites, Antibody**, which is defined as "local surface sites on antibodies which react with antigen determinant sites on antigens", is positioned in the tree structure as a narrowing of **Antibodies**, **Binding Sites**, and **Antigen-Antibody Reactions** [28]. Clearly, the strict *is a* subsumption only holds for the relationship with **Binding Sites**. The relationship of **Binding Sites, Antibody** with **Antibodies** is compositional, and the relationship with **Antigen-Antibody Reactions** is locational.

We used the properties in the UoT ontology to disambiguate the subsumption relationships between MeSH terms in the MeSH term tree structure. For example, we define **Binding Sites, Antibody** in the DL ontology to be a subclass of **Binding Sites**, and we use existential restrictions (the OWL-DL "someValuesFrom" restriction) to specify that each instance of the class **Binding Sites, Antibody** is the part of an instance of the class **Antibodies** and the location of an instance of the class **Antigen-Antibody Reactions,** as shown in Figure 1.

Original version:

Binding Sites, Antibody *is a specialization of*
Binding Sites,
Antigen-Antibody Reactions, and
Antibodies

Disambiguated version:

Binding Sites, Antibody *is a kind of*
Binding Sites
that *is the location of* some instances of **Antigen-Antibody Reactions** and
that *is the structure part of* some instances of **Antibodies**

Fig. 1. Disambiguation of subsumption relationships in MeSH for the term "Binding Sites, Antibody"

In this way, we disambiguated the subsumption relationships for all of the MeSH terms that are subsumed by more than one parent term in the MeSH tree structure. The disambiguation step resulted in the definition of about 800 specific relationships between MeSH terms, such as the locational relationship between **Binding Sites, Antibody** and **Antigen-Antibody Reactions** described in the previous paragraph. In addition, we specified relationships between MeSH terms and the additional classes, such as roles and types, in order to add more structure to the ontology for supporting logical inference. For example, the class **Carrier Proteins** is specified in the ontology as being connected to some values of **Transport Roles** via the property *has role*.

A schematic diagram of the resultant knowledge model for the UoT ontology is shown in Figure 2. The figure shows the main upper classes that we have used to structure the MeSH terms, including **Process**, **Physical Object**, **Role** and **Characteristic**. In fact, the class **Process** is actually subsumed by a higher level class **Phenomena**, which subsumes a number of concepts from MeSH that we did not feel were actually processes in the context of the knowledge model that we constructed, such as **Acclimatization** and **Genetic Speciation**. **Regulation** is a class that reifies the regulation of a process by some physical object, such as an enzyme.

The main properties in the UoT ontology that can connect different upper level classes are also shown in Figure 2. Instances of the class **Process** can be connected to each other temporally with *occurs during, occurs before* and *occurs after*. They can also be connected compositionally with *process part of*, and by similarity with *has process homology*. Instances of **Physical Object** can be connected structurally with *has structure part*, *in contact* and *connects*, functionally with *interacts with* and *origin structure of*, and by similarity with h*as structure homology*. Both instances of **Physical Object** and **Process** can be specified as members of instances of the respective family classes, which can turn can be members of other family class instances.

Physical Object instances can be described as active participants of particular **Process** instances by using *regulating agent of* or *actor of* and as passive participants by using *transported agent of*, *consumed agent of*, *produced agent of*, and *unaffected agent of*. In addition, process regulation can be reified using instances of the class **Regulation** linked to regulating **Physical Objects** with *regulation actor of* and regulated processes with *regulation of*. Reification of process regulation makes it possible to specify the manner in which the regulation occurs, e.g. the participation of cofactor molecules. Locations at specific **Physical Objects** can be specified for all instances of **Process**, **Physical Object** and **Regulation** using the *has location* property. **Physical Objects** and **Processes** can also be targets of **Investigative Techniques** using *analysis object of*.

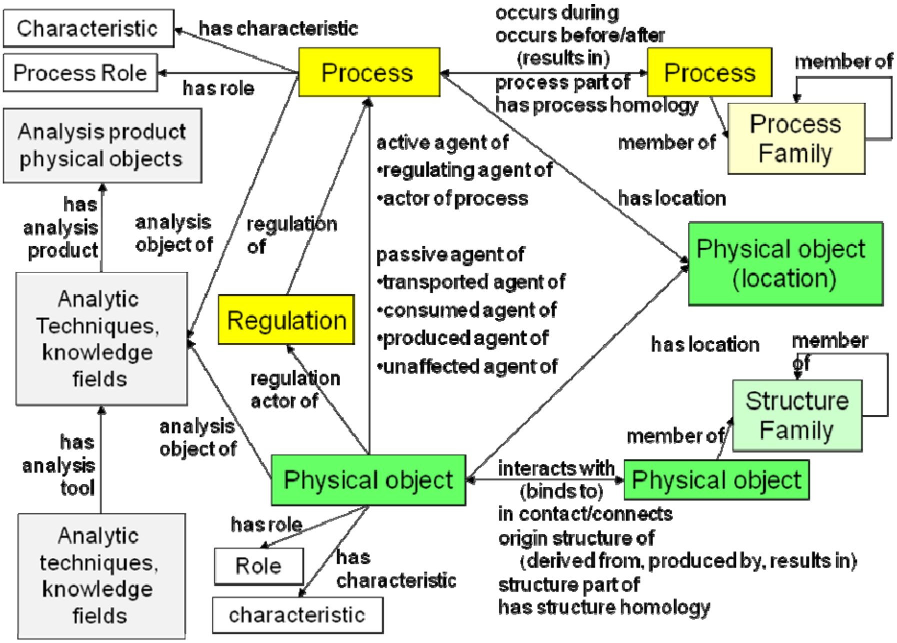

Fig. 2. Schematic diagram of the UoT ontology knowledge model showing top level classes and properties. Classes are shown in boxes. Physical objects, processes and analysis techniques are shown as green, yellow, and grey boxes. Directed arrows show properties that can connect a domain class to a range class. All boxes with the same name are interchangeable – e.g. all Physical Objects can be a passive agent of all Processes.

All properties are subsumed by the top level property *associated with*, which provides a way to describe an undetermined relationship between two instances. Therefore, any assertion of a specific relationship between two instances also produces an undetermined relationship usable in query matching. The logical characteristics and restrictions for all of the properties are specified using standard mechanisms provided by OWL-DL. In addition, as described in the Methods section, the OWL-DL restrictions “someValuesFrom” and “allValuesFrom” are used to define the allowable and descriptional usage of properties with specific classes.

3 Building the Corpus

We used the UoT ontology to create a corpus of semantic statements for a set of 392 research articles that were selected from MEDLINE for the UoT Project. The statements were created by curators having at least undergraduate degrees in life sciences, using the EKOSS (Expert Knowledge Ontology-based Semantic Search) system [23]. Each statement took about 3 to 4 hours to create and contains on average 26 class instances and 34 relationship triples of the form (domain instance, property, range instance). The entire corpus contains 13,283 semantic triples. An example of a complete semantic statement for the research article entitled "Oncogenic role of MPHOSPH1, a cancer-testis antigen specific to human bladder cancer" [39] is shown in Figure 3. Details on how to create semantic statements can be obtained at the EKOSS website at: www.ekoss.org.

The segment of the graph shown in Figure 3 that is circled corresponds to the following simple DL statement, where we use the notation of [9]:

- **Gene Expression**(colocalization)
- **Protein**(PRC1)
- *has produced agent*(colocalization, PRC1)

We can paraphrase this statement in natural language as follows:

"Colocalization" is a **Gene Expression** that *produces* a **Protein** called "PRC1."

As before, we show classes in bold font, properties in italics, and instance labels in quotes. The *is a* represents class instantiation, so "Colocalization" is an instance of the class **Gene Expression**.

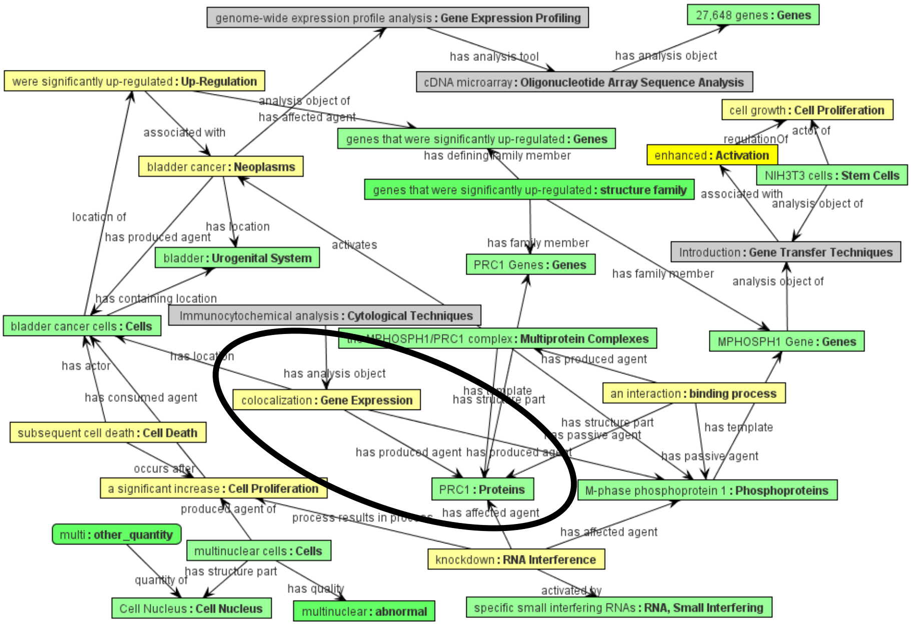
genome-wide expression profile analysis : Gene Expression Profiling
27,648 genes : Genes
has analysis tool
has analysis object
cDNA microarray : Oligonucleotide Array Sequence Analysis
were significantly up-regulated : Up-Regulation
analysis object of
has affected agent
cell growth : Cell Proliferation
associated with
genes that were significantly up-regulated : Genes
regulationOf
actor of
has defining family member
enhanced : Activation
bladder cancer : Neoplasms
NIH3T3 cells : Stem Cells
genes that were significantly up-regulated : structure family
associated with
analysis object of
location of
has location
has produced agent
activates
has family member
Introduction : Gene Transfer Techniques
bladder : Urogenital System
PRC1 Genes : Genes
has family member
analysis object of
has containing location
Immunocytochemical analysis : Cytological Techniques
bladder cancer cells : Cells
the MPHOSPH1/PRC1 complex : Multiprotein Complexes
MPHOSPH1 Gene : Genes
has location
has produced agent
has actor
has analysis object
has template
has structure part
an interaction : binding process
has consumed agent
colocalization : Gene Expression
subsequent cell death : Cell Death
has structure part
has passive agent
has template
has produced agent
has produced agent
has passive agent
occurs after
a significant increase : Cell Proliferation
PRC1 : Proteins
M-phase phosphoprotein 1 : Phosphoproteins
produced agent of
process results in process
has affected agent
multi : other_quantity
has affected agent
multinuclear cells : Cells
knockdown : RNA Interference
quantity of
has structure part
has quality
activated by
Cell Nucleus : Cell Nucleus
multinuclear : abnormal
specific small interfering RNAs : RNA, Small Interfering

Fig. 3. Graph view of the semantic statement based on [39]. Boxes show instances of classes from the domain ontology. The text to the left of the colon in a box is the instance label, and the bold text to the right of the colon is the class name of that instance. Arrows show properties expressing the asserted relationships between instances. Colors are as described in Fig 2. The semantic relationship described in the text is circled.

Of the 1762 classes in the ontology, 906 were used in the corpus at least once, and 210 classes were used 10 or more times. Of the 906 classes used, 751 were from the MeSH CV. Of the 210 commonly used classes, 156 were from the MeSH CV. The top 30 classes are listed in Table 1. The table shows that the most commonly used classes from the UoT ontology were high level classes that had been imported from other ontologies rather than from the MeSH CV. However, there were several MeSH terms that were used more than 40 times in the corpus.

Table 1. Usage counts for the 30 most often used classes in the UoT ontology. Source is the CV or ontology from which the term was drawn: "MeSH" for MeSH, "scinthuman" for the previous version of the ontology "scinthuman", "upper class" for a newly introduced upper level class, and "new concept" for a newly introduced domain level class.

Class Name	(source)	Usage Count
Regulation	(scinthuman)	535
characteristics	(upper class)	376
molecular processes	(scinthuman)	360
Activation	(scinthuman)	274
status	(upper class)	269
Genes	(MeSH)	269
Inhibition	(scinthuman)	208
Gene Expression	(MeSH)	159
molecule parts	(scinthuman)	157
organism processes	(scinthuman)	155
binding processes	(scinthuman)	148
Proteins	(MeSH)	140
Investigative Techniques	(MeSH)	118
quantity	(upper class)	112
cell processes	(scinthuman)	107
processes	(upper class)	106
Cells	(MeSH)	90
Humans	(MeSH)	85
Organic Chemicals	(MeSH)	83
absence	(new concept)	78
Diseases	(MeSH)	76
Mammals	(MeSH)	72
physical objects	(upper class)	66
Mice	(MeSH)	66
structure family	(scinthuman)	65
Neurons	(MeSH)	61
Enzymes	(MeSH)	49
Cellular Structures	(MeSH)	48
Metabolism	(MeSH)	46
Cells, Cultured	(MeSH)	45

Of the 151 properties in the ontology, 123 were used at least once, and 61 were used 50 or more times. The properties used most often were *regulation actor of* (932), *has regulation* (927), *has structure part* (834), *has location* (657), *has passive agent* (566), *characteristic of* (525), *structure part of* (500), and *has analysis object* (474).

Finally, we extracted associations between specific relationships of concepts using the technique that we reported previously [40], [41]. A relationship association is analogous to concept association, such as that evidenced by term co-occurrence in article titles, except that instead of being between singleton concepts, the association is between semantic relationships of the form “A has specific directed relationship X with B.” Therefore, a relationship association is a special kind of association rule that states “if concept A has relationship R1 with concept B, then it is likely that concept A has relationship R2 with concept C.”

4 Generating New Hypothetical Relationship Associations

In order to generate new and potentially interesting scientific discoveries from the relationship associations described in the previous section, we use a modification of the ABC open discovery model developed by Swanson and his colleagues [2], [6]. We first choose a small number of the most potentially “interesting” relationship associations that were extracted from the semantic statements. Then, for each of the relationship associations, irrespective of the “interestingness” criteria, we create all of the possible A-C relationship associations from the (A-B, B-C) pairs where the B triples match. Finally, we check that they are indeed “new” discoveries by searching for a match for each of the A-C relationship associations in the corpus of semantic statements. We consider an A-C relationship association that did not match with any of the semantic statements to be a potential discovery.

In the following subsections, we briefly describe each of the three main steps in generating potential knowledge discoveries: 1) matching the B triples of A-B and B-C relationship associations, 2) generating A-C relationship associations, and 3) searching for a match for each of the A-C relationship associations to the full set of semantic statements in the corpus. More detailed descriptions are given in [42].

4.1 Matching B Triples

The basic assumption in Swanson’s literature-based knowledge discovery model is that associations between concepts are transitive: if there is an association between concept A and B and between concept B and C, we can infer that there may be an association between concept A and C via the intermediary concept B. We consider two specific relationship triples to be associated if they are collocated in a particular semantic statement and there is an instance from each relationship triple that belongs to the same class, which we call the “connecting class”. The defined classes for the two instances do not have to be the same; we only need to show that they are semantically equivalent. Furthermore, unlike the original Swanson ABC model, relationship associations support directionality in the form of “if Triple 1 occurs in a semantic statement, then it is likely that Triple 2 will occur” [41]. Therefore, we also include the inverses of the relationship associations in the B-C set, which doubles the size of the B-C set.

4.2 Generating A-C Relationship Associations

Next, we create a new A-C relationship association from each pair of matching A-B and B-C relationship associations by connecting the non-matching triples in the two relationship associations, the A and C triples, via the connecting class in each relationship association. This means that in addition to having a matching B triple, the A-B and B-C relationship associations must also have matching connecting classes. We can think of this matching criterion as follows: a relationship association is essentially an association of two typed relationships that apply to one entity, which is represented by the connecting class. If two relationship associations can be found that describe two typed relationships for the same "connecting class" and one of those typed relationships are the same, we can create a new relationship association that associates the two non-matching typed relationships.

For example, consider the A-B relationship association No 3 in Table 2, "if a **neoplasm process** involves a **cell** then the **cell** is likely to be the actor of a **cell proliferation process**." The connecting class of this relationship association is **cell**, so this relationship association can only be matched to a B-C relationship association that also has **cell** (or a subclass or superclass of **cell**) as the connecting class. Therefore, the B-C relationship association "if a **bone marrow cell** is involved in a **neoplasm process**, then the **bone marrow cell** is likely to contain an **oncogene protein**" can match because it has **bone marrow cell** as the connecting class, which is a subclass of **cell**. However, the B-C relationship association "if a **bone marrow cell** is involved in a **neoplasm process**, then the **neoplasm process** is likely to involve an **oncogene protein**" cannot match because it has **neoplasm process** as the connecting class. In the case where the connecting class in one relationship association is a subclass of the connecting class of the other we create two new relationship associations using each class. Both of these are considered to be potential scientific discoveries.

4.3 Matching A-C Relationship Associations to the Semantic Statement Corpus

We look for matches for each of the A-C relationship associations in the entire semantic statement corpus using the standard semantic search algorithm based on RacerPro that we have reported in previous papers [23]. Note that the use of logic and rules makes it possible to find matches to relationship associations that are only implied at a semantic level because the reasoner can infer relationships between instances that are implied but not explicitly stated in the semantic statement. Any A-C relationship association that is found to match with at least one semantic statement is discarded from the set of knowledge discovery candidates.

5 Case Study

We have applied the process described above to the corpus of 392 semantic statements that we created using the UoT ontology. Because the corpus is small, our goal is only to demonstrate the potential effectiveness of the approach of generating hypotheses from relationship associations that could be realized with a larger set of semantic statements. The following subsections detail the application to the semantic statement set of each of the steps of the process for generating knowledge discovery candidates.

5.1 Selecting the A-B Set

For the A-B set, we chose five of the 984 relationship associations that met the relevance criteria for "interestingness" that we specified in our previous work: the first criterion is that the first triple must occur in no more than 40 semantic statements, and second criterion is that the probability that the association query occurs when the first triple occurs must be twice the probability that the second triple occurs when the connecting class occurs [41]. The five relationship associations are shown in Table 2.

Table 2. The five relationship associations we extracted previously [41]. Each triple is shown in the form "domain class | property | range class". The conditional triple is separated from the consequent triple using ">". The connecting class is shown in bold type.

No.	Relationship association
1	Flagella | has structure part | **Cytoplasmic Structures** > physical objects | interacts with | **Cytoplasmic Structures**
2	**Cytoplasmic Structures** | has structure part | Microtubules > Chlamydomonas | has structure part | **Cytoplasmic Structures**
3	**Cells** | passive agent of | Neoplasms > Cell Proliferation | has active agent | **Cells**
4	**Gene Expression** | has passive agent | Receptors, Cell Surface > **Gene Expression** | has location | Neurons
5	**organism parts** | structure part of | Drosophila > Growth and Development | has passive agent | **organism parts**

5.2 Creating the B-C Set

As discussed in the previous section, we want to use as many relationship associations as possible for the B-C set, even ones that might not be so interesting. Therefore, we used all 4821 of the relationship associations extracted from the corpus of semantic statements together with their inverses, for a total of 9642 B-C relationship associations to match with the five A-B relationship associations shown in Table 2.

5.3 Creating the Candidate Discovery A-C Set

We created all of the A-C relationship associations that results from pairing each of the 9642 B-C relationship associations with each of the five A-B relationship associations, both from pairs where the A-B relationship association is first and from pairs where the B-C relationship association is first. The number of A-C relationship associations generated for each A-B varies from 18 to 29, with an average of 24. Therefore, on average, just 0.25 percent of the B-C relationship associations match with each A-B relationship association. We suggest that the small number of B-C relationship associations matching with each A-B relationship association together with the relatively small variance in the matches for each A-B relationship association may be indicative of the diversity of the triples making up the B-C relationship associations because each different A-B relationship association matched with at least 18 B-C relationship associations.

5.4 Matching the A-C Relationship Associations to the Corpus of Semantic Statements

On average, 53% of the A-C relationship associations were found to already exist in the initial set of semantic statements, which disqualifies them as knowledge discovery candidates. The remaining A-C relationship associations are potential “discoveries”. However, as we noted earlier, the number of semantic statements is far too small to cover all of the semantic relationships that have been reported in the literature. We expect that with a larger corpus of semantic statements, many more of the A-C candidate relationship associations will be found to occur in the existing literature. In the following, we examine some of the knowledge discovery candidates that were generated.

One example of an A-C relationship association generated by the third A-B relationship association:

Cells | passive agent of | Neoplasms > Cell Proliferation | has active agent | **Cells**

and the B-C relationship association

Cell Proliferation | has active agent | **Cells, Cultured**
> Cell Differentiation | has passive agent | **Cells, Cultured**

that did not appear in any of the statements is:

Cells, Cultured | passive agent of | Neoplasms
> Cell Differentiation | has passive agent | **Cells, Cultured**

Here we express the relationship associations with the notation used in Table 2: “triple1 > triple2”, where each triple is expressed as “domain class | property | range class” and the connecting class is shown in bold type.

We can interpret this relationship association to mean that if a researcher happens to be studying cells involved in neoplasm processes, then it might be interesting for that researcher to look at the cell differentiation processes of those cells.

An example resulting from the fourth A-B relationship association:

Gene Expression | has passive agent | Receptors, Cell Surface
> **Gene Expression** | has location | Neurons

combined with the B-C relationship association:

Gene Expression | has location | Neurons
> **Gene Expression** | has passive agent | Carboxy-Lyases

is the hypothetical relationship association:

Gene Expression | has passive agent | Receptors, Cell Surface
> **Gene Expression** | has passive agent | Carboxy-Lyases

The hypothesis generated here is that if a researcher is studying gene expression involving cell surface receptors, it might be interesting to look for carboxy-lyase enzymes also involved in the gene expression.

An example resulting from the fifth A-B relationship association:

organism parts | structure part of | Drosophila
> Growth and Development | has passive agent | **organism parts**

combined with the B-C relationship association:

Growth and Development | has passive agent | **Synapses**
> Gene Expression | has location | **Synapses**

is the hypothetical relationship association:

Synapses | structure part of | Drosophila
> Gene Expression | has location | **Synapses**

The resulting hypothesis is that if a researcher is studying the synapses of *Drosophila*, it might be interesting to look at the gene expression located at those synapses.

We hope that these three examples have provided a clear demonstration of the type of scientific hypotheses that can be generated using the approach of literature-based knowledge discovery from relationship associations. With a larger corpus of semantic statements, it should be possible to extract more interesting potential discoveries of new relationship associations and to check more thoroughly that those relationship associations do not already occur in the published literature. We are currently exploring ways to increase the size of the semantic statement corpus, e.g. by integrating the statement authoring tools into the scientific paper publication process.

6 Related Work

The goal of the work presented in this paper is to discover new knowledge or hypotheses from the literature. Several previous research studies have attempted to attain this goal as we mentioned earlier. However, there are only a few studies that look at knowledge discovery about specific relationships between concepts.

Natarajan et al. (2006) used a combination of microarray experiments and NLP methods for extracting specific gene and protein relationships, such as inhibits and phosphorylates, from full-text research articles, in order to discover gene interactions linked to the protein S1P and the invasivity phenotype. However, their sentence-based text mining results had to be manually checked, and the problem of gene name polysemy was noted as being particularly difficult to resolve. They also did not appear to use any kind of inference.

Hristovski et al. used the natural language processing tool, BioMedLEE, to extract relationships between genotypic and phenotypic concepts in research articles, expressed in the form of "associated with change" [43]. They also used another NLP system, SemRep, to extract semantic relationships in the form of "treats". They then used the extracted relationships to construct a "discovery pattern", which they defined as a "set of conditions to be satisfied for the discovery of new relations between concepts." The conditions are given by combinations of relations between concepts that were automatically extracted from articles on MEDLINE. Finally, they conducted a novelty check to find discovery patterns that actually do not occur in the medical literature. However, their approach suffers from the low accuracy of automatically extracted semantic relationships and the limited number of relationship types that could be handled.

Another technique for extracting and interconnecting knowledge at the relationship level is automatic text summarization based on relationship extraction. The CLEF (clinical e-sciences framework) project aims to generate summaries or “chronicles” of patient medical histories based on relationships that are extracted from individual medical records [44]. The authors indicate that inference is used in assembling individual events into chronicles, but it is not clear if the inference is done at the level of specific relationships between events and entities in the records. MIAKT (Medical Imaging and Advanced Knowledge Technologies) is another system for automatically summarizing knowledge in medical examination reports that focuses on image annotations [45].

7 Discussion and Future Directions

Literature-based knowledge discovery is a technique that can be used to assist researchers in making scientific hypotheses that are well-based in the existing literature but have not been reported by any previous articles. A “discovery”, or more accurately a “potentially interesting hypothesis”, is generated in the form of an association between a pair of key terms in the literature that have not actually appeared together in any article but that have each occurred multiple times in the literature with the same intermediary concepts terms.

Existing techniques for literature-based knowledge discovery only consider associations between singleton concepts. Because most scientific knowledge takes the form of specific binary relationships between concepts rather than just unnamed associations, hypotheses that are generated from the implied associations of pairs of relationship triples, consisting of two concept instances and a typed and directed relationship between them, are potentially more interesting and meaningful.

A well formulated “heavy weight” ontology based on a description logic (DL) can function as a formalized knowledge representation language for expressing descriptions of knowledge resources that contain not only lists of key concepts, but also explicit assertions of specific relationships between pairs of concepts. Using a DL reasoner, one can even infer relationships that have not been explicitly stated but that are implied by the asserted relationships.

In order to test the effectiveness of such an ontology to realize more accurate literature-based knowledge discovery, we have constructed a DL ontology from a subset of the MeSH CV. The “thesaurus-type” relationships specified between terms in the MeSH CV were disambiguated manually by human experts with only a limited amount of automatic preprocessing based on identification of multiple superclasses and use of simple regular expressions. We then used that DL ontology to create a corpus of 392 semantic statements describing research articles in life sciences.

Next, we described an algorithm that we have developed for generating potential discoveries in the form of relationship associations that are implied by the extracted relationship associations but that do not appear in any of the semantic statements in the corpus. A relationship association is analogous to concept association, such as that evidenced by term co-occurrence in article titles, except that instead of being between singleton concepts, the association is between relationship triples. We applied the algorithm to the relationship associations extracted previously from the 392 semantic statements [40], [41]. Each semantic statement contains an average of 34 properties, and the corpus contains more than 13,000 semantic triples, which is comparable to the size of other major corpora used for testing knowledge discovery applications. In fact, the number of triples that are logically entailed is easily more than 100,000. However, even this corpus is too small to provide a good guarantee that a new relationship association has not actually been reported in the literature. Therefore, the aim of this case study has been to provide a demonstration of the kind of knowledge discoveries that could be possible if more semantic statements become available. We were able to find several implied relationship associations that at least appear to be somewhat novel and of interest in life sciences.

There are two major conditions for producing interesting knowledge discoveries using relationship associations. First, the classes and properties in the ontology must be sufficiently detailed to be able to express meaningful relationship associations. Second, the corpus of semantic statements must be large enough to check that a potential discovery has not already been reported in the literature. Unfortunately, we only have 392 semantic statements to work with, which is insufficient to satisfy the second condition. The EKOSS system is based on the idea that if the task of authoring the semantic statements could be distributed over the entire scientific community, the problem of scalability would be solved [12], [17]. However, here we have a typical "chicken and egg" problem: in order to convince scientists to make the effort to create the semantic statements, we must show their utility, but in order to show the utility of the semantic statements, we need a certain minimum number of statements to work with. Still, we believe that our corpus of 392 semantic statements will be sufficient to indicate the kind of discovery process that might be possible with a larger corpus of statements, thereby helping to "jump-start" a virtuous cycle of creating and applying semantic statements representing research articles.

There are several areas in which to continue this research. First, it would be useful to expand the DL ontology to cover a larger part of the MeSH CV. Regular expressions have been used to resolve relationships between terms in the Gene Ontology (GO) [46]. That was possible in part due to the particular concept labeling convention in GO together with the highly specific focus of GO on genes and gene products. Unfortunately, the MeSH vocabulary, with its broader concept coverage, is less amendable to this kind of approach. An alternative might be to use our manually disambiguated results to train a machine algorithm for disambiguating similar relationships in MeSH by using grammatical expressions in the term definitions as features for machine learning. For example, application of a part-of-speech tagger, named entity recognition, and grammatical analysis to the definition of "Binding Sites, Antibody" can identify that the term refers to "sites" that are spatially related to "antibodies" and that are participants of some "reaction" process involving antigens. This might be enough to enable a computer algorithm to disambiguate the relationships between some MeSH terms. However, applications to other terms may be less effective (for example "Antigen-Antibody Reactions" does not have any definition).

Other future tasks include 1) establishing additional measures of "interestingness" for the generated relationship associations that mirror the measures that we developed in our previous work and 2) building a larger corpus of semantic statements. In order to facilitate the process of creating semantic statements and reduce the cognitive overhead for the human authors, we are developing semi-automatic methods, including incorporation of natural language processing and machine learning algorithms into the semantic statement authoring tools. Finally, we would like to investigate the possibility for integrating the semantic statement authoring approach into the research article publication process in order to leverage the potential for network effects in the scientific community [12], [17], [47].

Acknowledgements. The authors thank the President's Office of the University of Tokyo for funding support. Hideo Ogimura assisted in the creation of the semantic statements in the UoT corpus.

References

1. Swanson, D.R.: Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine 30, 7–18 (1986)
2. Swanson, D.R.: Somatomedin C. and Arginine: Implicit connections between mutually isolated literatures. Perspectives in Biology and Medicine 33(2), 157–179 (1990)
3. Weeber, M., Kors, J.A., Mons, B.: Online tools to support literature-based discovery in the life sciences. Briefings in Bioinformatics 6(3), 277–286 (2005)
4. Racunas, S.A., Shah, N.H., Albert, I., Fedoroff, N.V.: HyBrow: a prototype system for computer-aided hypothesis evaluation. Biofinformatics 20(suppl. 1), i257–i264 (2004)
5. Natarajan, J., Berrar, D., Dubitzky, W., Hack, C., Zhang, Y., DeSesa, C., Van Brocklyn, J.R., Bremer, E.G.: Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. BMC Bioinformatics 7, 373 (2006)
6. Srinivasan, P.: Text Mining: Generating Hypotheses From MEDLINE. JASIST 55(5), 396–413 (2004)
7. van der Eijk, C.C., van Mulligen, E.M., Kors, J.A., Mons, B., van den Berg, J.: Constructing an associative concept space for literature-based discovery. JASIST 55(5), 436–444 (2004)
8. Yamamoto, Y., Takagi, T.: Biomedical knowledge navigation by literature clustering. Journal of Biomedical Informatics 40(2), 114–130 (2007)
9. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F.: The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, New York (2003)
10. Erhardt, R.A.-A., Schneider, R., Blaschke, C.: Status of text-mining techniques applied to biomedical text. Drug Discovery Today 11(7-8), 315–325 (2006)

11. Rinaldi, F., Schneider, G., Kaljurand, K., Hess, M., Romacker, M.: An environment for relation mining over richly annotated corpora: the case of GENIA. BMC Bioinformatics 7(suppl. 3), S3 (2006)
12. Ceol, A., Chatr-Aryamontri, A., Licata, L., Cesareni, G.: Linking Entries in Protein Interaction Database to Structured Text: the FEBS Letters Experiment. FEBS Letters 582(8), 1171–1177 (2008)
13. Rebholz-Schuhmann, D., Kirsch, H., Couto, F.: Facts from text–is text mining ready to deliver? PLoS Biol. 3(2), e65 (2005)
14. Gerstein, M., Seringhaus, M., Fields, S.: Structured digital abstract makes text mining easy. Nature 447, 142 (2007)
15. Seringhaus, M., Gerstein, M.: Manually structured digital abstracts: a scaffold for automatic text mining. FEBS Lett. 582, 1170 (2008)
16. Mons, B., et al.: Calling on a million minds for community annotation in WikiProteins. Genome Biol. 9(5), R89 (2008)
17. Pico, A.R., Kelder, T., van Iersel, M.P., Hanspers, K., Conklin, B.R., Evelo, C.: WikiPathways: Pathway Editing for the People. PLoS Biol. 6(6), e184+ (2008)
18. Hartley, J., Betts, L.: The effects of spacing and titles on judgments of the effectiveness of structured abstracts. JASIST 58(14), 2335–2340 (2007)
19. Cafarella, M.J., Re, C., Suciu, D., Etzioni, O.: Structured Querying of Web Text Data: A Technical Challenge. In: Proceedings of CIDR 2007 (2007)
20. O'donnell, M., Mellish, C., Oberlander, J., Knott, A.: ILEX: an architecture for a dynamic hypertext generation system. Nat. Lang. Eng. 7(3), 225–250 (2001)
21. Hunter, L., Cohen, K.B.: Biomedical language processing: what's beyond PubMed? Mol. Cell. 21, 589–594 (2006)
22. Natarajan, J., Berrar, D., Hack, C.J., Dublitzky, W.: Knowledge discovery in biology and biotechnology texts: A review of techniques, evaluation strategies, and applications. Critical Rev. in Biotech. 25, 31–52 (2005)
23. Kraines, S.B., Guo, W., Kemper, B., Nakamura, Y.: EKOSS: A Knowledge-User Centered Approach to Knowledge Sharing, Discovery, and Integration on the Semantic Web. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 833–846. Springer, Heidelberg (2006)
24. Kraines, S.B., Makino, T., Guo, W., Mizutani, H., Takagi, T.: Bridging the Knowledge Gap between Research and Education through Textbooks. In: Proc. 9th Intl Conference on Web Learning, Shanghai, China (2010)
25. Soualmia, L.F., Golbreich, C., Darmoni Soualmia, S.J.: Representing the MeSH in OWL: Towards a Semi-Automatic Migration. In: Proceedings of the KR 2004 Workshop on Formal Biomedical Knowledge Representation, Whistler, BC, Canada (2004)
26. OWL Web Ontology Language Overview, `http://www.w3.org/TR/2004/REC-owl-features-20040210`
27. Life Science Dictionary Project, `http://lsd.pharm.kyoto-u.ac.jp/en/service/weblsd/index.html`
28. U.S. National Library of Medicine, `http://www.nlm.nih.gov/pubs/factsheets/mesh.html`
29. McCray, A.T.: An upper-level ontology for the biomedical domain. Comparative and Functional Genomics 4, 80–84 (2003)
30. Batres, R., West, M., Leal, D., Price, D., Masaki, K., Shimada, Y., Fuchino, T., Naka, Y.: An upper ontology based on ISO 15926. Computers & Chemical Eng. 31, 519–534 (2007)
31. Niles, I., Pease, A.: Towards a Standard Upper Ontology. In: Welty, C., Smith, B. (eds.) Proc. 2nd Intl Conf. on Formal Ontology in Information Systems, Ogunquit, Maine (2001)

32. Rector, A., Bechhofer, S., Goble, C., Horrocks, I., Nowlan, W., Solomon, W.: The GRAIL concept modelling language for medical terminology. Artificial Intelligence in Medicine 9, 139–171 (1997)
33. Kraines, S.B., Iwasaki, W., Usuki, H., Yamamoto, Y.: A description logics ontology for biomolecular processes (poster). In: Bio-Ontologies SIG Workshop, Vienna, Austria (2007)
34. Yee, K.P., Swearingen, K., Li, K., Hearst, M.: Faceted metadata for image search and browsing. In: CHI 2003: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 401–408 (2003)
35. Kashyap, V., Borgila, A.: Representing the UMLS Semantic Network using OWL (Or "What's in a Semantic Web Link?"). In: Proceedings of the Second International Semantic Web Conference, Sanibel Island, Florida (2003).
36. Allemang, D., Hender, J.: Semantic Web for the Working Ontologist. Morgan Kaufmann, Burlington (2008)
37. Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., Wang, H., Wroe, C.: OWL Pizzas: Practical Experience of Teaching OWL-DL: Common Errors & Common Patterns. In: Motta, E., Shadbolt, N.R., Stutt, A., Gibbins, N. (eds.) EKAW 2004. LNCS (LNAI), vol. 3257, pp. 63–81. Springer, Heidelberg (2004)
38. Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in biomedical ontologies. Genome Biol. 6(5), R46 (2005)
39. Kanehira, M., Katagiri, T., Shimo, A., Takata, R., Shuin, T., Miki, T., Fujioka, T., Nakamura, Y.: Oncogenic role of MPHOSPH1, a cancer-testis antigen specific to human bladder cancer. Cancer Research 67, 3276–3285 (2007)
40. Guo, W., Kraines, S.B.: Discovering Relationship Associations in Life Sciences Using Ontology and Inference. In: Proceedings of 1st International Conference on Knowledge Discovery and Information Retrieval 2009, Madeira, Portugal, pp. 10–17 (2009)
41. Guo, W., Kraines, S.B.: Extracting Relationship Associations from Semantic Graphs in Life Sciences. In: Fred, A., Dietz, J.L.G., Liu, K., Filipe, J., et al. (eds.) IC3K 2009. CCIS, vol. 128, pp. 53–67. Springer, Heidelberg (2011)
42. Kraines, S.B., Guo, W., Hoshiyama, D., Mizutani, H., Takagi, T.: Generating Literature-Based Knowledge Discoveries in Life Sciences Using Relationship Associations. In: Proc. 2nd Intl. Conf. on Knowledge Discovery and Information Retrieval, Valencia, Spain (2010)
43. Hristovski, D., Friedman, C., Rindflesch, T.C., Peterlin, B.: Exploiting Semantic Relations for Literature-Based Discovery. In: AMIA Annu. Symp. Proc. 2006, pp. 349–353 (2006)
44. Taweel, A., Rector, A., Rogers, J.: A collaborative biomedical research system. Journal of Universal Computer Science 12, 80–98 (2006)
45. Bontcheva, K., Wilks, Y.: Automatic Report Generation from Ontologies: The MIAKT Approach. In: Meziane, F., Métais, E. (eds.) NLDB 2004. LNCS, vol. 3136, pp. 324–335. Springer, Heidelberg (2004)
46. Wroe, C.J., Stevens, R., Goble, C.A., Ashburner, M.: A methodology to migrate the Gene ontology to a description logic environment using DAML+OIL. In: Pacific Symposium on Biocomputing, vol. 8, pp. 624–635 (2003)
47. Berners-Lee, T., Hendler, J.: Publishing on the Semantic Web. Nature 410, 1023–1024 (2001)