

# An Ontology to Integrate Multiple Information Domains in the Patent System

Siddharth Taduri, Gloria T. Lau, Kincho H. Law  
Engineering Informatics Group  
Stanford University  
Stanford, CA, USA  
{staduri, glau, law}@stanford.edu

Hang Yu, Jay P. Kesan  
College of Law  
University of Illinois, Urbana-Champaign  
IL, USA  
[hangyu@illinois.edu](mailto:hangyu@illinois.edu), [kesan@illinois.edu](mailto:kesan@illinois.edu)

**Abstract**— In recent years, there has been an explosive growth in scientific and regulatory documents related to the patent system. Relevant information is siloed into many heterogeneous information domains making it very challenging to retrieve information across multiple domains. In this paper, we develop an ontology for the patent system to integrate information from the patent and court case domains. Through a use case erythropoietin, we demonstrate how this ontology can be used to enhance information retrieval across multiple domains.

**Keywords**—Ontology, Patent, Court cases, Information Retrieval, Knowledgebase

## I. INTRODUCTION

In recent years, there has been an explosive growth in scientific and regulatory documents. The advancement in IT and the extensive reliance on the internet has largely influenced initiatives such as Government 2.0, United States Patent and Trademark Office's (USPTO) full-text database and PubMed's comprehensive technical database. The impact falls directly on many sectors, organizations and institutions in a way that allows each of these independent entities to collaborate and cooperate. The current state of science and technology is however very distributed and deeply buried under the regulatory system. The information is maintained by completely independent entities, each of which functions heterogeneously resulting in many information silos. Comprehensive information cannot be found in any one single silo and it takes significant effort to gather the desired information from multiple sources.

In this paper, we focus on the patent system, which involves many diverse information silos. The patent system is a two stage system where the first stage includes the acquisition of patents, and the second includes their enforcement. In the acquisition phase, a patent application is prosecuted by the USPTO and finally issued or rejected based on the patent examiner's decision. The prosecution history is documented and is also known as the file wrapper for that issued patent or application. The various documents involved in the acquisition phase are the patent applications, file wrappers, issued patents and any form of prior art such as scientific publications. The enforcement stage of the patent system comes into play once the patent is issued. In case of infringement of patent claims, the infringer of a patent can be tried in court in a patent litigation. The enforcement stage can revisit the steps taken in

acquisition stage, and can invalidate an entire patent based on its findings. The documents involved in the enforcement stage include patent applications, issued patents, file wrappers, court cases, other forms of prior art including scientific publications, and appropriate chapters of the United States Code (U.S.C) and the Code of Federal Regulations (C.F.R). Clearly, the patent system involves information which is very heterogeneous and siloed into separate information buckets.

A significant effort needs to be taken in order to gather relevant information across all the diverse information silos. For example, a start-up company wanting to patent their technology in the field of Global Positioning System will want to search the patent databases, scientific publications and perform an infringement analysis which requires studying patent litigations. A patent examiner is faced with the challenge of thoroughly searching for any form of printed publication which may validate or invalidate a claim made by the patent applicant. In both cases, they are faced with a common challenge of searching through many heterogeneous information sources namely (1) patent documents; (2) court cases; (3) scientific publications; (4) file wrappers and (5) other regulatory documents. The sheer volume of these documents makes this an almost impossible task and the effects fall disproportionately on smaller organizations and individuals. A majority of this task is still performed manually requiring hours of expensive labor. Moreover, the data trapped in these information silos are heterogeneous at various levels. They are different in scope, in formats, in structure, in semantics and syntax, in interfaces available to the databases amongst many others.

In this paper, we propose an ontology for the patent system which attempts to provide a standardized formal representation of the information contained in the patent system. The goal is to clearly define the semantics expressed in the information domains and integrate the information from the various information silos to provide a unified knowledgebase. The knowledgebase will form the basis for many applications which will drastically reduce the time spent searching through the domains and manually co-relating them. For example, if an inventor or an organization is trying to gather information pertaining to a particular technology, and only has initial knowledge of certain companies involved with that technology,

the knowledgebase can be used to converge to a set of relevant information must faster than existing systems.

Specifically, we provide an ontology for the patent and the court case domains of the patent system. Other documents such as file wrappers and scientific publications will be included in future. We discuss in detail the current drawbacks and challenges associated with today's technologies, and provide a step by step methodology for developing an ontology as a potential solution to the problem. We choose to construct a use case in the bio domain which involves "erythropoietin", a hormone responsible for the production of red blood cells in living organisms. With the help of the use case, we will construct scenarios representing real-life situations faced by individuals and organizations.

Section II provides a background study on two information silos, namely patents and court cases, and the current state-of-art tools that allow us to access the information and related work in this area. Section III introduces the use case, and describes the structure of patents and court cases to help understand the differences and also the co-relation between the two document domains. Section IV describes the methodology followed to develop the ontology and Section V presents real-life scenarios, such as patent prior art research, to show the application of the ontology. Section VI concludes the paper by discussing some drawbacks and limitations of the study.

## II. BACKGROUND

In this section, we will review some of the challenges faced with respect to patent and court case research. We will also review relevant literature and the available state-of-the-art tools for information mining and integration of the information silos.

### A. Patents and Court Cases

There are currently over 7 million issued U.S Patents. In 2009 alone, 485,312 patent applications were filed with the USPTO [25]. In addition, there are over 40 different patent issuing authorities across the world, including the European Patent Office, Japanese and German Patent offices. Some of the online databases used to access legal information include the USPTO for patents, copyrights and trademarks, HeinOnline, LexisNexis and WestLaw for other IP related legal information [33]-[35]. Recently, Google and USPTO entered into a deal to make all USPTO products freely available online [23]. Thomson Innovation and Dialog LLC provide tools to help in information mining of patent documents and other scientific literature through services such as Delphion and Web of Science [36]. The Derwent World Patents Index (DWPI) is one of the largest patent databases with documents indexed from 41 patent-issuing authorities. There are 94 District Courts and one Court of Appeals (CAFC). PACER (Public Access to Court Electronic Records) is one electronic system to access databases for US Court cases [37]. Manually scanning each of these 95 databases is not a feasible option. Currently, PACER requires one to know the party name or the case number; in other words, it does not allow keyword-based search.

### B. Related Work

A variety of methods have been proposed for integrating diverse knowledge domains [14], [15], [16], [21]. One method suggests that a single ontology be defined, which integrates the semantics of all knowledge domains. A potential drawback of such an approach is its lack of scalability to a very large set of knowledge domains. Also, depending on the application, such a huge knowledgebase may be unnecessary and inefficient. Alternatives include having separate ontologies representing each knowledge domain, and integrating them through either the application directly, or via a top level ontology. Several ontology development methods have been proposed and are widely used [16], [17], [19]-[22].

There are other Information Retrieval (IR) techniques for both patents and case law which are not ontology-based [2], [6], [8], [12]. Due to the large amounts of unstructured information available online, such techniques need to be made more efficient. Several information retrieval methods have made use of domain specific ontologies such as bio ontologies to capture domain knowledge and in turn enhance retrieval [1], [3], [9], [10], [13]. Specifically related to the domain of patent documents, the PATEXPART project has developed an ontology for the patent domain which focuses on the European patent system [4], [5], [11]. However, these methodologies focus on a single domain of knowledge, and hence are not applicable to a larger set of heterogeneous domains. To address the issue of IR across a diverse set of information domains, firstly there is a need to construct an universal ontology, or to construct individual ontologies and subsequently integrate them. Secondly, the IR techniques need to be modified to collectively mine information from all domains.

## III. INTRODUCING THE USE CASE

We choose to demonstrate the working of the ontology by constructing a use case – erythropoietin. Erythropoietin is a hormone responsible for the production of red blood cells in the body through a process known as erythropoiesis. The deficiency of red blood cells results in lower haemoglobin levels than normal, which is also known as anemia. The synthetic production of the hormone erythropoietin has been a crucial discovery for the treatment of severe diseases such as anemia. Amgen Inc. was the first pharmaceutical company to commercialize the production of the synthetic version of erythropoietin in the form of Epogen. They own five core patents related to the production of erythropoietin, namely U.S. Patents 5,547,349, 5,618,698, 5,621,080, 5,756,349 and 5,955,422. We followed the forward and backward citations of the 5 core patents and identified 135 closely related U.S. patents.

BioPortal is a source for bio domain knowledge with a collection of over 150 bio-ontologies [24]. A search for an exact match of the term "erythropoietin" returned around 11 ontologies. From these ontologies, we identified 43 closely related concepts to erythropoietin, by extracting related concepts such as the synonyms, children, parents and grandparents of "erythropoietin". For each of the 43 extracted concepts including erythropoietin, we downloaded the top 50-100 patents to create a database of 1150 U.S. patents. The

database of 1150 patents contains patents both related and unrelated to the use case and acts as our test database. The 135 related patents identified will serve as the gold standard for any performance tests.

Our corpus also includes around 30 U.S. federal court cases which involve Amgen and the 5 core patents spanning from the late 1980s to date. Furthermore, these patents collectively cite over 3000 scientific publications. In addition, each patent document comes with a corresponding file wrapper. All put together, the use case provides us with documents which span multiple domains representative of the problem we seek to solve.

### A. Structure of the Documents

In our use case, we focus on patents issued in the U.S. which are publicly available on the USPTO website. The full-text documents (1973-present) are available for download as HTML files. Although no specific web service is provided by the USPTO, a simple ‘wget’ script is written to automatically fetch the required patent documents from the server. The downloaded patent documents have a standard structure which clearly distinguishes the various fields of interest such as the title, inventor, assignee etc. We exploit this structure and developed a script to parse out all the information that pertains to us.

We downloaded court cases from the LexisNexis database by searching for erythropoietin in the federal court database. The search resulted in 30 court cases which are closely related to the use case. It is difficult to automate the download of court cases since none of the systems mentioned in Section II.B provide an API or a web service to do so. Since the structure of court cases is not as well defined as patent documents, parsing these documents is more of a challenge. The important fields, such as title, abstract, claims etc. are thus extracted using a carefully coded script.

### B. Cross-Referencing

Patent and court litigations are highly inter-related. As shown in Fig. 1, the documents cross-reference information contained in one another. For example, the court case directly refers to a number of patents. The cross-referencing can be very crucial when relevancy has to be established between patents and court cases. Current systems take little or no advantage of this form of cross-referencing between the heterogeneous information silos to enhance multi-domain IR. When developing the ontology for the patent system, such factors must be taken into account.

## IV. METHODOLOGY

There is a large community working towards the development of ontologies, knowledge representation and engineering. Several ontology development methodologies have been proposed and implemented over the years [16], [21]. We reviewed some of the methodologies which are most applicable to the development of our patent system ontology [17], [19], [20]. In general, the development of ontologies consists of several steps starting from the conceptualization of the domain, defining the properties inter-relating the defined

classes, instantiating the classes with physical objects and the verification of the constructed ontology. In their paper Ontology 101, Noy and McGuinness state that ontology development is essentially an iterative approach where the ontology evolves to satisfy the requirements of the application it is being designed for [17]. We follow the 101 development methodology to (1) define the scope and the application of the ontology; (2) conceptualize our domain and build a hierarchy of classes; (3) define properties and relations on each of the classes; and (4) instantiate the classes with physical documents. This section goes over the development of the patent ontology and reviews appropriate literature.

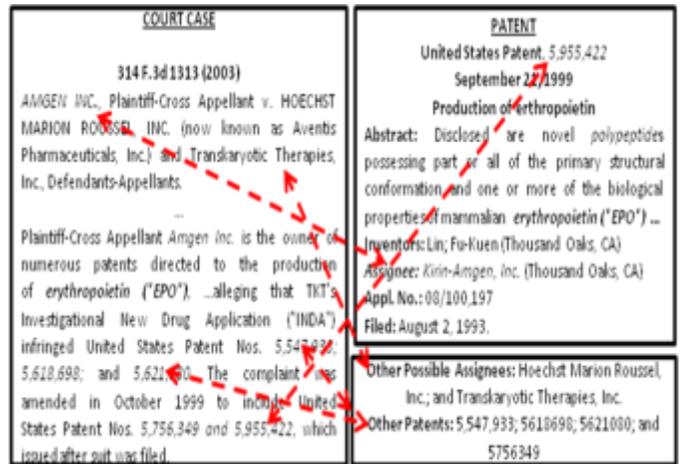


Figure 1. Cross-Referencing Between Patents and Court Cases.

### A. Defining Scope of the Ontology

Gruninger and Fox suggested that a set of competency questions be developed; these are questions that the ontology is expected to answer [22]. Developing these questions not only helps define the scope of our ontology but also allows us to verify the power and competency of the ontology both throughout and after the development phase. The main application for which we are developing the patent ontology is to integrate multiple heterogeneous domains in the patent system. Keeping this primary goal in mind, we will define a set of competency questions which confine to a single domain such as patents, and also span multiple domains. In Section V.B, we will build scenarios for information mining across patents and court cases and demonstrate how the ontology performs. The competency questions in no way limit the applications of this ontology to information retrieval alone, rather they are examples of questions the ontology must be capable of answering at a minimum.

#### 1) Patent Domain:

- Return all patent documents which contain the keyword “erythropoietin” in the “claims”
- Return all the patent documents which contain the keyword “erythropoietin”, at least 3 claims, issued prior to “a date” and assigned to “OrganizationA”

#### 2) Court Case Domain:

- Return all court cases which contain the keyword – “erythropoietin”

- Return all court cases which involve “OrganizationA” either as the plaintiff, defendant of both, and from the court “courtA”

### 3) Multi-domain:

- Return all patents which contain the keyword – “erythropoietin” in the “claims”, which has been challenged in the courts at least once.
- Return the number of times a patent by inventor “InventorA” has been challenged in the court.

The questions can get more complex depending on the requirement of the user. The results of one query can be re-filtered with more constraints too:

- Return all court cases with the term “erythropoietin”. From these court cases, return the patents involved. From these patents, follow the backward and forward citations to identify more important patents.

Notice that the last bullet point is the method we followed to identify the 5 core patents assigned to Amgen, and the 135 patents related to those as mentioned in Section III.

### B. Selection of Specification Language and Tool

It is important to determine the specification language in which the ontology will follow. Several specification languages have evolved over the years including frame based languages such as F-Logic and OIL, and descriptive logic based languages such as DARPA Agent Markup Language and Ontology Inference Layer (DAML+OIL), Resource Description Framework (RDF) and Web Ontology Language (OWL) [26]-[28]. Description Logic based languages were developed to overcome the lack of formal logic-based semantics in frame based languages. Several factors need to be considered when choosing a specification language for the ontology which include expressivity, semantics, reasoning capabilities, availability of tools, re-use and personal preference. RDF is a very widely used language to conceptualize domains. OWL is a W3C recommendation which is built on top of the semantics of RDF to provide higher expressivity levels. These higher expression levels allow us to define disjoint classes, sameAs or different individuals and class property restrictions amongst others.

Several tools have also been developed for the construction and modeling of ontologies such as Protégé and Chimaera [31], [32]. Protégé is very widely used in the ontology engineering community. Protégé supports both OWL and RDF, and provides useful features and plugins allowing us to query and visualize the ontology. Taking into account the above mentioned considerations, we choose OWL as the specification language and Protégé-3.4 as our development tool for the patent system ontology.

### C. Classes and Object Properties

In order to define classes, we must first enlist the terminology pertaining to the patent and court case documents such as title, inventor, plaintiff, judge etc. After enlisting the terms pertaining to the documents, they are then grouped into

classes representing a group of entities, for example, a class of all inventors, or a class of all court litigations. To develop the class hierarchy, we follow a bottom-up approach where first, all the classes and their corresponding definitions are generated. We then start to group classes which fall under the same parent class and build up to the highest level of abstract concept. Fig. 2 shows the asserted class hierarchy.

Relations between classes are defined through properties or slots. We define both annotation properties for the classes which contain the class definition, and object properties relating or mapping one class to another. Table 1 gives a detailed description of the object properties, their domains and ranges as defined in the ontology. These properties not only relate entities within the patent domain or the court case domain, but also cross-relate entities from both domains. The ontology can be queried based on any of the classes shown in Fig. 2 or the object properties shown in Table 1.

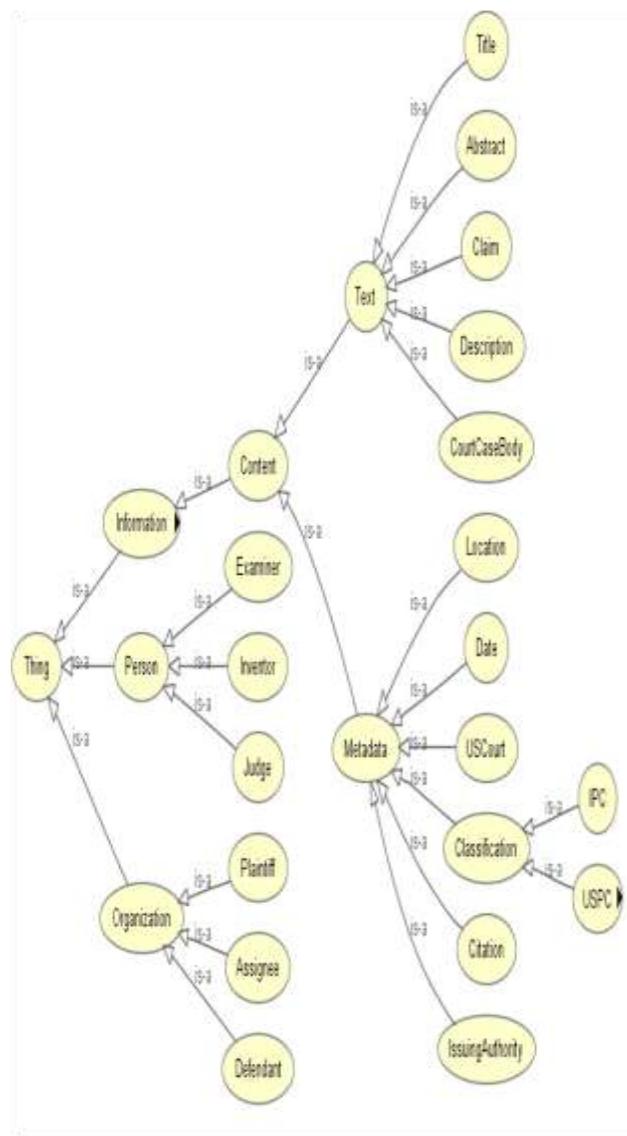


Figure 2. Class Hierarchy.

TABLE I. OBJECT PROPERTIES, DOMAINS AND RANGES

| Object Property | Domain               | Range            |
|-----------------|----------------------|------------------|
| claimedIn       | Claim                | Patent           |
| Examined        | Examiner             | Patent           |
| has Abstract    | Patent               | Abstract         |
| has Assignee    | Patent               | Assignee         |
| hasBody         | CourtCase            | CourtCaseBody    |
| hasCitation     | Patent               | Patent           |
| hasClaim        | Patent               | Claim            |
| hasDefendant    | CourtCase            | Defendant        |
| hasDescription  | Patent               | Description      |
| hasExaminer     | Patent               | Examiner         |
| hasInventor     | Patent               | Inventor         |
| hasIPCClass     | Patent               | ICPCClass        |
| hasUSClass      | Patent               | USClass          |
| Invented        | Inventor             | Patent           |
| isLocated       | Inventor or Assignee | Location         |
| issuedBy        | Patent               | IssuingAuthority |
| patentsInvolved | CourtCase            | Patent           |
| precededBy      | CourtCase            | Judge            |
| hasPlaintiff    | CourtCase            | Plaintiff        |
| hasTitle        | Patent or CourtCase  | Title            |

#### D. Instantiation

Instantiating the classes with actual physical entities is a crucial task in the development of the ontology. In this stage, the actual contents of the documents are parsed and the corresponding classes are populated. After instantiation, the ontology can now perform the role of a knowledgebase of which instances and classes can be queried along the defined object properties shown in Table 1.

The first step to creating instances is to parse the patent and court case documents. We have to account for the differences in the structure, format and language used in these documents. We built a regular expression based parser which parses through each document and extracts individuals pertaining to classes including Title, Abstract, Claim, Inventor etc. Once a particular entity, say ‘‘InventorA’’ is extracted, we first assign it to its class – Inventor, followed by any relations to other entities. To avoid confusion, the term ‘entity’ in the above context is used to refer to owl:Individuals and not classes or object properties as the definition of an entity would suggest.

For classes such as Inventor or Plaintiff, we use the extracted text itself as the name of the inventor or plaintiff. For example, an inventor InventorA will have the name ‘InventorA’ in the knowledgebase. For classes which include text content such as the Abstract, Description or CourtCaseBody, we extract the text and store it as a string under a custom defined annotation property ‘‘j.0:ext’’. A naming convention is followed for these classes of the form <patent#\_class\_#>. An example for this would be

5955422\_claim\_2, which means this extracted entity is the second claim of U.S. patent 5955422. To conform to the ontology syntax, entity names are not allowed to have spaces or certain special characters. Dealing with court cases is more difficult than parsing patent documents. However, with the use of well crafted regular expressions, the entities belonging to the classes Title, Plaintiff, Defendant, Judge and CourtCaseBody etc. were all successfully extracted.

We used the protégé-owl-3.4 Java API to automate the instantiation of the ontology. The parser feeds its output to the code responsible to populate the ontology. A total of 884 patent documents and 30 court cases are represented using 28 classes, 23 object properties, and 23942 individuals. However, some issues may arise during the instantiation process. Examples of such problems are when two different individuals, say two different inventors have the same name, or when a single company changes names over the years. To address these issues, we may have to modify our naming conventions to accommodate two different individuals of the same name.

## V. RESULTS

### A. Querying the Ontology

Over the years, several query languages for ontologies have been developed. Some of the commonly used ontology query languages include RDF Query Language (RDQL), SPARQL Protocol and RDF Query Language (SPARQL) and Semantic Web Rule Language (SWRL) [29], [30]. RDQL and SPARQL are query languages for RDF, which are syntactically very similar to Structured Query Language (SQL), a language commonly used to query relational databases. SPARQL is an extension of RDQL which provides many additional clauses overcoming some of the shortcomings of RDQL such as CONSTRUCT, DESCRIBE and ORDER BY clauses amongst many others. A detailed description of the SPARQL query language is available in the W3C documentation [29]. Although SPARQL is a query language for RDF, since OWL is built over the RDF semantics, SPARQL can be used to query OWL ontologies as well. SWRL is a language that combines OWL with Rule Markup Language (RuleML). Both languages have their own advantages while also suffering some drawbacks. Since SPARQL is a language for RDF, it is not possible to query OWL anonymous classes. However, it works well to query individuals instantiated in the named classes in an OWL ontology. SPARQL does not have an understanding of the semantics underneath, although this can be overcome by querying an inferred version of the ontology and by using CONSTRUCT queries. SWRL on the other hand may need many additional classes and clauses in-built in the OWL ontology for facilitating the same. While it is possible to construct queries which can be answered by one language and not the other, the simplicity and ease of use of SPARQL, which is in-built in the OWL API, has encouraged us to use SPARQL to query our OWL ontology.

### B. Use Case Scenario: Find important patents and court cases

In this section, we will answer the competency questions through a use case scenario. A patent prior art research is a

very time consuming task due to the sheer volume of information available online. A prior art search is required both in the acquisition phase and the enforcement phase of the patent system. A patent examiner may want to do a prior art research in order to examine a patent application; an inventor may need to determine the patentability of an invention; a startup company may need to carry out infringement analysis; a major company may want to stop others from infringing their patents and so on. In all cases, one would follow a general framework which involves identifying the important prior art, identifying the important court cases and expanding from them iteratively in various dimensions until a satisfiable point is reached. Fig. 3 shows a general patent prior art research [38]. In this section, we will simulate the patent prior art research by constructing a string of questions to query the ontology. Patent prior art can be any printed publication in the form of patents, scientific publications and even a PhD thesis. Since our knowledgebase currently includes patent documents and court cases, we will only consider patent documents and ignore other forms of prior art.

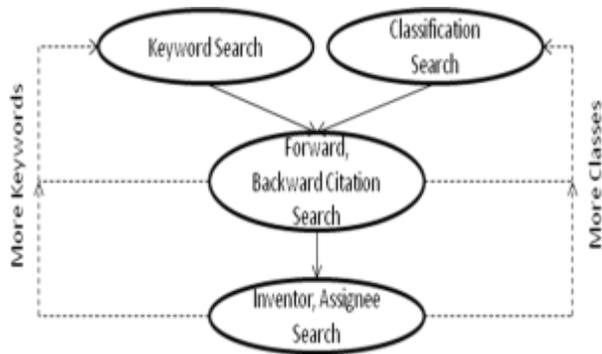


Figure 3. General Patent Prior Art Search.

Generally, the first step in patent prior art research is to use a keyword search. We will assume one has limited knowledge of what keywords to use at the beginning. Considering the volume of patent documents, a search for a single keyword could result in several thousands of patents. At this point one would realize that words can be used in different contexts meaning entirely different things, which result in irrelevant patents. By adding more keywords and restrictions such as fields to search (title, claims etc. instead of the entire document) the size of the search results will tend to be more manageable. It is now possible to scan the abstracts of some patents and identify the important classes that the inventions fall under. Searching for the keywords under those specific classes will result in more patent documents which may or may not be relevant. After identifying some relevant patents, the next step would be to follow the forward and backward citations, study the patents of the most relevant inventor and assignee to get a more concise and relevant result set. At every stage, new keywords can be added and these steps can be repeated until the result set starts to converge. The search is then independently applied to the application database.

A search for the term erythropoietin on USPTO results in over 7,000 patents. For a user who is unfamiliar with the terminology, or for a user who is unsure how to filter these

documents, it is a very daunting task to browse through the 7000+ documents that are returned. We will add another dimension to the patent prior art search in the form of court cases. Patents which have been involved in court cases have an obvious importance and can provide a good starting point for conducting the patent prior art search. We serve two purposes by adding this extra dimension (a) search through two heterogeneous information silos; and (b) provide a good starting point for searching patent prior art.

When dealing with heterogeneous information silos such as these, it is very difficult to independently search each silo and develop a set of documents which are related and span these silos. A higher layer of abstraction will provide a view into both the silos facilitating cross-referencing of information. This additional information regarding one set of documents is derived from a completely different set of documents.

*a) Search for all court cases containing the term "erythropoietin"*

The query shown in Fig. 4 performs the required search operation. To perform this search, first the description bodies of the court cases are retrieved. We use the FILTER REGEX clause to search the extracted text via the j.o:ext property to only extract those court cases which contain the term "erythropoietin". In this case, all 30 court cases are returned.

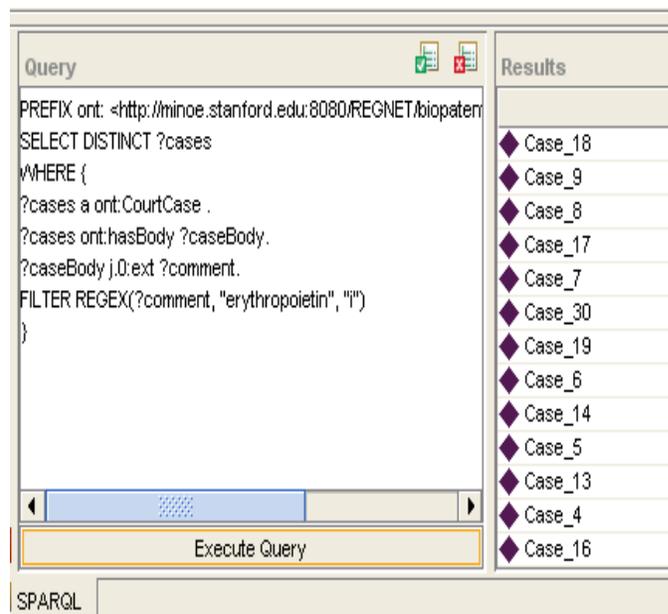


Figure 4. All Court Cases Containing – "erythropoietin".

*b) Enlist the patents involved in these court cases*

The query in Fig. 5 requests for all the patents which have been involved in these 30 court cases. The size of the patent list is around 17. It must be noted that not all 17 court cases may be present in our database of 1150 patents and hence during our instantiation process, no further information about these patents is updated in the knowledgebase. However, an ideal situation would assume all the patents are instantiated in the knowledgebase.

| Query   | Results  |
|---|--|
| <pre> PREFIX ont: &lt;http://minoe.stanford.edu:8080/REGNE SELECT DISTINCT ?patents WHERE{ ?cases a ont:CourtCase. ?cases ont:hasBody ?caseBody. ?caseBody j.0:ext ?comment. FILTER REGEX(?comment, "erythropoietin", "i"). ?cases ont:patentsInvolved ?patents. } </pre> | <ul style="list-style-type: none"> <li>◆ #5618698</li> <li>◆ #5547933</li> <li>◆ #5955422</li> <li>◆ #4377513</li> <li>◆ #5621080</li> <li>◆ #5756349</li> <li>◆ #4047496</li> <li>◆ #4667016</li> <li>◆ #3623712</li> <li>◆ #4703008</li> <li>◆ #5441868</li> <li>◆ #5641670</li> </ul> |

Figure 5. Patents Involved in the Retrieved Court Cases.

| Query   | Results   |
|---|---|
| <pre> PREFIX ont: &lt;http://minoe.stanford.edu:8080/REGNE SELECT DISTINCT ?assignee WHERE{ ?cases a ont:CourtCase. ?cases ont:hasBody ?caseBody. ?caseBody j.0:ext ?comment. FILTER REGEX(?comment, "erythropoietin", "i"). ?cases ont:patentsInvolved ?patents. ?patents ont:hasAssignee ?assignee } </pre> | <ul style="list-style-type: none"> <li>◆ Kirin_Amgen_Inc</li> <li>◆ Hayashibara_Ken</li> <li>◆ Ashida_Shin</li> <li>◆ Amgen_Inc</li> <li>◆ Kiren_Amgen_Inc</li> <li>◆ Genetics_Institute_Inc</li> <li>◆ Board_of_Trustees_of</li> </ul> |

Figure 8. Identified Assignees from the Selected Patents.

c) Identify the U.S. class, inventors and assignees of these patents

For all the patents which are available in the knowledgebase, we identify the inventors, the assignees and the U.S. classes the patents fall under. Figs. 6, 7 and 8 show the results of this search. In each case, we add another query triplet that questions individuals along the hasUSClass, hasInventor and hasAssignee properties on the patents returned by the query in Fig. 5. By removing the distinct clause, it is possible to get an estimate of which of these results are the most occurring. This can give a sense of the more important U.S. classes, or assignees and so on.

d) Search the forward and backward citations of these patents

In this step, the backward U.S. patent citations are extracted for each of the 17 patents returned by the query in Fig. 5. Many of these patents can have overlapping backward citations; however, with the DISTINCT clause the size of the resulting list of patents is around 40. If we also search the forward citations, we will generate a larger list of patents, some of which may be highly relevant.

| Query   | Results  |
|---|--|
| <pre> PREFIX ont: &lt;http://minoe.stanford.edu:8080/REGNE SELECT DISTINCT ?class WHERE{ ?cases a ont:CourtCase. ?cases ont:hasBody ?caseBody. ?caseBody j.0:ext ?comment. FILTER REGEX(?comment, "erythropoietin", "i"). ?cases ont:patentsInvolved ?patents. ?patents ont:hasUSClass ?class. } </pre> | <ul style="list-style-type: none"> <li>◆ #536</li> <li>◆ #435</li> <li>◆ #530</li> <li>◆ #514</li> <li>◆ #930</li> </ul> |

Figure 6. US Classes of the Selected Patents.

| Query   | Results  |
|---|--|
| <pre> PREFIX ont: &lt;http://minoe.stanford.edu:8080/REGNE SELECT DISTINCT ?pat2 WHERE{ ?cases a ont:CourtCase. ?cases ont:hasBody ?caseBody. ?caseBody j.0:ext ?comment. FILTER REGEX(?comment, "erythropoietin", "i"). ?cases ont:patentsInvolved ?patents. ?patents ont:hasCitation ?pat2 } </pre> | <ul style="list-style-type: none"> <li>◆ #4677195</li> <li>◆ #4394443</li> <li>◆ #4757006</li> <li>◆ #3865801</li> <li>◆ #4303650</li> <li>◆ #4264731</li> <li>◆ #4503151</li> <li>◆ #4397840</li> <li>◆ #4558006</li> <li>◆ #4695542</li> <li>◆ #4710473</li> <li>◆ #4667016</li> <li>◆ #4703008</li> <li>◆ #3033753</li> <li>◆ #4442205</li> <li>◆ #4338397</li> <li>◆ #4358535</li> </ul> |

Figure 9. Backward Citations from the Selected Patents.

| Query   | Results   |
|---|---|
| <pre> PREFIX ont: &lt;http://minoe.stanford.edu:8080/REGNE SELECT DISTINCT ?inv WHERE { ?cases a ont:CourtCase . ?cases ont:hasBody ?caseBody. ?caseBody j.0:ext ?comment. FILTER REGEX(?comment, "erythropoietin", "i"). ?cases ont:patentsInvolved ?patents. ?patents ont:hasInventor ?inv } </pre> | <ul style="list-style-type: none"> <li>◆ Lin_Fu_Kuen</li> <li>◆ Sugimoto_Kaname_</li> <li>◆ Hayashibara_Yasushi</li> <li>◆ Lai_Por_Hsiung_</li> <li>◆ Strickland_Thomas_WV</li> <li>◆ Hewick_Rodney_M_</li> <li>◆ Seehra_Jasbir_S</li> <li>◆ Cohen_Stanley_N_</li> <li>◆ Boyer_Herbert_WV</li> <li>◆ Seenra_Jasbir_S</li> </ul> |

Figure 7. Identified Inventors of the Selected Patents.

From the results obtained in Figs. 4-9, we can continue searching the knowledgebase for more patents by a particular inventor, or search under specific classes. Another possibility is to go back to the court case silo and gather information about other patents assigned to a particular assignee which have been involved in litigation and in turn use those results to search the patent silo. Hence, the integration of the information, and the ontology presented in this paper allows us to move back and forth between the information silos with fair ease. However, this is still a daunting task to be performed manually. The semantics of the ontology, which include the classes and the properties, provide a platform to build automated tools. The scenario presented above is a mere example of the application of the ontology as it was designed for. The goal of developing

this ontology is to provide integrated knowledgebase which can be extended for use in a variety of other applications as well.

## VI. CONCLUSION

Information pertaining to the patent system is available in multiple silos of heterogeneous information domains. To gather relevant information, one must broadly search the (1) patent documents and file wrappers; (2) scientific literature and (3) court litigations. In this paper, we developed an ontology to standardize the representation of documents in the patent and court case domains and demonstrated how this ontology can act as a knowledgebase to answer queries spanning these multiple domains. The resulting ontology consists of 28 classes, 23 object properties and individuals from 1150 patent documents and 30 court cases.

We developed a use case around the hormone “erythropoietin”. Through a use case scenario of searching for patent prior art, and relevant court litigations in Section V.B, we demonstrate the potential use of such an ontology. Since the ontology expresses semantics from patent and court case domains, one can go back and forth between the two information domains to make inferences based on the cross-referenced information.

Due to the varying structure and formats of the documents, they need to be parsed separately. We realize that court cases are harder to parse and limit the extent to which the information from them can be automatically extracted. Better techniques and stronger regular expressions may be required. We also notice that the naming convention could lead to issues especially when two completely different individuals have the same name. To avoid this, the naming convention of the individuals explained in Section IV. D may have to be modified. A friendly user interface for querying the ontology will be provided. However, to make full use of the ontology, one may have to know the syntax for querying in SPARQL, or any other query language of their choice.

A wide range of users involved in the different stages of the patent system including start-up companies, patent examiners and litigators will benefit from the ontology developed in this paper. Many tools can be built around the knowledgebase to aid the users in the IR process.

### A. Future Work:

Our future work has two parallel directions. First, we will review existing ontologies and create ontologies representing the other information domains in the patent system such as patent file wrappers, scientific literatures and regulatory documents such as the Manual of Patent Examining Procedure (MPEP), Code of Federal Regulations (CFR) etc. We will explore different possibilities of integrating this information by either merging them into a single global ontology, or mapping concepts only as needed.

Our second goal will focus on developing automated tools which will use the ontologies developed to enhance the information retrieval process from all these multiple heterogeneous information silos. The techniques developed will account for varying language and make maximum

utilization of the cross-referenced information [39], [40]. The required domain knowledge is available in the form of bio-ontologies at BioPortal. The developed tools will also integrate this domain knowledge to aid the information retrieval process.

## ACKNOWLEDGMENT

This research is partially supported by NSF Grant Number 0811975 awarded to the University of Illinois at Urbana-Champaign and NSF Grant Number 0811460 to Stanford University. Any opinions and findings are those of the authors, and do not necessarily reflect the views of the National Science Foundation.

## REFERENCES

- [1] Codina J., Pianta E., Vrochidis S. and Papadopoulos S., “Integration of semantic, metadata and image search engines with a text search engine for patent retrieval”, Proceedings of the Workshop on Semantic Search at the 5th European Semantic Web Conference, pp. 14-28, June 2008.
- [2] Fujii, A., “Enhancing patent retrieval by citation analysis”, In Proceedings of the 30th Annual international ACM SIGIR Conference on Research and Development in information Retrieval, New York, pp. 793-794, 2007.
- [3] Ghoula, N., Khelif, K., and Dieng-Kuntz, R., “Supporting Patent Mining by using Ontology-based Semantic Annotations”, Proceedings of the IEEE/WIC/ACM international Conference on Web intelligence, Washington, DC, pp. 435-438, November 2007.
- [4] Giereth M, Brüggemann S, Stähler A, Rotard M, Ertl T., “Application of semantic technologies for representing patent metadata”, In proceedings of the first international workshop on applications of semantic technologies, 2006.
- [5] Giereth M., Koch, S., Kompatsiaris, Y., Papadopoulos, S., Pianta, E., Serafini, L., and Wanner, L., “A Modular Framework for Ontology-based Representation of Patent Information”, Proceeding of the 2007 Conference on Legal Knowledge and information Systems: JURIX 2007, Vol. 165, 49-58, 2007.
- [6] Jackson, P., Al-Kofahi, K., Kreilick, C., and Grom, B., “Information extraction from case law and retrieval of prior cases by partial parsing and query generation”, In Proceedings of the Seventh international Conference on information and Knowledge Management, Bethesda, Maryland, pp. 60-67, Nov 1998.
- [7] Jaffe, Adam B., Trajtenberg, Manuel and Fogarty, Michael S., “The Meaning of Patent Citations: Report on the NBER/Case-Western Reserve Survey of Patentees”, NBER Working Paper No. W7631. Available at SSRN:<http://ssrn.com/abstract=228106>, 2000.
- [8] Kang, I., Na, S., Kim, J., and Lee, J., “Cluster-based patent retrieval”, Information Processing and Management, pp. 1173-1182, vol. 43 (5), Sep. 2007.
- [9] Mukherjea, S. and Bamba, B., “BioPatentMiner: an information retrieval system for biomedical patents”, In Proceedings of the Thirtieth international Conference on Very Large Data Bases, pp.1066-1077, vol. 30. 2007.
- [10] Soo VW, Lin SY, Yang SY, Lin SN, Cheng SL, “A cooperative multi-agent platform for invention based on patent document analysis and ontology”, Expert Systems with Applications, Volume 31(4), pp. 766-775, November 2006.
- [11] Wanner L., Baeza-Yates R., Brugmann S., Codina J., Diallo B., Escorsa E., Giereth M., Kompatsiaris Y., Papadopoulos S., Pianta E., Piella G., Puhlmann I., Rao G., Rotard M., Schoester P., Serafini L. and Zervaki V., “Towards content-oriented patent document processing”, World Patent Information, Volume 30(1), pp. 21-23, March 2008.
- [12] Xue, X. and Croft, W. B., “Automatic query generation for patent search”, In Proceeding of the 18th ACM Conference on information and Knowledge Management, Hong Kong, China, pp. 2037-2040, Nov 2009.

- [13] Yang SY, Lin SY, Lin SN, Cheng SL and Soo VW, "An Ontology-based Multi-agent Platform for Patent Knowledge Management", International Journal of Electronic Business Management, Vol. 3 (3), pp. 181-192, 2005.
- [14] Mitra, P., Wiederhold, G. and Jannink, J., "Semi-automatic Integration of Knowledge Sources", In 2nd International Conference on Information Fusion (FUSION 1999), July 6 - 8, Sunnyvale, CA, 1999.
- [15] Wiederhold, G., & Jannink, J., "Composing diverse ontologies" (Technical Report), Stanford University, Scalable Knowledge Composition (SKC) Project, 1999.
- [16] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner, "Ontology-based integration of information – a survey of existing approaches", In IJCAI-01 Workshop: Ontologies and Information Sharing, pp. 108–117, 2001.
- [17] Noy, N.F. and McGuinness, D.L. "Ontology Development 101: A Guide to Creating Your First Ontology". Development Stanford K, pp. 1-25, 2001.
- [18] Thomas R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing", Int. J. Hum.-Comput. Stud., 43(5-6), pp. 907–928, November 1995.
- [19] Mike Uschold and Michael Grüninger, "Ontologies: principles, methods, and applications", Knowledge Engineering Review, 11(2), pp. 93–155, 1996.
- [20] Mariano F. Lopez, Asuncion G. Perez, and Natalia Juristo, "METHONTOLOGY: from Ontological Art towards Ontological Engineering", In Proceedings of the AAAI '97 Spring Symposium, pp. 33–40, Stanford, USA, March 1997.
- [21] Asunción Gómez-Pérez, Mariano Fernández-López, and Oscar Corcho., "Ontological Engineering: With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web"., (Advanced Information and Knowledge Processing). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.
- [22] Gruninger, M. and Fox, M.S. "Methodology for the Design and Evaluation of Ontologies". In: Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95, Montreal, 1995.
- [23] Google's deal with USPTO. Accessed on 11/10/2010. <http://www.google.com/googlebooks/uspto.html>
- [24] BioPortal. Accessed on 10/25/2010. <http://bioportal.bioontology.org>
- [25] USPTO. Accessed on 10/25/2010. <http://www.uspto.gov>
- [26] I. Horrocks, "DAML + OIL: A description logic for the Semantic Web", IEEE Bull. Technical Committee Data Engrg, Vol. 25 (1), pp. 4–9, 2002.
- [27] RDF W3C Documentation. Accessed on 11/10/2010. <http://www.w3.org/RDF/>
- [28] OWL W3C Documentation. Accessed on 11/10/2010. <http://www.w3.org/TR/owl-ref/>
- [29] SPARQL W3C Documentation. Accessed on 11/10/2010. <http://www.w3.org/TR/rdf-sparql-query/>
- [30] SWRL W3C Documentation. Accessed on 11/10/2010. <http://www.w3.org/Submission/SWRL/>
- [31] Protégé Website. Accessed on 11/10/2010. <http://protege.stanford.edu/>
- [32] Chimaera Website. <http://www.ksl.stanford.edu/software/chimaera>
- [33] LexisNexis Website. <http://www.lexisnexis.com/>
- [34] HeinOnline IP Library. <http://home.heinonline.org/>
- [35] WestLaw Website. <http://www.westlaw.com/>
- [36] Thomson Innovation. <http://www.thomsoninnovation.com>
- [37] PACER. <http://www.pacer.gov/>
- [38] Jeffrey Schox, "Not So Obvious: A Guide to a Patent Law and Strategy for Inventors and Entrepreneurs".
- [39] Hang Yu, S. Taduri, J. P. Kesan, G. T. Lau and K. H. Law, "Retrieving Information Across Multiple, Related Domains Based on User Query and Feedback: Application to Patent Laws and Regulations", International Conference on Theory and Practice of Electronic Governance (ICEGOV2010), 2010.
- [40] S. Taduri, Hang Yu, G. T. Lau, K. H. Law and J. P. Kesan, "Developing a Comprehensive Patent-Related Information Retrieval Tool", Journal of Theoretical and Applied Electronic Commerce Research, special issue on E-government interoperability, enterprise architecture and strategies, Unpublished.