Developing an Ontology for the U.S. Patent System

Siddharth Taduri, Gloria T. Lau
Civil and Environmental Eng.
Stanford University
Stanford, CA, USA
staduri, glau@stanford.edu

Kincho H. Law
Civil and Environmental Eng.
Stanford University
Stanford, CA, USA
law@stanford.edu

Hang Yu, Jay P. Kesan
College of Law
University of Illinois at Urbana-Champaign
IL, USA
hangyu, kesan@illinois.edu

ABSTRACT
The past few years have experienced an explosive growth in scientific and regulatory documents related to the patent system. Relevant information is siloed into many heterogeneous information domains making it a challenging task to gather information. In this paper, we develop an ontology to standardize the representation of the patent system in order to overcome the heterogeneity and integrate information from the patent document, court case and file wrapper domains. Through a use case in the bio domain erythropoietin, we demonstrate how this ontology can be used as a tool to improve the learning curve of users gathering information across these multiple information domains. The proposed ontology provides the required semantics to develop automated tools for a variety of purposes including Information Retrieval (IR) and analytics.

Categories and Subject Descriptors
D.2.13 [**Software Engineering**]: Reusable Software – *Domain Engineering.*
H.3.4 [**Information Storage and Retrieval**]: System and Software – *Question-answering (fact retrieval) systems.*

General Terms
Design, Standardization, Management.

Keywords
Ontology, Patent, Court Cases, File Wrapper, Information Retrieval, Knowledgebase.

1. INTRODUCTION
The past few years have seen a revolutionary change in the way scientific and regulatory information is created, stored and processed. The explosive growth of these documents has led to the rise of intelligent applications to manage and process this information. However, in order to build such an application, one requires a thorough understanding of both the organization of information and the requirements of the targeted users.

In this paper, we focus on the patent system, which involves many such 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 United States Patent and Trademark Office (USPTO) and finally issued or rejected based upon the patent examiner's decision.

The amount of information available is enormous and very highly distributed. Information pertaining to a particular subject is maintained by independent entities in the regulatory system, each enforcing different standards which results in a very heterogeneous set of documents segregated into information silos. Therefore, one requires to simultaneously search multiple information silos in order to gather comprehensive information relating to a particular subject. The prosecution history is documented and is also known as the file wrapper for that issued patent or application. 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, or just a single claim, based on the findings. The various documents involved in both the acquisition phase and the enforcement stage are (a) patent applications; (b) file wrappers; (c) issued patents; (d) any form of prior art such as scientific publications and printed publications; (e) litigations of similar patents; and (f) regulations and laws involved i.e., appropriate chapters of the Code of Federal Regulations (C.F.R.) and the United States Code (U.S.C.).

The two stages of the patent system often function independent of each other i.e., the enforcement stage comes into picture only when the acquisition phase is complete. Both stages involve different users and entities. The requirements of each user or entity drastically vary as per the task. For example, a start-up company (the entity) will need to conduct a thorough patentability search before filing a patent application for their invention with the patent office. The company is mainly concerned with satisfying the utility, novelty and non-obviousness clauses of the U.S.C. which requires a thorough analysis of prior art and prior patent descriptions. As a second example, an established firm with a profitable patent may want to conduct an infringement analysis to enforce their rights during which, they will pay thorough attention to the patent claims, and the file wrappers. In both cases, a significant effort needs to be taken in order to gather relevant information across all the information silos, which are diverse in structure, in syntax, in semantics and in format. Clearly, the patent system is not only diverse in the information it contains, but also in the requirements of the users and entities involved.

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 ontology will define the semantics expressed in the information silos and serve as a platform to integrate the information. We propose to develop a knowledge base by populating the classes of the ontology with

Dg.o'11, June 12–15, 2011, College Park, MD, USA.

information and appropriately relating them. The knowledge base provides the semantics and the representation needed to build automated tools to perform a variety of actions such as analytics and IR. The knowledge base will also serve as a basis for interactive tools to guide and improve the learning curve of users gathering information.

Our current implementation spans three information domains namely – issued patents, court cases and patent file wrappers. As we make progress with these documents, we intend to include other information sources such as scientific publications and regulations. We discuss in detail the current drawbacks and challenges associated with today's technologies. 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. Through this use case, we develop a simple scenario, demonstrating how the ontology can be queried to perform IR.

Section 2 provides a background study on the information silos, namely patents, court cases and file wrappers, and the current state-of-art tools that allow us to access the information and related work in this area. Section 3 introduces the use case and describes the test corpus. Section 4 describes the structure of the documents to help understand the challenges faced with respect to the diversity of the information domains. Section 5 describes the methodology followed to develop the ontology and Section 6 presents a mock scenario to show the application of the ontology. Section 7 concludes the paper by discussing some drawbacks and limitations of the study.

2. 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 IR and integration of the information silos.

2.1 Challenges and State-of-the-Art Tools

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, Japanese and German Patent Offices. The USPTO maintains a database for issued patents, patent applications, copyrights and trademarks. HeinOnline, LexisNexis and WestLaw are libraries for other IP related legal information [31]-[35]. In a recent deal, Google is now 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 [34]. The Derwent World Patents Index (DWPI) is one of the largest patent databases with documents indexed from 41 patent-issuing authorities. Public Access to Court Electronic Records (PACER) is an electronic system to access the databases of the 94 District Courts and 13 Courts of Appeals (CAFC) [35]. Currently, PACER requires one to know the party name or the case number; in other words, it does not allow keyword-based search. Also, manually scanning each of these databases is not a feasible option.

In 2003, the USPTO introduced the Image File Wrapper (IFW) system to replace the paper based system. The Image File Wrappers are available for more recent patents on the Patent Application Information Retrieval (PAIR) website. However, several challenges are to be overcome to make these documents computer accessible. The USPTO does not permit automated crawling of the IFWs and requires one to enter a CAPTCHA verification code to access the documents. Google has recently started indexing these documents and provides a web service to download these files [38]. However, the files are still available as images, which means additional processing and smart OCR algorithms are required to extract text from them. To access file wrappers prior to 2003, a 3rd party agent is currently the best solution to convert the paper based file wrappers to text-readable file wrappers [34]. IFW Insight is a tool which has indexed over a 1000 IFWs and allows one to navigate and search for critical information contained within them [39]. However, a strong integration with other information domains in the patent system is still lacking. There are several structural and organizational challenges associated with IFWs which are addressed in the later sections.

2.2 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. Alternative architectures suggest 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 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 are required to be made more efficient. Several IR 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 PATEXPERT project has developed an ontology for the patent document domain which focuses on the European patent system [4], [5], [11]. However, the above mentioned methodologies focus on a single information silo, 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 standardize the representation of the information either through a single ontology, or to construct individual ontologies and subsequently integrate them. Secondly, the IR techniques need to be improved to take advantage of the implicit cross-referencing between the various information domains.

3. USE CASE

The working of the ontology is demonstrated by constructing a use case in the bio domain – 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 hemoglobin 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. 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. These 135 related patents identified will serve as the gold standard for any performance tests.

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.

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, the 135 closely related 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.

4. 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. (see Figure 1). We exploit this structure and developed a script to automatically 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. Also, since the structure of court cases is not as well defined as patent documents, parsing these documents is more of a challenge (see Figure 2). The important fields, such as the plaintiff, the defendant, the court etc. are thus extracted using a carefully coded script.

As mentioned in Section 2.1, file wrappers for patents dated earlier than 2003 are only available in paper form. We requested a

Primary Examiner: Martinell; James
Attorney, Agent or Firm: Bell, Boyd & Lloyd

Parent Case Text

This is a continuation of application Ser. No. 07/957,073, filed Oct. 6, 1992, abandoned, which is a continuation of application Ser. No. 07/609,741, filed Nov. 6, 1990, now abandoned, which is a continuation of application Ser. No. 07/113,179, …

Claims

What is claimed is:

1. A pharmaceutical composition comprising a therapeutically effective amount of human erythropoietin and a pharmaceutically acceptable diluent, adjuvant or carrier, wherein said erythropoietin is purified from mammalian cells grown in culture.

2. A pharmaceutically-acceptable preparation containing a therapeutically effective amount of erythropoietin wherein human serum albumin is mixed with said erythropoietin.

Description

BACKGROUND
…

Figure 1. Sample Patent Document

927 F.2d 1200 (1991)

AMGEN, INC., Plaintiff/Cross-Appellant,
v.
CHUGAI PHARMACEUTICAL CO., LTD., and Genetics Institute, Inc., Defendants-Appellan

Nos. 90-1278, 90-1275.
United States Court of Appeals, Federal Circuit.
March 5, 1991.
Suggestion for Rehearing Declined May 20, 1991.
…
Before MARKEY, LOURIE and CLEVENGER, Circuit Judges.
…
THE PATENTS
On June 30, 1987, the United States Patent and Trademark Office (PTO) issued to Dr. Rodney Hew U.S. Patent 4,677,195, entitled "Method for the Purification of Erythropoietin and Erythropoie Compositions" (the '195 patent). The patent claims both homogeneous EPO and compositions ther and a method for purifying human EPO using reverse-phase high performance liquid chromatograp The method claims are not before us. The relevant claims of the '195 patent are:
1. Homogeneous erythropoietin characterized by a molecular weight of about 34,000 daltons SDS PAGE, movement as a single peak on reverse phase high performance liqu chromatography and a specific activity of at least 160,000 IU per absorbance unit at 2 nanometers.
* * * * * *
3. A pharmaceutical composition for the treatment of anemia comprising a therapeutically effect amount of the homogeneous erythropoietin of claim 1 in a pharmaceutically acceptable vehicle.
4. Homogeneous erythropoietin characterized by a molecular weight of about 34,000 daltons SDS PAGE, movement as a single peak on reverse phase high performance liqu chromatography and a specific activity of at least about 160,000 IU per absorbance unit at 2 nanometers.
* * * * * *

Figure 2. Sample Court Case

copy of the file wrapper for one of the core patents U.S. 5,955,422 for the purpose of processing the information contained within it. We will demonstrate the use of the proposed tool using this file wrapper, and include more file wrappers in future. File wrappers are highly unstructured documents which make it very hard to automate parsing of these documents. In general, the prosecution history contains of an initial application, an amendment from the applicant and the examiner's response with either a rejection, or an approval for issue.

However, due to the nature of the prosecution, this process can continue over several transactions until the application is finally accepted, or withdrawn. Hence, the first challenge is identifying each of these transactions since every patent file wrapper will have a different number of transactions and could be out of order. Several miscellaneous documents such as the fee structure are ignored in our model. The next challenge is that each of these transactions does not have a standardized format and is generally in the form of a letter with important information such as restricted claims, allowed claims, rejected claims and corresponding arguments are expressed in a mixed form within the text (see Figure 3). Some file wrappers include special events such as an interference. Due to the above mentioned reasons, we have manually parsed the file wrapper, extracting information we

During a telephone conversation with Mr. Kokulis on March 25, 1992 provisional election was made with traverse to prosecute the invention of Group VI claims 61-63. Affirmation of this election must be made by applicant in responding to thi Office action. Claims 1-60 are withdrawn from further consideration by the Examiner, 3 CFR 1.142(b), as being drawn to a non-elected invention.

Claim 63 is rejected under 35 U.S.C. S 112, second paragraph, as bein indefinite for failing to particularly point out and distinctly claim the subject matter whic applicant regards as the invention.

Claim 63 is vague and indefinite in the recitation of "recombinan erythropoietin". The specification discusses several different recombinant systems fo production of EPO. It appears that different recombinant systems produce differen modifications of the protein. It is not clear that all different modifications are intended t be encompassed by the claims.

Claims 61 and 62 are allowed.

Figure 3. Excerpt of a Rejection Letter

need. However, a significant effort will be directed towards developing intelligent parsers to extract critical information from file wrappers in future.

5. MODELING THE DOMAINS

5.1 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 [17]. Keeping our primary goal in mind, which is enabling integration of data from the information domains in the patent system, we will define a set of competency questions which – (1) confine to a single domain such as patents; and (2) span multiple domains. The competency questions in no way limit the applications of this ontology, rather they are examples of questions the ontology must be capable of answering at a minimum. The following competency questions are not necessarily useful questions, but they show the variety of queries that can be answered by the ontology which include numeric filters, regular expressions, cross-domain questions, ordering etc. In Section 6, through a scenario some of the competency questions will be answered to demonstrate the working of the ontology.

5.1.1 Patent Document Domain:

- Return all inventors who have 3 or more patents.
- Return all the patent documents which contain the keyword "erythropoietin" in at least 3 claims and assigned to "Amgen_Inc".

5.1.2 Court Case Domain:

- Return all court cases which contain the keyword – "erythropoietin"
- Return all court cases which involve "Amgen_Inc" either as the plaintiff, defendant of both, and from the United States District Court for the District of Massachusetts.

5.1.3 File Wrappers:

- Display all the events contained within the file wrapper of a particular U.S. patent, arranged in chronological order.
- Enlist all the claims of the initial application, identifying claims which were rejected and the claims that were finally allowed.

5.1.4 Multi-domain:

- Return all patents which contain the keyword – "erythropoietin" in the "claims", which has been challenged in the courts at least once.
- From a file wrapper, identify the patents involved in an interference, display information about the inventor, assignee, and claims of that patent. Further, enlist the other patents the inventor owns, if any.

5.2 Conceptualizing

Patents, court litigations and file wrappers are highly inter-related documents. There is a significant amount of cross-referencing between these documents. For example, the court litigations directly reference the patent numbers, the assignee, the claims of concern etc. The court litigations highly rely on the information from the file wrapper. The cross-referencing can be crucial when relevancy has to be established between these documents. Present systems take little or no advantage of this form of cross-referencing. The proposed ontology will provide these additional semantics between the information domains via properties.

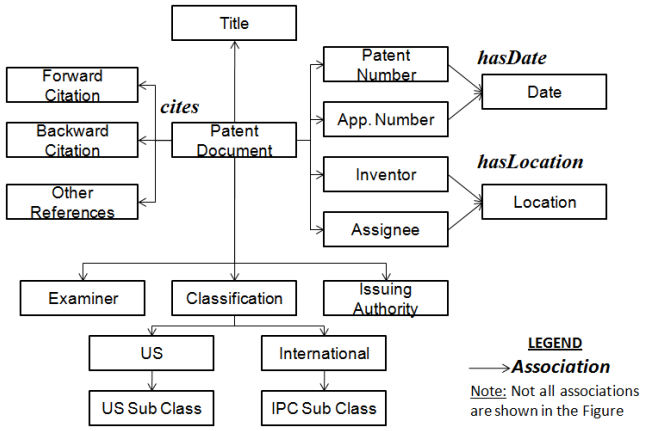

Figure 4. A Conceptual view of the Patent Document Domain

When designing the ontology, we model each information domain separately, and then arrange them together in a poly-axial hierarchy. We generate properties or relations between classes from within an information domain, and with other information domains which allows users to reason across the multiple information domains. We recognize that some applications may require access to a small fragment of the semantics and information. Hence, applications should be allowed to work with only what is needed. For example, a patent analytics tool may only consider the patent document domain and ignore the semantics with other domains.

5.2.1 Patents and Court Cases

A patent document consists of information in both the textual content and metadata. Figure 4 gives a conceptual view of the patent document domain and the important "terms" from an IR perspective. We follow a bottom-up approach when creating the class hierarchy. The lower level classes such as Abstract, Inventor, Plaintiff etc. are appropriately grouped into general classes. Also, we must define the abstraction levels or the boundary between instances and classes in our model. For example, the textual content includes the abstract, the claims, the description etc. We consider Claim to be a general class, and every claim of a patent is an individual of type Claim. Similarly from the metadata, information such as the filing date, issuing date, inventor, assignee, citations, international classification, U.S. classification etc. are identified as concepts or classes. We arrange these concepts in a hierarchical manner as shown in Figure 5. The classes Documents, Information and Events are the root nodes. Two other root nodes, Person and Organization are excluded from the figure to ensure clarity. A similar conceptual view of court cases can also be studied to identify classes, individuals, properties etc. Important information in the court cases is contained within the text of the body which includes the claims under concern, the patents involved and the analysis. Extracting the patents involved in the court case, the parties involved i.e. the defendant and/or the plaintiff, along with identifying key terms

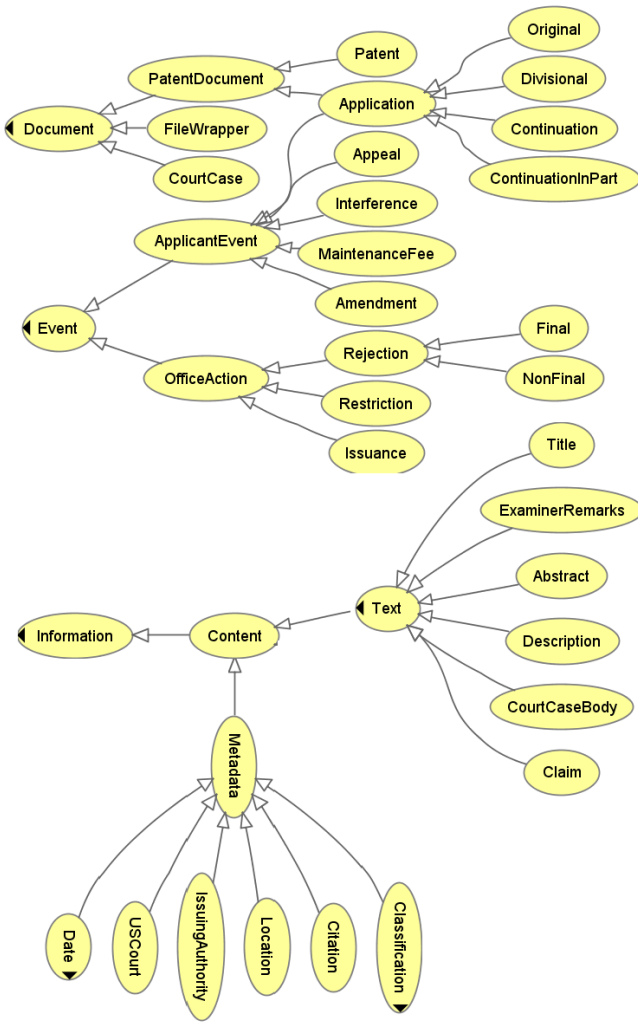

Figure 5. Class Hierarchy of the Patent System Ontology

can be a very effective way to co-relate court cases and patent documents.

5.2.2 File Wrappers

A file wrapper includes details of every communication that happened between the patent office and the applicant (or the applicant's attorneys). These details include the initial application, applicant's amendments, examiner's response – either as a rejection, or an approval, with appropriate explanations, interference records (if any), facsimile transactions and other miscellaneous documents such as fee structure, extension of time etc. We group all the transactions broadly under the class Event. To model the file wrapper domain, a thorough understanding of the terminology is essential [25]. The examiner's responses such as a rejection (final or non-final), an approval, or any other office action are grouped together under a single parent node called OfficeAction. Any event occurring from the applicant's side such as an appeal for interference or an amendment is grouped under another common parent class called ApplicantEvent. Figure 5 shows this hierarchy of the concepts from the three information domains.

In addition to the events, more concepts and properties are defined at a finer level of granularity. For example, an interference record consists of the following information – (1) The patent or application which is interfering; (2) The date of the interference; (3) The interfering claims in the other patent application; (4) The interfered claims in the current application; (5) The corresponding count; (6) The decision made, i.e. favorable or not favorable (see Figure 6). Each of the above parts of the interference document is encoded into the interference class via object and data type properties.

Table 1. Summary of properties relating various classes.

Object Property	Domain	Range
contains	FileWrapper	Event, Patent Document
allowedClaim	Rejection	Claim
hasDate	Patent Document, Event	Defined by the sub-properties
withdrawnClaim	Rejection	Claim
hasAssignee	Patent Document	Assignee
hasBody	CourtCase	CourtCaseBody
hasCitation	Patent Document	Patent
hasClaim	Patent Document	Claim
hasDefendant	CourtCase	Defendant
hasUSClass	Patent Document	USClass
isLocated	Inventor or Assignee	Location
patentsInvolved	CourtCase	Patent
precededBy	CourtCase	Judge

As mentioned in earlier in this Section, the file wrappers are highly correlated to the other sets of documents. Figure 6 shows an excerpt of an interference contained within the file wrapper for U.S. patent 5,955,422. Notice that there is a much smaller subset of the originally filed 63 claims which are involved in the interference. However, in order to see the claims, a user would have to search for the application in the USPTO database and refer to the specified claims. The user would then have to open the issued patent database to refer to the claims of the infringing patent. When manually reading a file wrapper, if one needs more information about the interfering patent or application than what is provided on the interference record, the patent or application will have to be looked up in another database, i.e. the information is not available hands on. Instead, the proposed ontology can be used to recognize the interfering patent as an individual in the Patent class. Hence, any information regarding that patent can be

Count 1

An erythropoietin-containing, pharmaceutically acceptable composition wherein human serum albumin is mixed with erythropoietin.

The claims of the parties which correspond to Count 1 are:

Lin: Claims 61-63

Shimoda et al.: Claims 3-4

Figure 6. Excerpt from an Interference Letter

easily looked up such as the interfering claims (1-4). Figure 7 shows a snapshot of an individual in the interference class. Notice how this particular instance points to the actual claims of the patents, instead of simply listing the text as in the interference letter.

"Special events" such as phone conversations between the examiner and the applicant are not sufficiently recorded on paper. This imposes an additional challenge when conceptualizing and extracting information from the file wrapper domain. Currently, such events are not modeled in the ontology. The methodology employed in designing this ontology is an iterative process. The hierarchy and concepts shown in Figure 5, along with the properties and relations are expected to undergo revisions and improvements several times. A summary of the properties relating the various classes along with the domain and range restrictions is shown in Table 1.

Several representation languages are available such as Resource Description Framework (RDF), RDF Schema (RDFS) and the W3C recommendation Web Ontology Language (OWL) [27]-[28]. We use OWL to encode our ontology since it provides richer semantics than RDF and RDFS. We use Protégé 3.4.x as the ontology editor [30]. We choose SPARQL Protocol and RDF Query Language (SPARQL) to query the knowledgebase [29].

The current version of the knowledgebase is populated with the 1150 U.S. patents and 30 court cases from our corpus which is described in Section 3. Other patent documents which may have been found in court cases or through patent citations, but not in the original 1150 documents are instantiated but contain no information about the patent since the original document itself is unavailable in our corpus. However, we ignore any documents which are not a part of our corpus when performing the tests. The file wrapper for U.S. patent 5,955,422 has also been partially incorporated into the knowledgebase. Currently, only the first amendment, rejection, interference and the original application from the file wrapper are populated. The instantiated OWL ontology is available online at [41]. Currently, we use Protégé to access and query the ontology. Since the ontology contains a large number of individuals, we plan to create an RDF store using tools such as Virtuoso in order to make it scalable [42]. A searchable Lucene index, with links to the original patent and court case documents is also available on the web site.

6. RESULTS

6.1 Evaluation of the Extracted Data

As mentioned in Section 4, the patent documents downloaded from USPTO have a fairly consistent structure. We developed a regular expression based parser to extract information from the patent documents in order to instantiate the ontology. However, a small fraction of documents may be structured differently, which can lead to inaccurate extraction of the data. In this section, the quality of the extracted data is evaluated and efforts needed to improve the quality of the extracted data are discussed.

For the purpose of the evaluation, a random sample of 50 patents is generated. The data from these 50 patents is automatically parsed using the parser. The true data is manually extracted from the original documents in order to be compared with the extracted data. No inconsistencies were found in the fields – Title, Abstract, Claims, Description, Patent Number, Publications, Inventors, Patent Citations and the US and International Classifications. However, 2 false negatives and 48 true positives were found for the Examiner field. Also, 1 false positive and 50 true positives

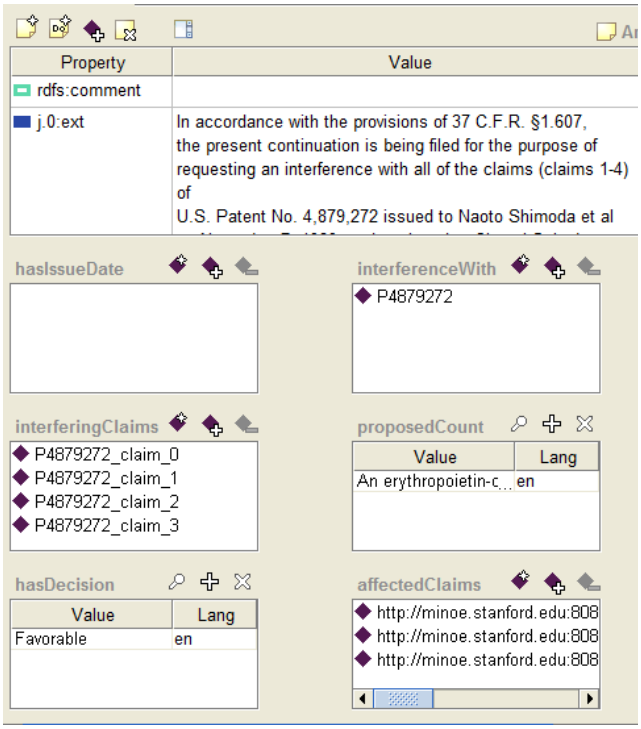

Figure 7. Individual from the Interference Class

were found for the Assignee fields. Table 2 gives the values for precision and recall over these 50 patents for the Assignee and Examiner fields. In future revisions of the parser, we wish to modify the regular expressions to handle the exceptions which lead to the false positives and negatives. Since the set of 50 patents were randomly chosen, we can assume the values for precision and recall hold for larger data sets as well.

Table 2. Evaluation of the Extracted Data

Field	Precision	Recall
Assignee	0.961	1
Examiner	1	0.96

In the original documents, it is possible that the true data, for example, an Inventor's name, is expressed differently in different documents. Possible variations can include upper and lower cases, inversion of first and last names etc. These differences can cause the same information to be considered distinct by the parser. We ignore these differences in our evaluation. However, some of these variations may be easily avoided by simple techniques such as converting all data to lower case. Given the high precision and recall values, the corpus of US patents can be mapped to the ontology with high quality using the existing parser.

Since the court cases are not structured as consistently as the patent documents (see Figure 2), regular expression based parsers do not perform very well resulting in lower precision and recall values for the extracted data. Due to the small number of court cases in our corpus, it is possible to manually correct the extracted data prior to instantiating them in the ontology. However, the current methods for parsing need to be improved upon in order to scale to much larger databases of court cases. Newer NLP techniques such as Named Entity Recognition (NER) and Hidden Markov Models (HMM) can be explored in addition to regular

expressions to accurately extract data from the court cases. Figure 2 highlights some of the relevant data from court cases. The improvement of the parsers is out of the scope of the current paper, although our future implementations will address this issue.

6.2 Querying the Ontology

In this section, we will demonstrate possibility of the patent ontology being used as a tool for improving the learning curve of a user wanting to gather information in the patent system related to the production of erythropoietin. Generally, file wrappers are referred to during the enforcement stage of the patent system, typically during infringement analysis or court litigation. When the claims of two patents do not literally infringe, it is important to determine the scope of each limitation of the claim under the "doctrine of equivalents". For this, the patent's entire file history will have to be studied, and the focus is set on the wordings of the claim and how they evolved. This is a very non-trivial task and involves tremendous amount of reasoning. There are several methods to perform a search in the patent system for relevant information. Each individual will employ a different method to gather information. We have generated a series of questions the knowledgebase will be queried with to serve an example.

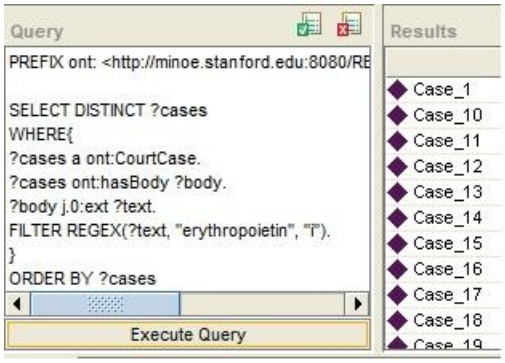

Figure 8. List of Court Cases Related to Erythropoietin

Assuming the user has little information available at the start to perform a thorough search, we would like an efficient way of finding important patents related to "erythropoietin". One possible method is to look at court litigations which have involved erythropoietin, and back track to find the patents involved. Figure 8 shows the list of court cases which have the term erythropoietin in their text.

All the court case documents in our corpus show up in this search since every one of them contains the term "erythropoietin". In the query shown in Figure 8, first all the individuals belonging to the CourtCase class are identified, i.e. the court cases themselves. We then extract the body of the court cases through the hasBody relation. The actual text of the court case body is stored under the annotation property "resourceVal". Hence, we first extract this text followed by a regular expression filter to select only those cases which contain the term "erythropoietin".

In the next step, the patents involved in these court cases are identified. To do this, we follow the "patentsInvolved" relation from the court case domain to the patent document domain. Figure 9 shows the list of patents involved in these court cases. In an expanded view, U.S. patent 5,955,422 is identified as a very frequently occurring patent, which also happens to be one of Amgen's core patents.

Further, we can choose to study the file wrapper of the patent U.S. 5,955,422. The query shown in Figure 10 displays all the events

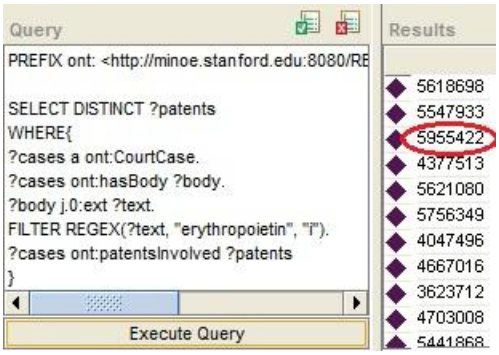

Figure 9. List of Patents Involved in Erythropoietin Related Court Cases

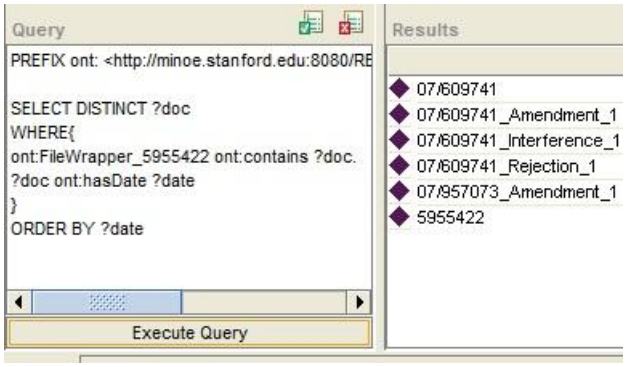

Figure 10. List of Events Contained within the File Wrapper

contained within the file wrapper. This list is obtained via the "contains" property. We order the results by the date in which the occurred. Notice that the initial application (07/609741) and the final issued patent (5,955,422) are both part of the file wrapper. It is possible to view the nature of the application, i.e. whether filed as a continuation, continuation-in-part, divisional or a fresh application. An example where such information is useful is when determining the priority date for certain parts of the application. If this application is a continuation or a divisional, in a more complex query, it will be possible to trace back to the root, i.e. the original application.

In the next few steps, we attempt to show how the patent system ontology can be used to gather information from the file wrapper in a more efficient way. The initial claims as filed by the applicant are generally very different from what is finally allowed. The final scope of the claims is determined by the added limitations which make the claim acceptable. The issued patent by itself will not contain the original claims. However from a file wrapper, this information can be extracted as shown in Figure 11.

Figure 12 provides a snapshot of the Rejection class. A rejection can possibly have restrictions, allowed claims, withdrawn claims, and appropriate arguments (not shown in Figure) etc. As explained in Section 4, the ontology captures these aspects of the documents. The Restriction class is defined as a grouping of claims under a certain U.S. class as advised by the patent examiner. Hence, each individual in the Restriction class points to a set of claims, and at the same time points to the U.S. class they are grouped under.

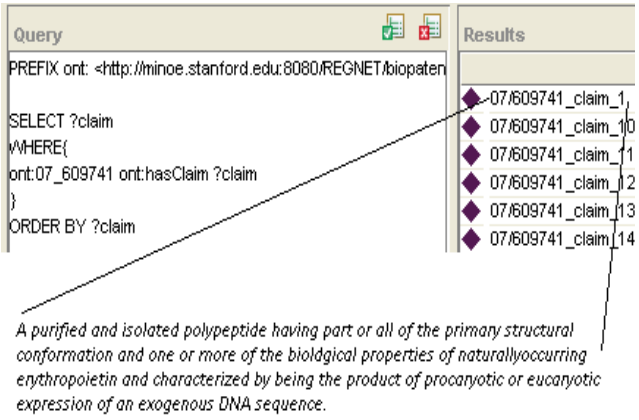

Figure 11. Initial Claims

Similarly, other restrictions can be viewed as well (see Figure 13) via the hasRestriction property. The actual text of the rejection letter is also included under the resourceVal annotation property. Due to the restriction the applicant is only allowed to pursue one of the groups of claims for approval. From Figure 12, we see several claims are withdrawn, and of the remaining 3 claims, two are accepted and one is rejected.

We can access the claims that are allowed via the allowedClaim property. The text of the claims can also be viewed as shown in Figure 14. In a similar fashion, we can compare the text of the claims at every stage of the prosecution of the application including the final claims (see Figures 11 and 14) to identify the added limitation which made the claims acceptable to the examiner.

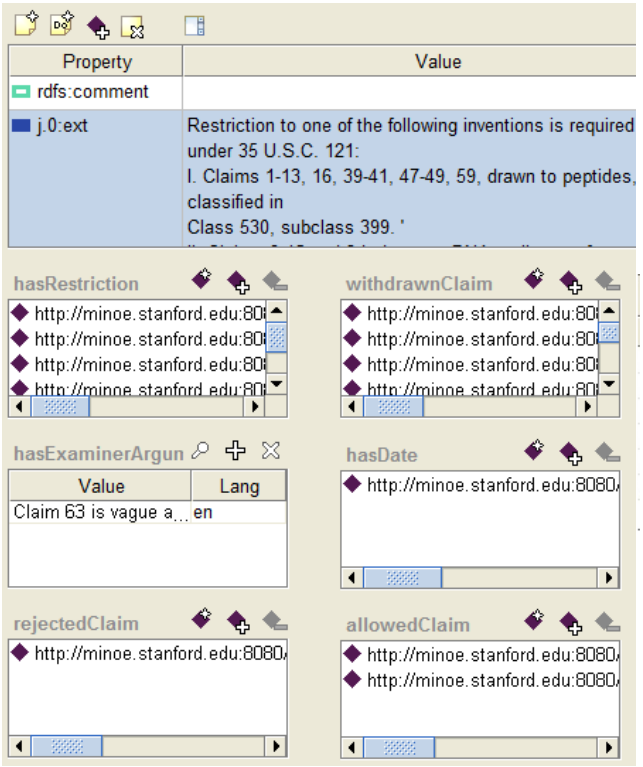

Figure 12. Snapshot of an Individual in the Rejection Class

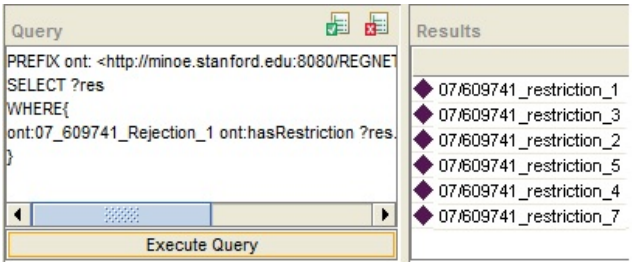

Figure 13. The Restrictions Imposed on the Initial Application

This process of querying the file wrapper can continue as long as required. Many similar scenarios can be constructed as desired by the user. The true potential of the ontology will be visible when complex queries spanning more than one information domain are presented. The ontology takes advantage of the highly cross-referenced information and provides the required semantics to jump from one domain to another with ease. However this is still a daunting task to perform manually. The semantics will allow machines to automatically process the information and perform highly complex tasks such as IR and analytics. The fine granularity of the ontology allows applications and tools to address different users requirements.

7. 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; (2) file wrappers; (3) scientific literature; (4) court litigations; and (5) corresponding regulations and laws. In this paper, we developed an ontology to standardize the representation of documents in the patent, court case and file wrapper domains and defined properties along which they are related. We demonstrated how this ontology can act as a knowledgebase to answer queries spanning these multiple domains. The resulting ontology consists of 54 classes, 36 object properties, 3 datatype properties, 2 attribute properties and over 15,000 individuals from 1150 patent documents, 30 court cases and 1 file wrapper. The instantiated OWL ontology is available online for download [41]. Also included on the web site is a searchable index for the patent and court case documents with a direct link to the original documents.

We developed a use case around the hormone “erythropoietin”.

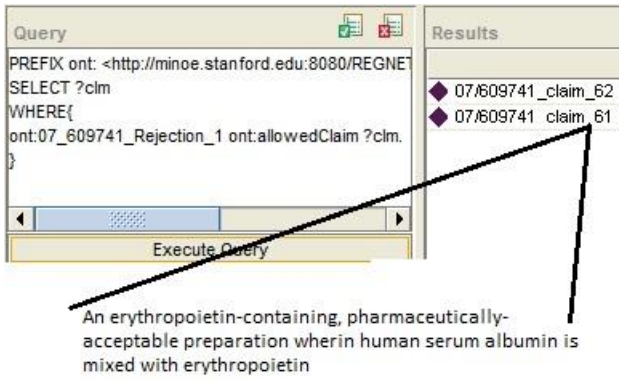

Figure 14. Text of the Claims which were Allowed in the Rejection

Through a use case scenario in Section 6, we demonstrate the potential use of such an ontology as a tool for improving the learning curve of a user wanting to gather information from the patent system. Since the ontology expresses semantics from multiple information domains, one can go back and forth between the information domains to make inferences based on the cross-referenced information. The proposed ontology serves as a base for (1) a tool that will expedite the learning process for someone researching in the patent system; and (2) automated tools intended for a variety of applications such as IR and analytics.

Due to the varying structure and formats of the documents, they need to be parsed separately. We realize that court cases and file wrappers 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. Although we have implemented a naming convention to avoid conflicts, 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 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 automated tools can be built around the knowledgebase to aid the users in their research.

7.1 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 scientific literatures and regulatory documents including the Manual of Patent Examining Procedure (M.P.E.P.), Code of Federal Regulations (C.F.R.) 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. Simultaneously, we propose to provide access to these ontologies and develop a tool allowing one to navigate through the concepts and data. Since the ontology contains a large number of individuals, we will implement an RDF store suing tools such as Virtuoso in order to make it scalable and provide efficient access.

Our second goal will focus on developing automated tools which will use the ontologies developed to enhance the IR 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 [36], [37]. The required domain knowledge is available in the form of bio-ontologies at BioPortal. This research is a continuation of our previous work [40].

Currently, a regular expression based parser is implemented to automatically extract data from patent and court case documents. However, it is challenging to achieve accurate extraction of data using automatic parsers, especially for court cases since the relevant information is not uniformly structures. In future, we wish to address this issue by exploring newer NLP techniques and providing tools that will allow new documents to conform to the structure defined by the ontology.

8. ACKNOWLEDGMENTS

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.

9. REFERENCES

[1] Codina J., Pianta E., Vrochidis S. and Papadopoulos S. 2008. 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*, 14-28.

[2] Fujii, A. 2007. 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, 14-28.

[3] Ghoula, N., Khelif, K., and Dieng-Kuntz, R. 2007. Supporting Patent Mining by using Ontology-based Semantic Annotations. *Proceedings of the IEEE/WIC/ACM international Conference on Web intelligence*, Washington, DC, 435-438.

[4] Giereth M, Brügmann S, Stäbler A, Rotard M and Ertl T. 2006. 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. 2007. A Modular Framework for Ontology-based Representation of Patent Information. *Proceedings of the 2007 Conference on Legal Knowledge and information Systems: JURIX 2007*, 165, 49-58.

[6] Jackson, P., Al-Kofahi, K., Kreilick, C., and Grom, B. 1998. 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, 60-67.

[7] Jaffe, A. B., Trajtenberg, M. and Fogarty, M. S. 2000. The Meaning of Patent Citations: Report on the NBER/Case-Western Reserve Survey of Patentees. *NBER Working Paper No. W7631*, 2000.

[8] Kang, I., Na, S., Kim, J., and Lee, J. 2007. Cluster-based patent retrieval. *Information Processing and Management*, 43, 5 (Sep 2007), 1173-1182.

[9] Mukherjea, S. and Bamba, B. 2007. BioPatentMiner: an information retrieval system for biomedical patents. *In Proceedings of the Thirtieth international Conference on Very Large Data Bases*, 30, 2007, 1066-1077.

[10] Soo VW, Lin SY, Yang SY, Lin SN, Cheng SL. 2006. A cooperative multi-agent platform for invention based on patent document analysis and ontology. *Expert Systems with Applications*, 31, 4 (November 2006), 766-775.

[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. 2008. Towards content-oriented patent document processing. *World Patent Information*, 30, 1 (March 2008), 21-23.

[12] Xue, X. and Croft, W. B. 2009. Automatic query generation for patent search. *In Proceeding of the 18th ACM Conference on information and Knowledge Management*, Hong Kong, China, Nov 2009, 2037-2040.

[13] Yang S. Y., Lin S.Y., Lin S. N., Cheng S. L. and Soo V. W. 2005. An Ontology-based Multi-agent Platform for Patent Knowledge Management. *International Journal of Electronic Business Management*, 3, 3 (2005), 181-192.

[14] Mitra, P., Wiederhold, G. and Jannink, J. 1999. Semi-automatic Integration of Knowledge Sources. *In 2nd International Conference on Information Fusion (FUSION 1999)*, July 6 – 8, Sunnyvale, CA.

[15] Wiederhold, G. and Jannink, J. 1999. Composing diverse ontologies (Technical Report). *Stanford University, Scalable Knowledge Composition (SKC) Project*.

[16] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster,H. Neumann, and S. Hübner. 2001. Ontology-based integration of information – a survey of existing approaches. *In IJCAI–01 Workshop: Ontologies and Information Sharing*, 108–117.

[17] Noy, N.F. and McGuinness, D.L. 2001. Ontology Development 101: A Guide to Creating Your First Ontology. *Development Stanford K*, 1-25.

[18] Thomas R. Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud*, 43, 5-6 (November 1995), 907–928.

[19] Mike Uschold and Michael Grüninger. 1996. Ontologies: principles, methods, and applications. *Knowledge Engineering Review*, 11, 2 (1996), 93–155.

[20] Lopez M. F., Perez A. G., and Juristo N. 1997. METHONTOLOGY: from Ontological Art towards Ontological Engineering. *In Proceedings of the AAAI '97 Spring Symposium*, Stanford, USA, 33–40.

[21] Pérez A. G., López M. F., and Corcho O. 2007. 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.

[22] Gruninger, M. and Fox, M.S. 1995. Methodology for the Design and Evaluation of Ontologies. *In: Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing IJCAI-95*, Montreal.

[23] Google's deal with USPTO. Accessed on 01/28/2011. http://www.google.com/googlebooks/uspto.html

[24] BioPortal. Accessed on 01/28/2011. http://bioportal.bioontology.org

[25] USPTO. Accessed on 01/28/2011. http://www.uspto.gov

[26] I. Horrocks. DAML + OIL: A description logic for the Semantic Web. *IEEE Bull. Technical Committee Data Engrg*, 25, 1, 2002, 4–9.

[27] RDF W3C Documentation. Accessed on 01/28/2011. http://www.w3.org/RDF/

[28] OWL W3C Documentation. Accessed on 01/28/2011. http://www.w3.org/TR/owl-ref/

[29] SPARQL W3C Documentation. Accessed on 01/28/2011. http://www.w3.org/TR/rdf-sparql-query/

[30] Protégé Website. Accessed on 01/28/2011. http://protege.stanford.edu/

[31] LexisNexis Website. http://www.lexisnexis.com/

[32] HeinOnline IP Library. http://home.heinonline.org/

[33] WestLaw Website. http://www.westlaw.com/

[34] Thomson Innovation. http://www.thomsoninnovation.com

[35] PACER. http://www.pacer.gov/

[36] Hang Yu, S. Taduri, J. P. Kesan, G. T. Lau and K. H. Law. 2010. 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)*.

[37] 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*. In Press.

[38] Google USPTO PAIR data. http://www.google.com/googlebooks/uspto-patents-pair.html

[39] IFW Insight. http://ifwinsight.com/

[40] S. Taduri, G. T. Lau, K. H. Law, Hang Yu, J. P. Kesan. 2011. An Ontology to Integrate Multiple Information Domains in the Patent System. *2011 IEEE International Symposium on Technology and Society (ISTAS)*, May 23-25, 2011, Accepted.

[41] Patent Ontology Dataset. http://minoe.stanford.edu:8080/REGNET/biopatent/

[42] Virtuoso. http://virtuoso.openlinksw.com/