# A Patent System Ontology for Facilitating Retrieval of Patent Related Information

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

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

## ABSTRACT

The recent years have seen a tremendous growth in research and developments in science and technology, and an emphasis in obtaining Intellectual Property (IP) protection for one's innovations. Information pertaining to IP for science and technology is siloed into many diverse sources and consists of laws, regulations, patents, court litigations, scientific publications, and more. Although a great deal of legal and scientific information is now available online, the scattered distribution of the information, combined with the enormous sizes and complexities, makes any attempt to gather relevant IP-related information on a specific technology a daunting task. This paper describes a knowledge-based software framework to facilitate retrieval of patents and related information across multiple diverse and uncoordinated information sources in the US patent system. The document corpus covers issued US patents, court litigations, scientific publications, and patent file wrappers in the biomedical technology domain.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – retrieval models,

J.1 [Administrative Data Processing]: law.

## General Terms

Algorithms, Design, Economics, Experimentation

## Keywords

Ontology, Patent, Information Retrieval, Knowledgebase.

## 1. INTRODUCTION

The recent years have seen a tremendous growth in research and developments in science and technology, and an emphasis in obtaining Intellectual Property (IP) protection for one's innovations. IPs are important assets of any organization. During the lifetime of a patent, from its initial filing, patent issuance to disputes and litigations, the patent system will constantly be searched for information. However, information pertaining to IP and the patent system for science and technology is siloed into many diverse sources and consists of laws, regulations, patents, court litigations, and more. Although a great deal of legal and government information is now available online, the scattered distribution of the information, combined with the enormous sizes and complexities, makes any attempt to gather relevant IP-related information, even on a specific technology, a daunting task. Currently, the task of gathering IP-related information is performed manually and is both laborious and expensive. This falls disproportionally on smaller firms, start-ups, and individual inventors who have very limited resources. This paper describes the development of a patent system ontology to facilitate retrieval of patents and related information across multiple diverse and uncoordinated information sources in the US patent system.

This paper is organized as follows: Section 2 discusses the background and motivation for this research. Section 3 briefly describes our use case and document corpus. Section 4 introduces the patent system ontology and its structure. Section 5 briefly describes the Information Retrieval (IR) framework. Illustrative examples are provided in Section 6. Section 7 concludes this paper by summarizing the current status.

## 2. BACKGROUND AND MOTIVATION

The following scenarios illustrate some of the issues faced with the current patent system:

- A company looking to patent its technology on medical imaging devices, for example, is required to perform an initial patentability search and establish the usefulness, novelty, and non-obviousness of the technology [1]. The patentability search involves a thorough study of prior art including scientific literature and patent databases, competitor analysis, existing litigations to similar technologies, and regulations issued by government agencies such as the Federal Drug Agency (or any agency enforcing laws with respect to medical imaging devices and related technologies).
- Similar to the patent applicant, a patent examiner performs patentability search when examining an application. As of 2009, the United States Patent and Trademark Office (USPTO) employs about 6,242 patent examiners and received over 456,106 utility patent applications [28].[1] Roughly, this translates to around 73 patents per examiner

[1] The USPTO's annual statistics can be accessed at http://www.uspto.gov/web/offices/ac/ido/oeip/taf/reports.htm (Accessed on 03/01/2012)

annually and approximately 1.5 patents per examiner per week. Although a patent examiner is generally well-versed with the technological domain of the patent application, the situation imposes a serious time constraint during the review process. Hence, each application receives lesser time and leads to insufficient examination, and possibly infringement or invalidation at a later stage.

- To protect IPs, companies may perform an infringement analysis to ensure that a particular patent's right is not being infringed. The consequences of an infringement can be severe and result in heavy losses. Infringement analysis involves a thorough search in the issued patents database, patent application database, prior court litigations, regulations, and any form of documented evidence to help assert the infringement or invalidate an infringing patent's claim as a defensive measure.

Irrespective of the scenario, whether a company intends to patent its technology or to perform an infringement analysis, or a patent examiner intends to perform a patentability search, several questions arise:

- What are the issued patents in related technologies?
- What is the legal scope of similar patents?
- Who are the competitors?
- Have any similar patents been challenged in court?
- How can one work around existing body of knowledge?
- Are there any scientific literatures, or regulations which can potentially be used to challenge and to invalidate a patent's claims?

Many existing methods attempt to improve IR within a single information source [9,12,14,16,26,27,30,31]. However, these questions cannot be answered from any single information source. An integration framework is needed to enable the retrieval of relevant information from diverse sources. This research explores a knowledge-based approach to address two fundamental information integration issues – (a) the lack of interoperability among the information sources in the current patent system; and (b) the varying information needs by the users of the patent system.

Interoperability between information sources is essential in order to perform multi-source IR [29]. E-government initiatives in the US and Europe are increasingly adopting interoperability frameworks [11]. An important step in achieving interoperability is to allow the information sources to communicate with one another. To achieve this, we propose a Patent System Ontology (PSO) to standardize the representation of the information sources and achieve interoperability. While the documents are vastly diverse, the information is implicitly cross-referenced. For example, a court document which involves a particular patent document reveals a high relevancy between the two documents. Such relevancy is central to our method for multi-source IR and is captured by the PSO.

Terminological variations such as synonymy and polysemy are a common source of problems which often hinder the effectiveness of traditional term based IR methods. We develop a knowledge-based method that uses external knowledge sources such as domain ontologies to provide the required semantics to resolve terminological inconsistencies and improve semantic interoperability between information sources. The IR framework then integrates the patent system ontology and the domain ontologies to retrieve a set of related documents across multiple sources.

# 3. USE CASE AND DOCUMENT CORPUS

We demonstrate our methodology through a use case in the biomedical domain – erythropoietin, a hormone responsible for the production of red blood cells. The synthetic production of erythropoietin has enabled the treatment of chronic diseases such as anemia. Epogen - the production brand of synthetic erythropoietin manufactured by the pharmaceutical giant Amgen Inc. is protected by five core patents namely – US 5,547,933, US 5,618,698, US 5,621,080, US 5,756,349, and US 5,955,422. These patents have been central to many related court cases involving other pharmaceutical companies such as Hoescht Marion Roussel and Transkaryotic Therapies, and heavily cite scientific literature from top journals. Our corpus includes a total of 1150 patent documents downloaded from the USPTO database related to erythropoietin [28]. We identified 135 relevant patents amongst the 1150 patents by following forward and backward citations from the five core patents that will serve as the ground truth. Several court litigations, involving these five patents and some others, date back to the late 1980's. Around 30 U.S. patent litigation documents are collected which are closely related to the use case from Public Access to Court Electronic Records (PACER), an electronic system to access the databases of the 94 District Courts and 13 Courts of Appeals (CAFC) [22]. The repository includes file wrappers for the core patents. The repository also contains the 2007 Text Retrieval Conference (TREC) Genomics data set, which consists of over 162,000 scientific publications from 49 prominent biomedical journals [32]. All in all, the repository contains a diverse set of information from different domains.

# 4. PATENT SYSTEM ONTOLOGY

In this section, we describe a patent system ontology which provides standardized representation and a shared vocabulary of the information sources to facilitate interoperability. The ontology will also provide the required declarative syntax to express multi-source queries, rules, and relevancy metrics.

Many ontology development methodologies have been proposed and implemented over the years [6,8,10,18]. 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. The development process is iterative as the ontology evolves to satisfy the requirements of the application it is being designed for [18]. The resulting ontology is instantiated with actual physical documents from the document repository.

Another practical aspect of ontology development is the specification language in which the ontology will be coded. 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) [21,24]. The factors for choosing a specification language

include expressivity, reasoning capabilities, availability of tools, re-use and personal preference. RDF is a 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 'differentFrom' axioms among others [21]. Several tools have also been developed for the construction and modeling of ontologies such as Protégé and Chimaera [7,23]. 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. However, it should be noted that not all OWL axioms are highly scalable; hence, to the extent possible we make maximum use of the RDF subset of the OWL axioms.

## 4.1 Scope

Ontologies are typically developed with specific applications as targets. Gruninger and Fox suggested that a set of competency questions be developed; these are questions that the ontology is expected to answer [10]. Developing these questions not only helps define the scope of our ontology but also allows us to verify the usefulness of the ontology both throughout and after the development phase [18]. In the patent system domain, the target applications may include patent claim invalidation, and patent infringement analysis. The following are selected examples of competency questions.

*Patent Domain:*

- Return all patent documents which contain the phrase *'recombinant erythropoietin receptor'* in the claims
- Return all the patent documents which contain the phrase *'recombinant erythropoietin receptor'*, at least 3 claims, issued before 02-02-1999 and assigned to Genetics Inc.

*Court Case Domain:*

- Return all court cases which contain the term – *'erythropoietin'*
- Return all court cases which involve the company Amgen Inc. either as the plaintiff or defendant, and from the District Court of Massachusetts

*Multi-domain:*

- Return all patents which contain the term – *'erythropoietin'* in their claims, which are involved in at least one court litigation.
- 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.

Note that the questions can get more complex depending on the requirement of the user. The results of one query can be further re-filtered with additional constraints. In each of the listed questions, the main terms (or objects) are underlined. First, these terms are grouped together into concepts or classes such that they represent a collection of items corresponding to that term. Second, relations are drawn between classes such that the competency questions can be sufficiently expressed as a query using those classes and relationships. The competency questions shown in no way limit the use of the ontology to these applications alone, rather they are examples of questions the ontology must be capable of answering at the minimum. Furthermore, the list of competency questions presented is not meant to be an exhaustive list, but to illustrate how the metadata and text fields parsed from the documents.

Relations in OWL are binary relations, i.e. they can be used to relate exactly two classes, two individuals or an individual to a value. These can be represented in triple form as {subject, predicate, object}. The values that the subject and object take on can be restricted by defining the domain and the range of the relation; where domain refers to the subject end of the relation and range refers to the object end of the relationship [13]. OWL additionally allows us to define logical characteristics such as transitivity and symmetry on these binary relations which enhance the meaning of this relation. For example, if the '=>' relation is defined as a transitive relation, then {A => B} can be used to infer {B => A}. Hence, if properly defined, new knowledge can be derived from existing knowledge. Additionally, we can define necessary and sufficient conditions on classes which can be used to logically classify instances into classes [13].

## 4.2 Conceptualization

Figures 1 and 2 show a conceptual view of the patent and court case documents respectively. The relations between two entities (shown as a black line) are directional from patents and court cases out to other classes, e.g. {Patent, hasTitle, Title}. The relations are not symmetric and hence the inverse {Title, hasTitle, Patent} does not hold true. As shown in the figures, the remaining classes can be grouped under either metadata or textual information. This form of classification helps to address all the metadata at once, instead of individually calling out to each one. For example, if an application requested for all metadata of a patent, using the ontology we can return all metadata entities such as Title, Date, Classification, etc.. We can further group metadata and textual information into a single parent node Information. When the patent and court case hierarchies are combined, classes which are common to both documents will refer to the same concept and not two different concepts.

This form of abstraction is not only possible for classes, but also for relations, made possible by the rdfs:subPropertyOf construct. Court cases and Patents are related to each of the classes shown in Figures 1 and 2. These relations, such as 'hasTitle', 'hasAbstract', and 'hasPlaintiff', etc., can also be abstracted into a common parent relation 'hasInformation'. This relation has a domain of either Patent or Court Case and Information as a range.

File wrappers are not documents themselves, but in fact a collection of documents. This makes modeling file wrappers trickier than the other documents such as patents and court cases. Firstly, a vocabulary of all kinds of documents contained within the file wrapper must be defined. Since each of these documents refers to a particular event of communication between the applicant and patent office, we will call it Event instead of document to avoid confusion between the class Document and a file wrapper event. The events of importance to us are shown in Figure 3. We group application events and office actions separately to allow representation of queries such as – "*Return all office actions for file wrapper A"*. Each file wrapper event must be individually modeled keeping in mind the information it contains. For example, each examiner Rejection contains critical information such as – the allowed claims, the rejected claims, and

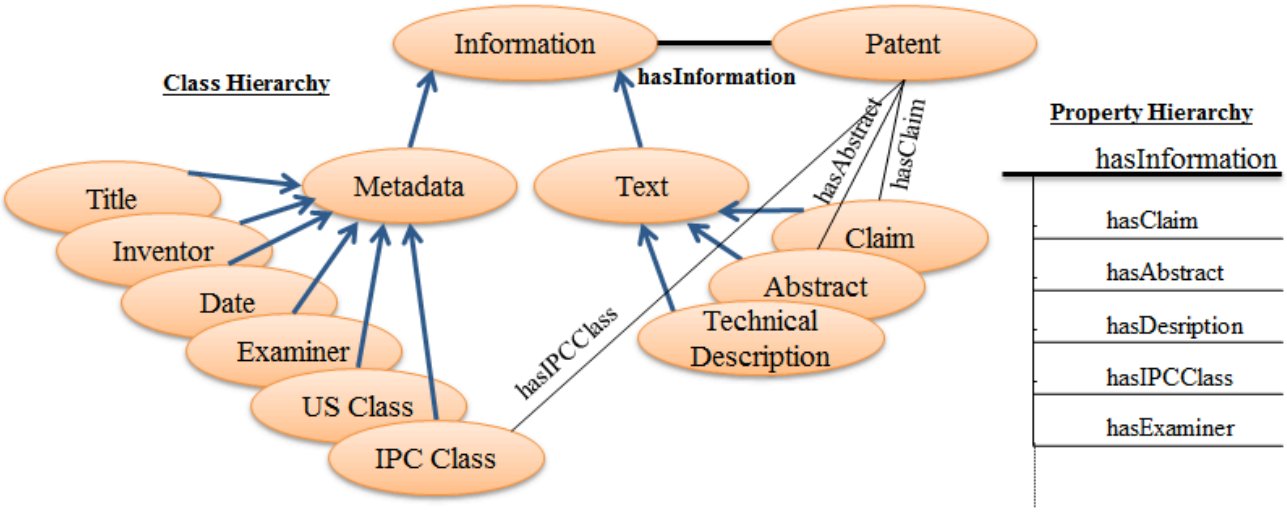


**Figure 1. Conceptual View of Patent Documents**

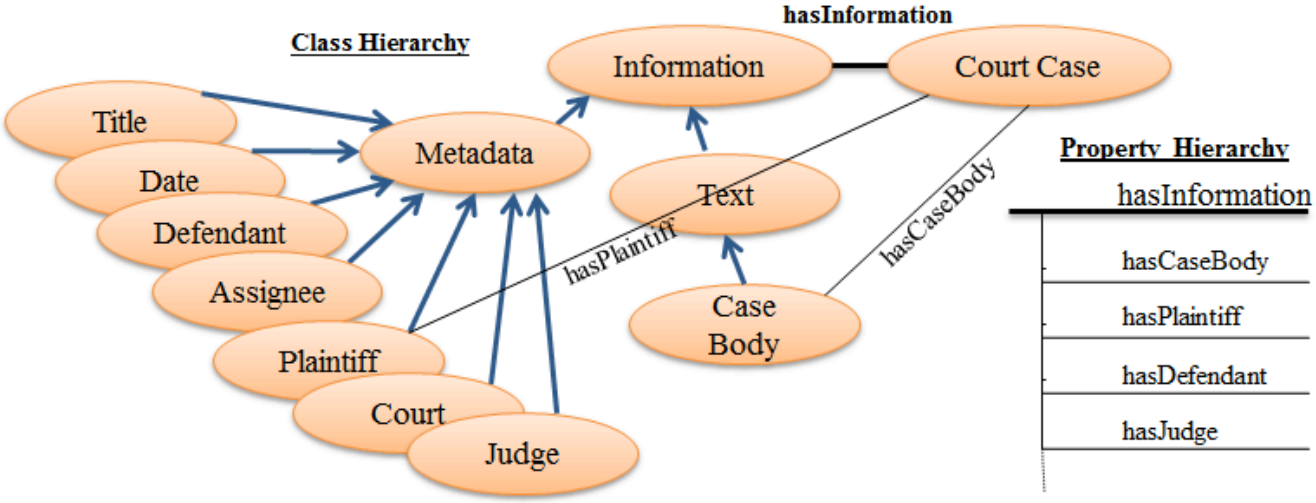


**Figure 2. Conceptual View of Court Case**

the withdrawn claims (see Figure 4). Similarly, other events such as Interference, Restriction, and Amendments can be modeled using the patent system ontology.

The Patent, Court Case, and File Wrapper classes shown in Figures 1-3 are different types of documents available from different information sources. The patent system comprises many such information sources and many such documents. In the top level ontology for the patent system (shown in Figure 5), all types of documents are abstracted into a single parent class (Document).

The Document class can be sub-classed any number of times to include other forms of documents such as regulations and laws which are currently not in the scope of our study. The classes Document, Information, and Event correspond to the three root nodes of the patent system ontology. Additionally, the classes Inventor, Examiner, Author, and Judge, etc., can be abstracted into a common parent node such as Person.

As mentioned earlier, information sources in the patent system implicitly cross-reference one another (see Figure 6). These implicit cross-references show relevancy for comparing documents from different information sources. When manually comparing two documents, these cross-references are rather obvious to the human eye. For example, a human could easily spot a reference to a patent document in the court case. These references can very quickly help identify relevant documents to a user query. The power of the patent system ontology lies in the ability to integrate information across multiple information sources by explicitly expressing such cross-references. Applications built around the patent system ontology can dynamically derive relevancy based on these pre-defined cross-references.

## 4.3 Populating the Ontology

The ontology is populated with information from actual physical documents from the document repository. The instantiation is done automatically using the standard Jena and Protégé Java libraries [23]. Once the instantiation is complete, a standard OWL reasoner such as Pellet is triggered to check for consistency and make inferences [33]. For example, an entity in the class Patent

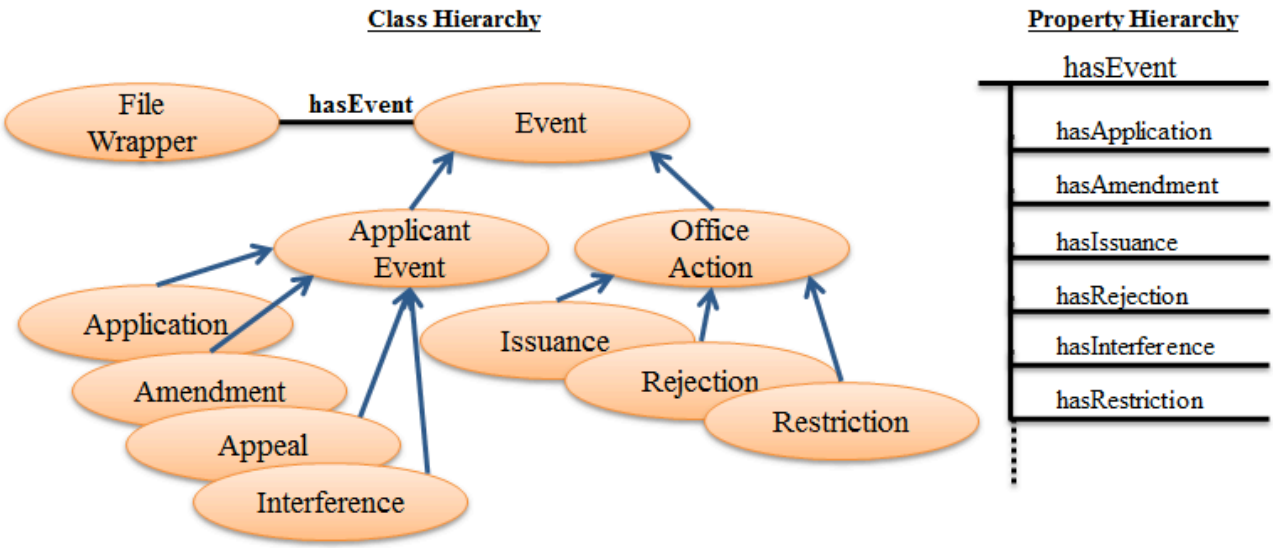


**Figure 3. Events Contained in a File Wrapper**

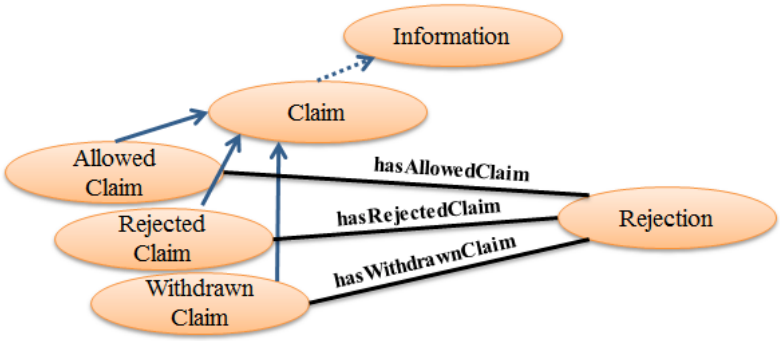


**Figure 4. Excerpt from the Patent System Ontology: Rejection class**

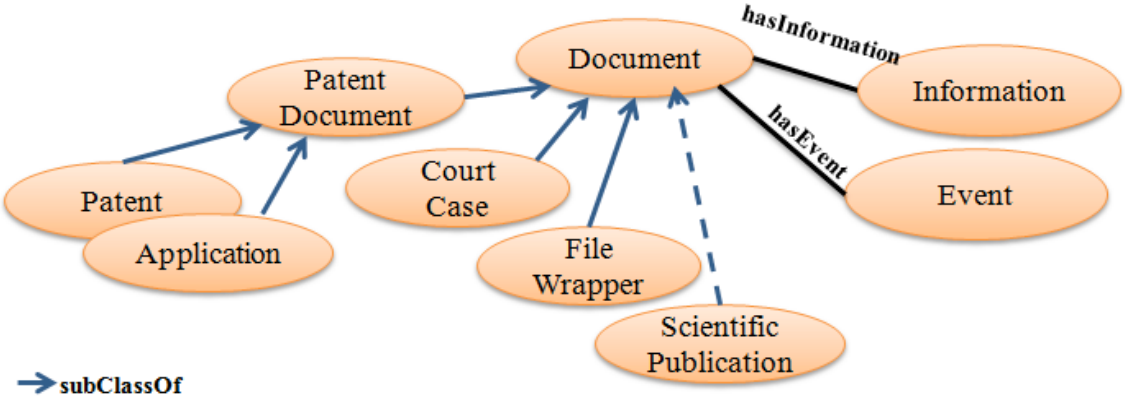


**Figure 5. Top Level Ontology for the Patent System**

will be additionally classified as a Document, since Patent is a subclass of Document. The current version of the knowledge-base is populated with the actual documents from our corpus described in Section 3. Other documents which may have been found in court cases or through patent citations, but not in the corpus are instantiated but contain no information about the patent since the original document itself is unavailable. A file wrapper has also been partially (including the first amendment, rejection, interference and the original application) incorporated into the knowledge-base.

RDF Triple stores are specialized databases to manage large amount of information written in RDF [5,17,20]. Due to the size of the ontology, we create a local instance of a triple store

**Table 1. Expressing Competency Questions in SPARQL**

| Competency Questions | SPARQL Query |
|---|---|
| Return all court cases which involve the company Amgen Inc. as the plaintiff and from the District Court of Massachusetts | SELECT ?case WHERE {<br>?case type CourtCase .<br>?case hasPlaintiff "Amgen Inc." .<br>?case hasCourt "District Court…"<br>} |
| Return all patents which contain the phrase *'recombinant erythropoietin receptor'* in the claims and IPC class "A61K" | SELECT ?pat WHERE {<br>?pat type Patent .<br>?pat hasClaim ?clm .<br>?clm hasTerm "recombinant …" .<br>?pat hasIPCClass "A61K" .<br>} |

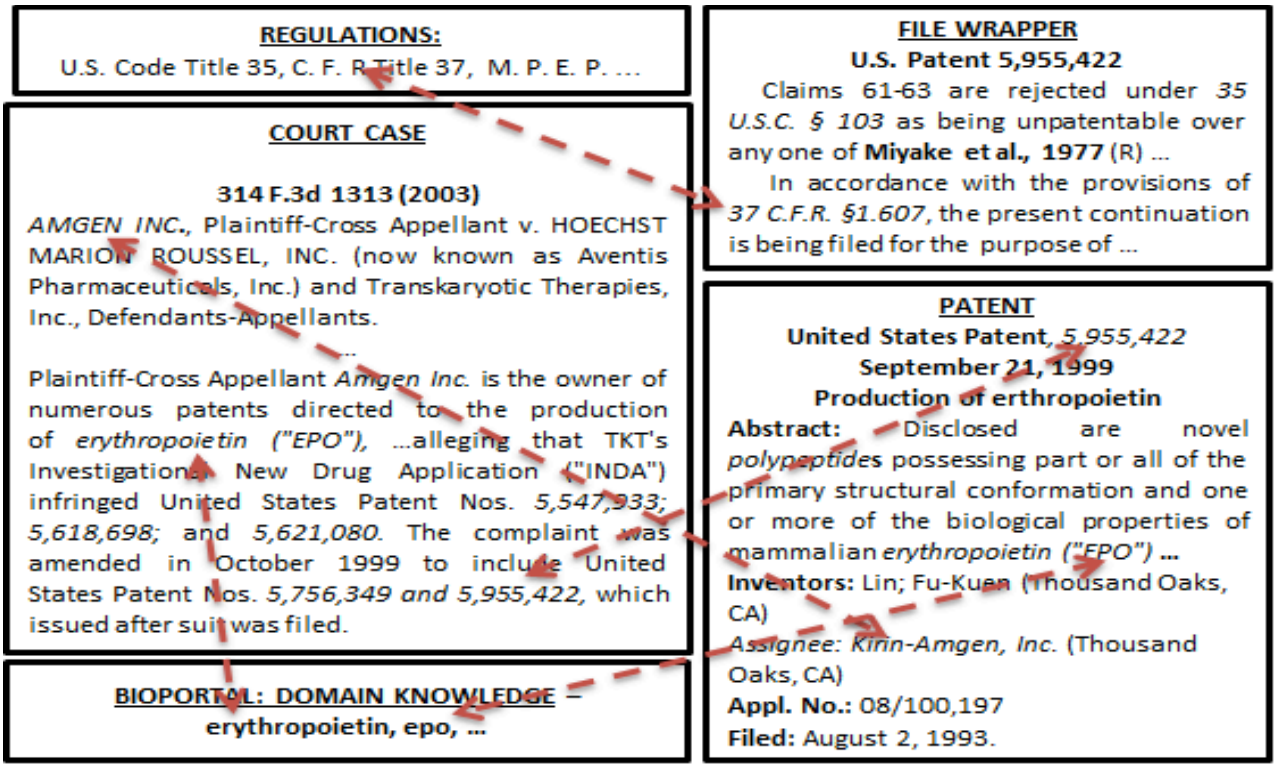


**Figure 6. Cross-Referencing between Documents in the Patent System**

(Virtuoso) and store all the triples in it. Using a triple store will allow us to scale our ontology to millions of instances (documents). Moreover, ontology editors such as Protégé require loading the ontology each time the application is executed. The triple stores provide a persistent store for the triples and significantly lower the loading time. The ontology can be queried using SPARQL through both Protégé and Virtuoso interfaces [20,23].

Table 1 shows examples of how we can represent any natural language question in SPARQL to query the ontology, as long as the classes and relations required to express the query are defined in the ontology. The queries do not always have to return documents, but can return other classes like Inventors or Examiners as well. These SPARQL queries will generally be handled at the application level and will be abstracted from users. Applications can request any information they want from the ontology. In fact, even the applications do not have to fully know the details of the ontology. The ontology can be queried for all its relations for a particular class or between two classes. For example, the query:

```
SELECT ?rel WHERE {
    ?pat   type      Patent   .
    ?pat   ?rel      Information
}
```

will return all relations (variable ?rel) which have the class Patent as the domain. In other words, all relations defined on patents such as hasTitle, hasAbstract, hasIPCClass, etc., will be returned.

Hence, updating the underlying ontology with new information will automatically update the application using it as well.

# 5. INFORMATION RETRIEVAL FRAMEWORK

In information retrieval, the information desired is seldom achieved with a single query. Queries are typically reformulated several times based on intermediate search results until the information need is satisfied [25]. This reformulation could include the addition of synonyms, new search terms, and other constraints. When performing multi-source search, information obtained from searching one domain is applied to another. The patent system ontology provides the backbone for automating this process by standardizing representation of the information sources. In this section, we present an IR framework which builds on top of the semantics of the patent system ontology in multiple stages to enhance multi-source IR (see Figure 7):

Step–I Expand Query: In this stage, the user's initial query is expanded using external knowledge such as dictionaries, thesauri, or domain ontologies. While the patent system ontology provides a framework for the structural interoperability between the information sources, domain ontologies provide semantic interoperability within a specific technical domain. The term expansion is based on several properties of domain ontologies such as abstraction, synonymy, and term mapping, etc..

Step–II Search Information Sources: Information sources are independently searched using the expanded query from Step–I. The required vocabulary and syntax for searching the information sources is contained in the patent system ontology. For example, the patent system ontology provides the syntax for searching the titles of documents – hasTitle:'erythropoietin'. The information sources are searched independently in this stage to retrieve highly relevant documents from each source.

Step–III Cross-Reference Information: The cross-referenced information is highly important for multi-domain retrieval. The cross-references explicitly defined in the patent system ontology are used as relevancy measures to correlate search results between information sources. For example, a relation defined in the patent system ontology – {caseA, patentsInvolved, patentA} will help the framework to extract patent numbers from the court case. These patent numbers can be used to repeat or enhance the search for the patent domain. Similarly, text from one document can be used to search an entirely different information silo. For example,

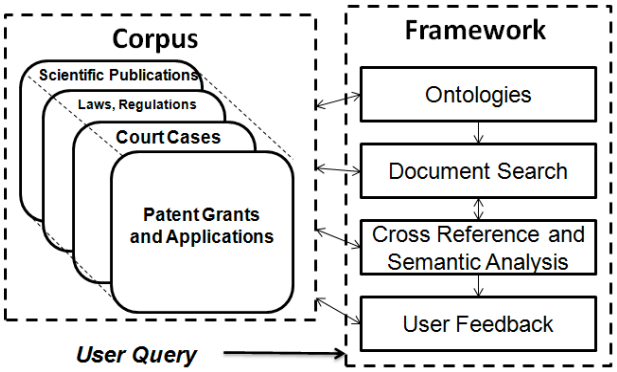


**Figure 7. Information Retrieval Framework**

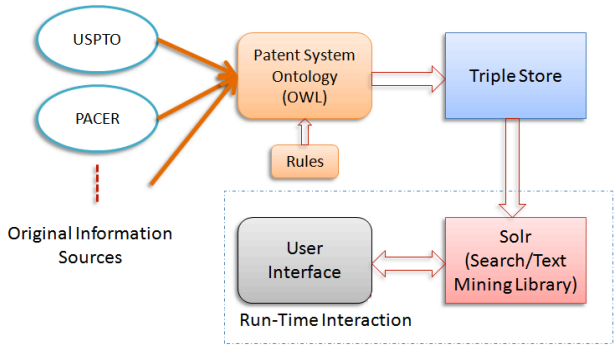


**Figure 8. System Implementation**

the abstract of a patent document can be used as a query to search for relevant scientific publications. In fact, the extracted text can be directly fed back into Steps 1 and 2.

Step–IV User Feedback: Besides the diverse information and knowledge sources, the users in the patent system domain area also come from a diverse background – scientific/technical, legal, business, and more. The intention of the user must be captured through the search process in order to ensure that the results retrieved are indeed relevant to the user. User-relevancy feedback has been an important part of IR research [4,15]. However, the user relevancy feedback stage is out of this paper's scope and will not be discussed.

## 5.1 Implementation Details

In this section, we provide a brief overview of the implementation of the IR framework and its basic features (see Figure 8). The IR framework is implemented entirely in Java with abstractions of several modules that are critical for the system. The actual documents are parsed in order to populate the patent system ontology and perform reasoning using standard Jena libraries. A persistent storage such as Virtuoso or Mulgara is used to store the RDF triples [17,20]. Apache Lucene is a widely used text mining library [2]. In order to provide text-based search, a simple user interface is developed which interacts with the lucene text index. A summary of the implementation is provided below:

- Jena libraries and triple store integration for modifying the patent system ontology through new constructs, cross-references, or rules.
- Solr and Lucene libraries to create, update, and query the text indexes [3].
- Generic API for integration with sources of domain knowledge such as BioPortal [19]
- Automatic query generation, abstracting the syntactic details from the user.

# 6. ILLUSTRATIVE EXAMPLES

As explained in Sections 1 and 2, during the lifetime of a patent, the patent system will constantly be searched for information. Examples include prior art searches, patent claim invalidations, and infringement analysis, etc.. In this section, we demonstrate how the patent system ontology enables queries across multiple sources in order to retrieve related information. Specifically, we focus on Step-III of the framework, i.e., the cross-referencing capabilities of the patent system ontology. We develop a

SPARQL query to search for information related to US patent 5,955,422 (see Figure 9). The individual clauses in the query are grouped into three categories each meant to retrieve related case documents, patents, and scientific publications respectively.

Category I: Court cases provide important information regarding the major competitors, and successful patents, etc.. A patent which has been challenged in court several times is considered very important in its respective technology class. Clauses 1-3 attempt to retrieve related documents by following the cross-references between court cases and patent documents. Court cases related to US patent 5,955,422 are identified by querying the ontology for all court cases that involve the patent. This search retrieves around 20 patent litigations from our corpus. In addition, other patents involved in the court case are also extracted as relevant documents using clause 3.

Category II: After identifying some relevant patents, some of the possible next steps could be to retrieve patents by following the forward and backward citations etc. (clause 4), to get more relevant results. The information extracted from the ontology can also include names of inventors (clause 5), assignees (clause 6), and technology classifications (clause 7) that appear in the patent documents, which can in turn be used to search the documents.

```
SELECT ?pat1 ?pat2 ?case ?pub ?inv ?assg ?class
WHERE {
        Category I:
        Clause 1: ?case  a  CourtCase .
        Clause 2: ?case  patentsInvolved  US5955422.
        Clause 3: ?case patentsInvolved ?pat1

        Category II:
        {Clause 4: ?pat1  hasCitation ?pat2 .}
        {Clause 5: ?pat1  hasInventor  ?inv .}
        {Clause 6: ?pat1  hasAssignee  ?assg .}
        {Clause 7: ?pat1   hasUSClass  ?class .}

        Category III:
        Clause 9:  ?pat  hasClaim  ?claim .
        Clause 10: ?pub a Publication .
        Clause 11: ?pub hasBody  ?body .
        Clause 12: FILTER REGEX (?body, ?claim, "i")
}
```

**Figure 9. SPARQL Query to Retrieve Information Related to U.S. Patent 5,955,422**

**Table 2. Summary of Extracted Information**

| Plaintiffs/Defendants | Patents Involved in Cases | US Class | Inventor | Assignee |
|---|---|---|---|---|
| Amgen Inc. | 5,955,422 | 514/8 | Lin, Fu-Kuen | Kirin-Amgen, Inc. |
| Chugai Pharmaceuticals | 5,547,933 | 530/350 | Hewick, Rodney, M. | Amgen, Inc. |
| Hoescht Marion Roussel | 5,621,080 | 536/23.51 | Seehra, Jasbir, S. | Kiren-Amgen, Inc. |
| Genetics Inc. | 5,618,698 | 435/325 | Seenra, Jasbir, S. | Genetics Institute, Inc. |

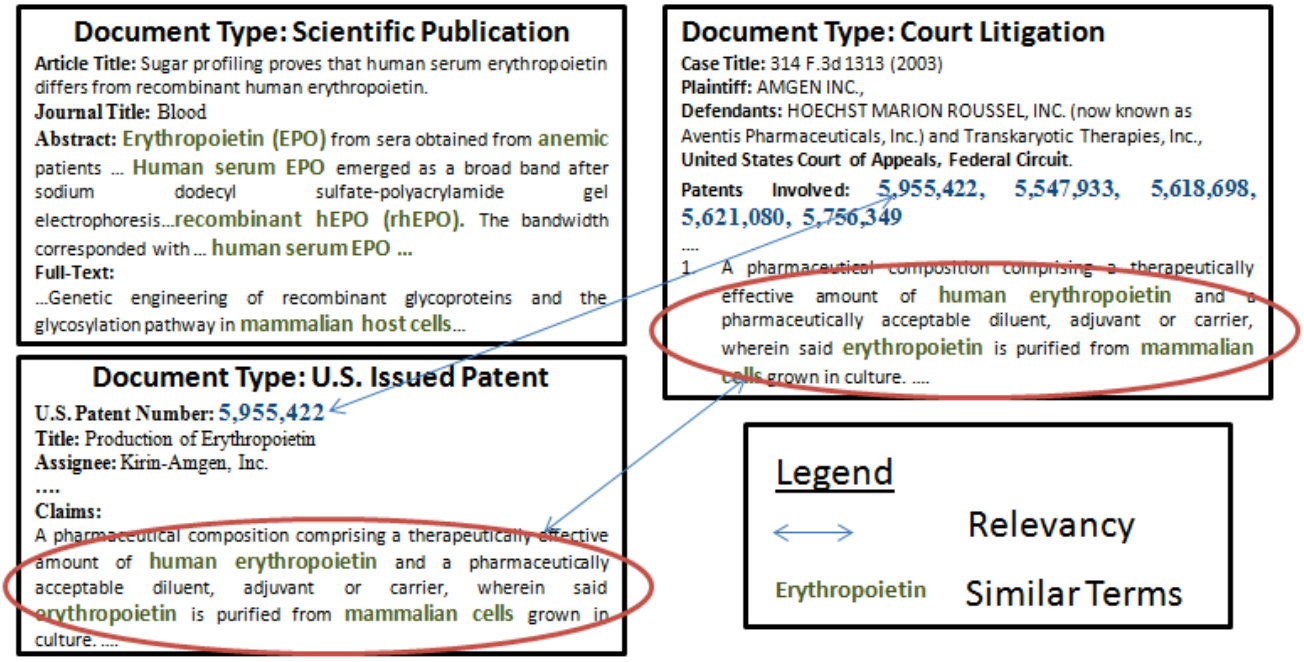


**Figure 10. Actual Documents Retrieved by Querying Patent System Ontology**

**Table 3. Precision of Retrieved Patent Documents Related to a Set of Inventors, Assignees or US Class**

| Query | Precision |
|---|---|
| Top 5 Technology Classes | 0.183 |
| Inventors | 0.8 |
| Assignees | 0.256 |
| Combined | 0.186 |

The extracted information from the documents is summarized in Table 2.

Category III: In addition to metadata such as inventors and assignees, the text of the patent documents can also be used to search for related documents. For example, clauses 9-12 show how the claims of a patent can be used to search for related scientific publications.

The query discussed in this section shows how the semantics provided by the patent system ontology can be used to integrate information across multiple domains, with a potential to improve search. Similar to the above example, in practical applications, SPARQL queries can be formulated to express extremely complex information needs. Figure 10 shows the actual documents retrieved from this query.

Precision, a common metric used to evaluate the quality of IR methodologies, is measured as the ratio of the number of relevant documents retrieved to the total number of documents retrieved. In addition to the query in Figure 9, we retrieve patent documents related to the top inventors, assignees, and technology classes from Table 2. We calculate the precision of the retrieved patent documents with respect to the ground truth discussed in Section 3. Table 3 summarizes these results.

# 7. CONCLUSION

Intellectual Property (IP) related information for science and technology is distributed across several heterogeneous information silos. The scattered distribution of information, combined with the enormous sizes and complexities, make any attempt to collect IP-related information for a particular technology a daunting task. Hence, there is a need for a software framework which facilitates semantic and structural interoperability between the diverse and un-coordinated information sources in the patent system. In this paper, we present a knowledge-based software framework to facilitate retrieval of patents and related information across multiple diverse and uncoordinated information sources in the US patent system. Specifically, we discuss the patent system ontology which provides standardized representation and a shared vocabulary of the information sources to facilitate interoperability.

Through an illustrative example in Section 6, we showed how the patent system ontology can be used to integrate information and query multiple information sources to retrieve related information. The patent system ontology provides the necessary semantics to allow users to develop complex declarative queries. The methodology presented will benefit many end users ranging from lawyers, start-up companies, to large corporations.

# 8. ACKNOWLEDGEMENTS

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] *35 U.S.C. Sec. 103 (United States Code)*. "Conditions for Patentability; Non-Obvious Subject Matter," 2010.

[2] *Apache Lucene*. http://lucene.apache.org/

[3] *Apache Solr*. http://lucene.apache.org/solr/

[4] Baeza-Yates, R. and Ribeiro-Neto. B., *Modern Information Retrieval*, ACM Press, 1999.

[5] Broekstra, J., Kampman, A. and Harmelen, F., V., "Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema", *The Semantic Web – ISWC 2002*, Lecture Notes in Computer Science, 2342:54-68, 2002.

[6] Bruijn, J., D. et al. State-of-the-art Survey on Ontology Merging and Aligning. *V1. SEKT-project report D4.2.1 (WP4)*, IST-2003-506826, 2003.

[7] *Chimaera Website*. http://www.ksl.stanford.edu/software/chimaera (Accessed on 03/01/2012).

[8] De Nicola, A., Missikoff, M. and Navigli, R., "A Software Engineering Approach to Ontology Building," *Information Systems*, 34(2):258-275, 2009.

[9] Fujii, A. Enhancing Patent Retrieval by Citation Analysis. Annual ACM Conference on Research and Development in Information Retrieval*, Proceedings of the 30th annual International ACM SIGIR*, Amsterdam, The Netherlands, 2007.

[10] 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.

[11] Guijarro, L. Interoperability Frameworks and Enterprise Architectures in e-Government Initiatives in Europe and the United States. *Government Information Quarterly, 24, 1*, January 2007, 89-101.

[12] Higuchi, S., Fukui, M., Fujii, A., Ishikawa, T., Iwayama, M., Fujii, A., Kando, N. An Empirical Study on Retrieval Models for Different Document Genres: Patents and Newspaper Articles. *Information Retrieval,* 2003, 251-258.

[13] Horridge, M. A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools. The University of Manchester, March 2011.

[14] Li, Y., Taylor, J., S. Advanced Learning Algorithms for Cross-Language Patent Retrieval and Classification. *Information Processing and Management*, Elsevier, 2007.

[15] Manning, C., D., Raghavan, P. and Schutze, H. *An Introduction to Information Retrieval*. Cambridge University Press, 2009.

[16] Mase, H., Matsubayashi, T., Ogawa, Y., Iwayama, M., and Oshio, T. Proposal of Two-Stage Patent Retrieval Method Considering the Claim Structure. *ACM Transactions on Asian Language Information Processing (TALIP) Archive, 4, 2*, June 2005, 190 – 206.

[17] *Mulgara triplestore*. Available online: http://www.mulgara.org/ (Accessed on 03/01/2012).

[18] Noy, N., F., and McGuinness, D. Ontology Development 101: A Guide to Creating your First Ontology. *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880*, March 2001.

[19] Noy, N., F., Shah, N., H., Whetzel, P., L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D., L., Storey, M., A., Chute, C., G. and Musen, M., A., "BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse," *Nucl. Acids Res., 37(2)*:W170-W173, 2009.

[20] *OpenLink Virtuoso*. http://virtuoso.openlinksw.com/ (Accessed on 03/01/2012).

[21] Dean, M. and Schreiber, G. (Eds.). *OWL Web Ontology Language Reference.* W3C Recommendation, 10 February 2004.

[22] *PACER*. http://www.pacer.gov/ (Accessed on 03/01/2012).

[23] *Protégé Website.* http://protege.stanford.edu/ (Accessed on 03/01/2012).

[24] *Resource Description Framework (RDF) Model and Syntax*, W3C Recommendation, 22 February 1999.

[25] Spink, A. A User-Centered Approach to Evaluating Human Interaction with Web Search Engines: An Exploratory Study. *Information Processing and Management, 38, 3*, May 2002, 401-426.

[26] Takaki, T. Associative Document Retrieval by Query Subtopic Analysis and its Application to Invalidity Patent Search. Conference on Information and Knowledge Management. *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, Washington, D.C., USA, 2004, 399 - 405.

[27] Tseng, H., Lin, C., J., Lin, Y., I. Text Mining Techniques for Patent Analysis. *Information Processing & Management,* 2007, Elsevier.

[28] *USPTO*. http://www.uspto.gov/ (Accessed on 03/01/2012).

[29] Wache, H., Vogele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann H., and Hubner, S. Ontology-Based Integration of Information - A Survey of Existing Approaches. In *Proceedings of IJCAI-01 Workshop: Ontologies and Information Sharing*, Seattle, WA, , 2001, 108-117.

[30] 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, 30, 1*, March 2008, 21-23.

[31] Xue, X., Croft, W., B. Automatic Query Generation for Patent Search. Conference on Information and Knowledge Management, *Proceeding of the 18th ACM Conference on Information and Knowledge Management*, Hong Kong, China, 2009.

[32] Hersh, W. and Voorhees, E. TREC Genomics Special Issue Overview. *Information Retrieval, Special Issue on TREC Genomics Track: Guest Editor: Ellen Voorhees, 12, 1*, 2009, 1-15.

[33] Sirin, E., Parsia, B., Grau, B., C., Kalyanpur, A., and Katz, Y. Pellet: A Practical OWL-DL Reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web, 5, 2*, June 2007, 51-53.

[34] Gruber, R., T. Toward Principles for the Design of Ontologies used for Knowledge Sharing. *Int. J. Hum-Comput. Stud, 43, 5-6,* November 1995, 907–928.

[35] Damiani, E., Fugazza, C. Toward Semantics-Aware Management of Intellectual Property Rights. *Online Information Review, 31, 1*, 2007, 59 – 72.