

A.T.E.I. OF THESSALONIKI - DEPARTMENT OF  
INFORMATION TECHNOLOGY

Development of a standard web search application  
for patents

(Ανάπτυξη πρότυπης διαδικτυακής εφαρμογής  
αναζήτησης σε πατέντες)

**Bachelor Thesis of**

Georgios Kotsios-Kontokotsios

**Supervisor Professor**

Michail Salampasis

**Thessaloniki, September 2012**

## Contents

Introduction.....	5
Chapter 1 <sup>st</sup> : Introduction .....	6
1.1. Motive .....	6
1.2. Objectives .....	6
1.3. Structure.....	7
Chapter 2 <sup>nd</sup> : Patents .....	8
Introduction.....	8
2.1. Before Applying For a Patent.....	8
2.2. Application.....	9
2.3. Filing and Formalities Examination.....	9
2.4. Search .....	10
2.5. Publication of the Application .....	10
2.6. Substantive Examination .....	10
2.7. The Grant of a Patent .....	11
2.8. Validation.....	11
2.9. Opposition .....	11
2.10. Limitation/Revocation .....	11
2.11. Appeal.....	12
Chapter 3 <sup>rd</sup> : Patent Search .....	13
3.1. Prior-Art.....	13
3.2. Validity/Invalidity.....	13
3.3. Freedom To Operate .....	14
3.4. Technology Landscape .....	14
3.5. Novelty Search.....	14
3.6. Search systems .....	15
Chapter 4 <sup>th</sup> : Federated Search & ezDL .....	17
4.1. How Content Is Accessed .....	17
4.2. How Typical Web Search Engines Work .....	17
4.3. Differences in Federated Search .....	18
4.4. Benefits of Federated Search .....	18
4.5. Quality of Results.....	18
4.6. Most Current Content .....	19
4.7. Federated Search in Depth.....	19

4.8.	Important Features.....	20
4.9.	EzDL - An Interactive Search System .....	21
4.10.	Architecture.....	22
4.11.	The Backend .....	22
4.12.	The Frontend .....	23
Chapter 5 <sup>th</sup> : The PerFedPat project.....		24
5.1.	Introduction.....	24
5.2.	Patent Data Sources .....	24
5.3.	New Query Form Component .....	25
5.4.	Data Representation.....	26
5.5.	Integration of Tools .....	27
Chapter 6 <sup>th</sup> : Technologies.....		29
6.1.	Introduction.....	29
6.2.	Repositories.....	29
6.3.	Lucene/Solr.....	30
6.4.	Java .....	30
6.5.	XML/HTML Parsing .....	31
Chapter 7 <sup>th</sup> : Conclusions.....		32
7.1.	Overall Conclusions .....	32
7.2.	Applications Of iPerFedPat.....	32
7.3.	Future Work.....	32
Publication References.....		33

## **Index of images**

Image 1 - The overall architecture of iPerFedPat .....	22
Image 2 - ezDL bibliographic query form .....	25
Image 3 - iPerFedPat patent query form .....	26
Image 4 - IPC Suggestions for Titanium .....	27
Image 5 - Entities extration from current search .....	28
Image 6 - Result clustering for current search.....	28

## **Index of tables**

Table 1 - Free patent search systems .....	15
Table 2 - Fee based patent search systems .....	16

# Introduction

The preparation of my thesis was a very interesting experience and a great opportunity for me to fathom computer programming, which I hope will be the trigger for a successful career in this field. During the development and completion of my diploma thesis, a variety of topics were investigated.

First and foremost I had to understand how the patent industry works and specifically I focused on what a patent actually is and what the application and publication processes include.

Also the thesis aims to research into a new generation of advanced patent search systems for the patent related industries and the whole spectrum of patent users by designing a new exciting framework for integrating multiple patent data sources, patent search tools and user interfaces. The actual goal application is based on an open source project called ezDL, which started from the University of Duisburg Essen. Information about ezDL and the way it works are going to be given at a later stage.

In the final chapters I describe all the different technologies and platforms that were used and the experience I gained during the development of the system.

# Chapter 1<sup>st</sup>: Introduction

## 1.1. Motive

Patent search is an economically important problem, central to the R&D operations of many industries including pharmaceuticals, biotechnology, automotive and many more. Besides the economic interest, from a technological perspective, patent search reveals important challenges for the field of information access. Even though there is a common number of important characteristics with web search some important differences exist, like lengthy search sessions, demand for high recall and high value documents. This thesis aims to research into a new generation of advanced patent search systems for the patent related industries and the whole spectrum of patent users by designing a new exciting framework for integrating multiple patent data sources, patent search tools and UIs.

## 1.2. Objectives

The iPerFedPat system, which will be the main result of this research, will have a pluggable architecture, providing core services and operations being able to search multiple patent data sources and streams, thus providing multiple patent search UIs while hiding complexity from the end user. The major objectives of this research are:

- Design a pluggable framework for federated multi-lingual search of large-scale patent information.
- Develop new algorithms (e.g. for source selection, results merging, personalized results presentation) and integrate these algorithms to components of such a pluggable framework based on existing open-source components.
- Achieve sufficient conceptual integration between approaches of heterogeneous fields (distributed information retrieval, human computer interaction, machine learning and semantic web) to enable the seamless integration of iPerFedPat components based on methods from these diverse fields.

Evaluate the effectiveness of the reference implementation through an application which will address the needs of real patent users.

### 1.3. Structure

Chapter 1 is an introductory. Chapter 2 revolves around the patent industry and the way it operates. There is a description of what a patent actually is, stages an idea goes through in order to be applied and published and an inside view of the professionals who have been assigned this task. Chapter 3 describes ezDL, the platform that what used as a base for the iPerFedPat application and Federated search, the information retrieval technology behind it. Chapter 4 provides extensive information, both theoretical and technical regarding iPerFedPat. Chapter 5 offers an inside view on the technologies that made this thesis-project possible.

# Chapter 2<sup>nd</sup>: Patents

## Introduction

Patent search is an economically important problem, central to the R&D operations of many industries including pharmaceuticals, biotechnology, automotive and many more. Besides the economic interest, from a technological perspective, patent search reveals important challenges for the field of information access. Even though there is a common number of important characteristics with web search some important differences exist, like lengthy search sessions, demand for high recall and high value documents. This thesis aims to research into a new generation of advanced patent search systems for the patent related industries and the whole spectrum of patent users by designing a new exciting framework for integrating multiple patent data sources, patent search tools and UIs.

## 2.1. Before Applying For a Patent

First, it is important to know what inventions and patents are. An invention can be, for example, a product, a process or an apparatus. To be patentable, it must be new, industrially applicable and involve an inventive step. Patents are valid in individual countries for specified periods. They are generally granted by a national patent office, or a regional one like the EPO. Patents confer the right to prevent third parties from making, using or selling the invention without their owners' consent.

Patents should not be confused with the other kinds of intellectual property rights available:

- Utility models can be registered in some countries, to protect technical innovations which might not qualify for a patent
- Copyright protects creative and artistic works such as literary texts, musical compositions and broadcasts against unauthorized copying and certain other uses
- Trademarks are distinctive signs identifying brands of products or services; they may be made up of two- or three-dimensional components such as letters, numbers, words, shapes, logos or pictures, or even sounds
- Designs and models protect a product's visual appearance, i.e. its shape, contours or color.



## 2.2. Application

There are different routes to patent protection and the best route for you will depend on your invention and the markets your company operates in. The European Patent Office accepts applications under the European Patent Convention (EPC) and the Patent Cooperation Treaty (PCT). If you are seeking protection in only a few countries, it may be best to apply direct for a national patent to each of the national offices.

A European patent application consists of:

- a request for grant
- a description of the invention
- claims
- drawings (if any)
- an abstract.

Applications can be filed at the EPO in any language. However, the official languages of the EPO are English, French and German. If the application is not filed in one of these languages, a translation has to be submitted. Although the services of a professional representative are mandatory only for applicants residing outside Europe, the EPO advises all applicants to seek legal advice.

## 2.3. Filing and Formalities Examination

The first step in the European patent granting procedure is the examination on filing. This involves checking whether all the necessary information and documentation has been provided, so that the application can be accorded a filing date.

The following are required:

- an indication that a European patent is sought
- particulars identifying the applicant
- a description of the invention or
- a reference to a previously filed application.

If no claims are filed, they need to be submitted within two months. This is followed by a formalities examination relating to certain formal aspects of the application, including the form and content of the request for grant, drawings and abstract, the designation of the inventor, the appointment of a professional representative, the necessary translations and the fees due.

## 2.4. Search

While the formalities examination is being carried out, a European search report is drawn up, listing all the documents available to the Office that may be relevant to assessing novelty and inventive step. The search report is based on the patent claims but also takes into account the description and any drawings. Immediately after it has been drawn up, the search report is sent to the applicant together with a copy of any cited documents and an initial opinion as to whether the claimed invention and the application meet the requirements of the European Patent Convention.

## 2.5. Publication of the Application

The application is published - normally together with the search report - 18 months after the date of filing or, if priority was claimed, the priority date. Applicants then have six months to decide whether or not to pursue their application by requesting substantive examination. Alternatively, an applicant who has requested examination already will be invited to confirm whether the application should proceed. Within the same time limit the applicant must pay the appropriate designation fee and, if applicable, the extension fees. From the date of publication, a European patent application confers provisional protection on the invention in the states designated in the application. However, depending on the relevant national law, it may be necessary to file a translation of the claims with the patent office in question and have this translation published.

## 2.6. Substantive Examination

After the request for examination has been made, the European Patent Office examines whether the European patent application and the invention meet the requirements of the European Patent Convention and whether a patent can be granted. An examining division normally consists of three examiners, one of whom maintains contact with the applicant or representative. The decision on the application is taken by the examining division as a whole in order to ensure maximum objectivity.

## 2.7. The Grant of a Patent

If the examining division decides that a patent can be granted, it issues a decision to that effect. A mention of the grant is published in the European Patent Bulletin once the translations of the claims have been filed and the fees for grant and publication have been paid. The decision to grant takes effect on the date of publication. The granted European patent is a "bundle" of individual national patents.

## 2.8. Validation

Once the mention of the grant is published, the patent has to be validated in each of the designated states within a specific time limit to retain its protective effect and be enforceable against infringers. In a number of contracting states, the patent owner may have to file a translation of the specification in an official language of the national patent office. Depending on the relevant national law, the applicant may also have to pay fees by a certain date.

## 2.9. Opposition

After the European patent has been granted, it may be opposed by third parties – usually the applicant's competitors – if they believe that it should not have been granted. This could be on the grounds, for example, that the invention lacks novelty or does not involve an inventive step. Notice of opposition can only be filed within nine months of the grant being mentioned in the European Patent Bulletin. Oppositions are dealt with by opposition divisions, which are normally made up of three examiners.

## 2.10. Limitation/Revocation

This stage may also consist of revocation or limitation proceedings initiated by the patent proprietor himself. At any time after the grant of the patent, the patent proprietor may request the revocation or limitation of his patent. The decision to limit or to revoke the European patent takes effect on the date on which it is published in the European Patent Bulletin and applies from the beginning to all contracting states in respect of which the patent was granted.

## 2.11. Appeal

Decisions of the European Patent Office – refusing an application or in opposition cases, for example – are open to appeal. Decisions on appeals are taken by the independent boards of appeal. In certain cases it may be possible to file a petition for review by the Enlarged Board of Appeal.

## Chapter 3<sup>rd</sup>: Patent Search

### 3.1. Prior-Art

Prior art is any body of knowledge that relates to your invention. A patent search is part of your search for prior art. Prior art would include previous patents, trade journal articles, publications (including data books and catalogs), public discussions, trade shows, or public use or sales anywhere in the world and helps prove the novel and nonobvious legal conditions that are required for a patent to be granted. Thus, a prior art search will help distinguish between what is already known (prior art) and what is new (invention). The secondary benefit of a prior art search is that an inventor can also use such a search to understand the prevailing state of art in his field of research. This will give an idea as to how the future scope of research could be. Also when an organization invests large sums of money in Research and Development activities, it seldom verifies if the technology it wants to develop already exists and if it is owned by someone else. To know what has been developed before you initiate your work you need to perform a prior art search to detect all existing similar developments or inventions.

### 3.2. Validity/Invalidity

The defense of invalidity argues that a patent should not have been issued as a patent in the first place because the invention is not novel or is obvious. One example of patent invalidity would be where the defendant can show a printed publication that completely describes the invention before the invention date of the patentee. This defense is usually more difficult to prove than noninfringement, because the patentee is given a presumption of validity on the patent once it issues.

A Validity Search is used to determine whether a patent can be invalidated because the invention was not novel and inventive when the patent was granted. For this reason, a Validity Search is also known as an Invalidity Search. Validity Search is different from a Patentability Search which is conducted before you take out a patent to establish the novelty of the invention. A Validity Search is carried out once a patent has been granted to test whether the invention truly satisfied the novelty provisions of the patent application process. If prior art can be discovered that was missed during examination by the Patent Office, the patent can be invalidated.

### 3.3. Freedom To Operate

Freedom to operate (FTO) is usually used to mean determining whether a particular action, such as testing or commercializing a product, can be done without infringing valid intellectual property rights of others. Freedom to Operate from a patent perspective means that you have established – with a reasonable certainty – that your product does not infringe the Intellectual Property rights of others. We say "with a reasonable certainty" because "freedom to operate" can never be determined with absolute certainty due to inherent features of the patent system. The first step in establishing FTO is to conduct a Clearance Search or Infringement Search to locate granted patents, or patent applications (which upon grant) determine whether your product would infringe. Most companies will engage a reputable IP Analytics firm to do this.

### 3.4. Technology Landscape

Patent landscapes describe the patent situation for a specific technology in a given country, region or on the global level. They usually start with a state-of-the-art search for the technology of interest in suitable patent databases. The results of the search are then analyzed to answer specific questions, e.g. to identify certain patterns of patenting activity or certain patterns of innovation (innovation trends, diversity of solutions for a technical problem, collaborations). An essential component of each patent landscape report is the visualization of these results in order to facilitate their understanding, and certain conclusions or recommendations based on the empirical evidence provided by the search and analysis. Patent landscapes can therefore be useful for policy discussions, strategic research planning or technology transfer. However, they provide only a snapshot of the patenting situation at a certain point in time.

### 3.5. Novelty Search

A Patent Novelty Search or Patentability Search is a Prior Art Search conducted before a patent application is prepared. This search will determine whether anyone else publicly disclosed the inventive concept prior to its critical date and provides a host of other advantage. Specifically, novelty is one of the requirements of a patent and if the patent is published before the application date or before the priority, if the patent requires priority, it will lose novelty. In some countries, such as China, America and Japan, if the inventor or its successor publishes the inventions before application date, they will gain a grace period. It is said that if the inventor or its successor has published the inventions, then he or she still can apply for this patent with novelty, assuming that the application

date will be within the grace period. The grace period of most countries is six to twelve months. Sometimes the limit of this type of novelty can also be called relative novelty. In some other countries, including majority of European countries, any invention makes an oral or writing publication, exposition or open for use before application for patent, no matter who or where it is used or published, the invention will lose its novelty and it won't gain certificate of patent. This kind of rule is called absolute novelty.

### 3.6. Search systems

There are many free and fee-based search tools available today. Selecting a search tool is usually based on data coverage, pricing, usability, and other features. A big set of tools exist ranging from specialized search tools that aid in chemical, genetic, mechanical, electronic, and other technology areas. All the available tools provide an important service because they are able to access huge amounts of data but in the end the experience level of a patent researcher is what makes the difference in providing reliable search results. The most popular systems are displayed in tables 1 and 2.

System Data	Espacenet	FreePatentsOnline	Google Patent Search	Patent Lens	SumoBrain	Surf-IP
Owner Name	European Patent Office (EPO)	Free Patents Online	Google	Cambia	Patents Online, LLC	Intellectual Property Office of Singapore
Full Text: Patent Authority Coverage	EP, WO/PCT	US, EP, WO/PCT	US	US, EP, WO/PCT, AU	US, EP, WO/PCT	US, WO/PCT
Current US Class	No	Yes	No	No	Yes	No
Original US Class	No	No	Yes	No	No	No
IPC - R	Yes	Yes	No	No	Yes	Yes
Original IPC data (v1-v7)	Yes	No	Yes	No	No	Yes
ECLA	Yes	No	No	No	No	No
Japanese File Index Terms	No	No	No	No	No	No
Japanese F-Terms	No	No	No	No	No	No
Other National Classification Systems	No	N/A	No	No	N/A	No

Table 1 - Free patent search systems

System Data	EAST	PatBase	PatBase Express	QPat	SureChem	WIPS Global
<b>Owner Name</b>	United States Patent and Trademark Office (USPTO)	Minesoft Ltd; RWS Group	Minesoft Ltd; RWS Group	Questel-Orbit	Macmillan Publishers Ltd.	WIPS Global
<b>Full Text: Patent Authority Coverage</b>	US	US, EP, WO/PCT, JP, BE, BR, CH, CN, DE, DK, ES, FI, FR, GB, IN, KR, SE, TW	US, EP, WO/PCT, JP, BE, BR, CH, CN, DE, DK, ES, FI, FR, GB, IN, KR, SE, TW	US, EP, WO/PCT, JP, AT, BE, BR, CA, CH, CN, DE, DK, ES, FI, FR, GB, IN, RU, SE, SU, TW	US, EP, WO/PCT	US, EP
<b>Current US Class</b>	Yes	Yes	Yes	Yes	Yes	Yes
<b>Original US Class</b>	Yes	No	No	Yes	No	Yes
<b>IPC - R</b>	Yes	Yes	Yes	Yes	Yes	Yes
<b>Original IPC data (v1-v7)</b>	Yes	Yes	Yes	Yes	No	No
<b>ECLA</b>	Yes	Yes	Yes	Yes	No	Yes
<b>Japanese File Index Terms</b>	No	Yes	No	Yes	No	Yes
<b>Japanese F-Terms</b>	No	Yes	No	Yes	No	Yes
<b>Other National Classification Systems</b>	Yes	Yes	No	Yes	No	No

Table 2 - Fee based patent search systems



## Chapter 4<sup>th</sup>: Federated Search & ezDL

### 4.1. How Content Is Accessed

Federated search facilitates research by helping users find high-quality documents in more specialized or remote corners of the Internet. Federated search applications excel at finding scientific, technical, and legal documents whether they live in free public sites or in subscription sites. This makes federated search a vital technology for students and professional researchers. For this reason, many libraries and corporate research departments provide federated search applications to their students and staff.

### 4.2. How Typical Web Search Engines Work

There are two basic approaches to finding content on the Web. The approach that all major search engines use is to “crawl” the Web. Over many years, search engines, has amassed a list of billions of Web sites. In the early days, it’s likely that owners registered their sites with them. Today, search engines can find new Web sites through links from sites they already know about and periodically visit the sites on their list and identify the links at that site. Then they follow each link they find to arrive at other pages where they start the process over to find more links. In doing this search engines discover sites they didn’t know about during previous visits. This process of going from one page to another and then to another is referred to as “crawling,” just like a spider crawls from one thread to another in its web. In fact, Web “spiders” are commonly referred to as “Web crawlers.” When you create a new site, just create a link to it from another site, or get someone to do it for you, and a Web crawler will discover it. The trouble with crawling is that this search technique doesn’t find everything. One might believe that through sufficient crawling, one could find all Web pages. In fact, only a small percentage of the Web’s content is accessible to search engines. The term “deep Web” refers to the vast portion of the Web that is beyond the reach of the typical “surface Web” crawlers. Surface Web search engines can’t easily fathom the deep Web because most deep Web content has no links to it.

A very nice example is the following:

If someone wants to research the effects of some chemical or hazardous substance on humans, it would be better to search the National Library of Medicine’s Toxicology Data Network. Most of the information someone would find there would not be available on a search engine. That’s because, in order to find these research

articles, someone must type one or more words in a search box and click on the “search” button. Almost none of these articles have links towards them from a Web site and that’s why it’s not possible to find those articles on a search engine. Search engines are not designed to fill out search forms and click “submit” the way humans do. In particular, search engines wouldn’t know what search words to put into the form and even if they did know what to enter into search forms and how to submit them, they wouldn’t be able to retrieve all of the documents from the source. This leaves search engines with very incomplete content from deep Web sources.

### 4.3. Differences in Federated Search

In most cases search engines doesn’t fill out search forms, while this is exactly what federated search applications do. That’s because it turns out that filling out forms is a difficult problem. Federated search engine builders have to customize their search software for each Web form they encounter. While search engines have a general approach to crawling links from any Web site, federated search engines are programmed with intimate knowledge of each search form. The specialized software must know not only how to fill out the form and how to simulate the pressing of the “search” button, but also how to read the results that the Toxicology Data Network (as in the example above), or any other source, provides. Both are difficult to do well.

### 4.4. Benefits of Federated Search

The essential benefits of federated search to its users include efficiency, quality of search results, and current, relevant content. Also, using a federated search engine can be a huge time saver for researchers. Instead of needing to search many sources, one at a time, the federated search engine performs the many searches on the user’s behalf. While federated search engines specialize in finding content that requires form submissions to retrieve, it isn’t the only criterion for being a federated search engine. A federated search engine also associates content from different sources. Federated search uses just one search form to cover numerous sources, and combines the results into a single results page.

### 4.5. Quality of Results

Federated search engines show their value best in environments in which the quality of results matters, such as libraries, corporate research environments, and the federal government. In the case of the federal government, the constituents of the government benefit greatly from such applications. A major difference between a

federated search engine and a standard search engine is that the client who contracts for the federated search service selects the sources to search. In almost every case, the sources will be authoritative. Search engines, on the other hand, have very minimal criteria for source selection. If a Web page doesn't look like outright junk (i.e., spam) they will present it among the search results. Thus, the federated search engine acts as a helpful librarian does, directing users to excellent quality.

#### 4.6. Most Current Content

In addition to filling out forms and combining documents from multiple sources, another important benefit of federated search engines is that they search content in real time. Real time data is crucial for researchers who are searching for up-to-the-minute content or for content which changes frequently. As soon as the content owner updates their source, the information is available to the searcher on the very next query. By contrast, with standard search engines, the results are only as current as the last time they crawled sites with content that matches your search words. Content you find via search engines might be days or weeks old, which can be fine depending on your situation, but can be problematic if you want the most current information.

#### 4.7. Federated Search in Depth

Federated search is the process of performing a simultaneous real-time search of multiple diverse and distributed sources from a single search page, with the federated search engine acting as intermediary.

The key words in this definition and their influence on the value of federated search is as follows:

- federated - Content is combined from different sources saving the effort of searching sources one at a time.
- simultaneous - Federated search queries all user-selected sources at once. It would be unacceptably slow if it waited for all of the results from one source before querying the next.
- real-time - Federated search occurs live and results are current. There's no stale content.
- multiple - The value of federated search to the researcher increases as the number of sources increases.
- diverse sources - Federated search engines typically can search sources containing documents of different types, e.g. PDF, Word, PowerPoint. The

process of extracting text from documents of different types is hidden from the user.

- distributed sources - Federated search engines expect to search content that lives in different locations.
- single search page - Federated search engines provide a single point of searching.
- federated search engine acting as intermediary - The federated search paradigm is such that the user doesn't communicate directly with the content sources when performing searches. The user submits a search to the federated search engine which, in turn, submits the search to each of the content sources. Each content source provides its results to the federated search engine which combines all of the results from all the sources into a single page of results. Note that federated search was developed independently of the Web, and therefore federated search engines need not be Web-based.

Lastly federated search goes by a number of different names. Metasearch, distributed search, directed search, broadcast search, deep web search, cross-database search, and universal search are often, but not always, used synonymously with "federated search." Metasearch is a term that is often used to refer to a search engine that searches other major search engines. Dogpile, for example, is dedicated to searching the four big search engines: Google, Yahoo!, Bing and Yandex. Some would argue that metasearch engines aren't federated search engines because, even though they search the underlying search engines in real time, the underlying search engines may not have the most current information since they themselves are "crawlers."

#### 4.8. Important Features

Three very important features that are very popular with federated search engines are aggregation, ranking, and de-duplication. The definition of these features is as follows:

- Aggregation - Aggregation is the process of combining search results from the different sources in some helpful way. A federated search engine might present all of the results from one source then, beneath those results, present the results from the next source, and so on. Aggregation may incorporate sorting (e.g., by date, title, or author), or it may involve ranking, also known as relevance ranking.
- Ranking - A researcher searching a couple of dozen sources via a federated search engine usually wants to know which results are most

relevant to his or her search from among all of the sources. Relevance ranking compares results from all sources against one another and displays the results in order. Surprisingly, not all federated search engines rank their results. This is largely because ranking is difficult to perform well.

- De-duplication - A federated search engine may retrieve the same result or document from multiple sources. Users are not interested in seeing duplicate results, yet it turns out to be difficult to remove duplicates effectively. Two documents may have the same title and author, but might actually be different revisions of one document.

#### 4.9. EzDL - An Interactive Search System

ezDL is an open-source software for building highly interactive search user interfaces with strategic support. It builds on the ideas developed and implemented within the Daffodil<sup>1</sup> project from 2000 to 2009, but uses more modern software technologies and interface design methods. The ezDL framework can be characterized by three main purposes. It is foremost

- a working interactive tool for searching a heterogeneous collection of digital libraries. In addition to that, it is
- a flexible software platform providing a solid base for writing customized applications as well as
- a system that can be used for many different types of user evaluations.

Today many systems covering one or more aspects of ezDL exist but unifying them into one single framework (like the ezDL project did) is unique.

---

<sup>1</sup> Daffodil is a digital library system targeting at strategic support during the information search process.

## 4.10. Architecture

ezDL is a continuation of the Daffodil project and therefore shares its main ideas: meta-search in digital libraries and strategic support for users. Its overall architecture likewise has inherited many features from Daffodil. Figure 2 provides a high-level overview of the system. The system architecture makes extensive use of separation of concerns to keep interdependencies to a minimum and make the system more stable. This is true on the system level where a clear separation exists between clients and backend, but also within the backend itself, where individual "agent" processes handle specific parts of the functionality, and even within these agents. The desktop client, too, is separated into multiple independent components called "tools". ezDL is completely written in Java using common frameworks and libraries.

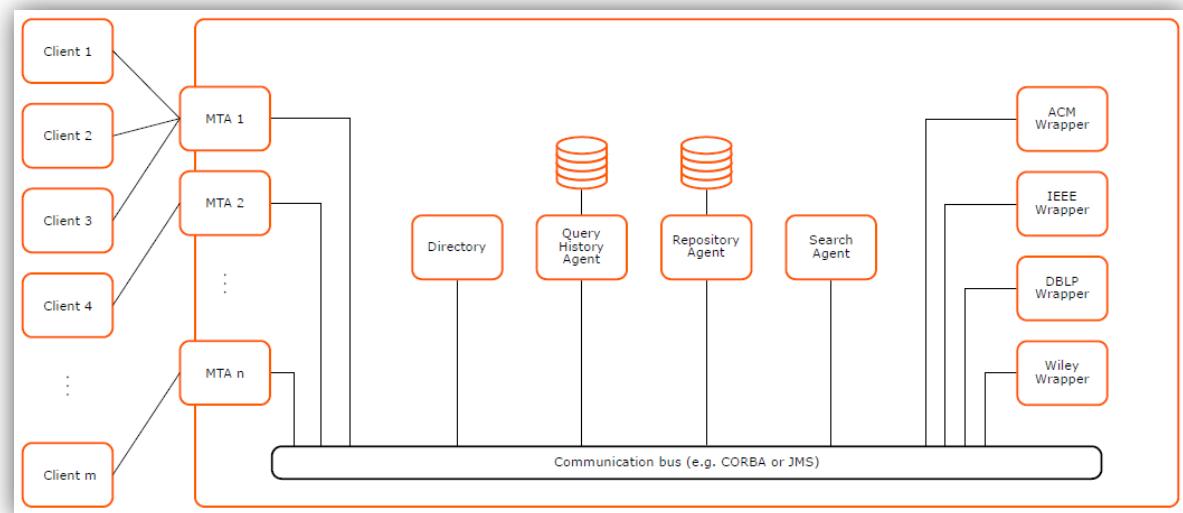


Image 1 - The overall architecture of iPerFedPat

## 4.11. The Backend

The backend provides a large part of the core functionality of ezDL: the meta-search facility, user authorization, a knowledge base about collected documents, as well as wrappers and services that connect to external services. Functionality that provides collaboration support and allows storing of documents and queries in a personal library is also located here. Lastly the backend is responsible for the communication with the database from which various functionalities are provided.

The following list briefly presents these functionalities.

- Storing of users
- Caching of results
- User logging
- Event logging
- Public library records
- User specific query history

#### 4.12. The Frontend

There are multiple frontends for ezDL, among them the basic desktop client and a web client. Specialized frontends exist for various applications. Clients for iOS and Android tablets are currently being developed. EzDL also involves tools and perspectives, where a tool comprises a set of logically connected functionalities and each tool has one or more tool views, interactive display components that can be placed somewhere on the desktop. While a configuration of available tools and the specific layout of their tool views on the desktop is called a perspective. Users can modify existing predefined perspectives as well as create custom perspectives. The desktop client already has many built-in tools and functionalities and can be easily extended.

## Chapter 5<sup>th</sup>: The PerFedPat project

### 5.1. Introduction

Patent search is an economically important problem, central to the R&D operations of many industries including pharmaceuticals, biotechnology, automotive and many more. Besides the economic interest, from a technological perspective, patent search reveals important challenges for the field of information access. This is because it has important differences (lengthy search sessions, demand for high recall, high value documents) despite the fact that it shares a number of important characteristics with web search. The PerFedPat project aims to research into a new generation of advanced patent search systems for the patent related industries and the whole spectrum of patent users by designing a new exciting framework for integrating multiple patent data sources, patent search tools and UIs. The iPerFedPat system, which will be the main result of the project, will have a pluggable architecture, providing core services and operations being able to integrate multiple patent data sources and patent related data streams, thus providing multiple patent search tools and UIs while hiding complexity from the end user. At the core of the system's architecture lies the idea of Personalized Federated Search. In iPerFedPat federated search is used as a method for retrieving information from distributed data sets into user's workbench, possibly operate and/or integrate, and finally deliver to the patent users for using them in a parallel, coordinated way. As a result the iPerFedPat system will be able to provide a rich, personalized information seeking experience for different types of patent search types, potentially exploiting techniques from diverse areas such as distributed information retrieval, machine learning and human-computer interaction.

### 5.2. Patent Data Sources

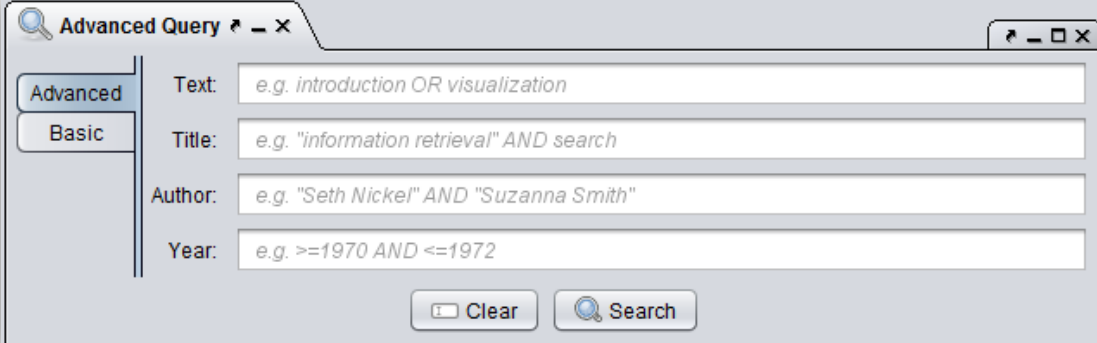
iPerFedPat currently provides access to three patent data source. These are Espacenet, the United States patent and trademark office, through Google patents and a collection called CLEF-IP. Espacenet offers the Open Patent Services - or OPS as it is also known – which is a web service providing access to the EPO's raw data via a standardized XML interface. The data is extracted from the EPO's bibliographic (EPODOC), legal status (Worldwide Legal Status Database/PRS), full-text (EPOQUE) and image (BNS) databases and is therefore from the same sources as the data in Espacenet and the European Patent Register providing free access to more than 70 million patent documents worldwide, containing information about inventions and technical developments from 1836 to today. With Google Patents, you can search the full text of the U.S. patent corpus and find patents that interest you. All patents



available through Google Patents come from the United States Patent and Trademark Office (USPTO). Patents issued in the United States are public domain documents, and images of the entire database of U.S. patents are readily available online via the USPTO website. To date, the USPTO has made available approximately 8 million patents and 3 million patent applications. Clef-IP is an extract of the MAREC dataset, containing over 3.1 million patent documents pertaining to 1.3 million patents from the European Patent Office with content in English, German and French, and extended by documents from the WIPO. iPerFedPat gains access to the content of these data sources using agents. These agents are responsible for three main actions. First and foremost is translating the query from internal application form, to the specific data source format. Second is actually querying against the dataset and lastly is parsing the response into a specified data structure which then is messaged along the available agents comprising the backend.

### 5.3. New Query Form Component

During development it was clear that new search fields must be available to the user in order to conduct a patent search. The available fields in ezDL are mostly specific to bibliographic search (see image 2) while the fields that were created for iPerFedPat are the ones commonly used in patent search (see image 3).



The image shows a web-based query form titled "Advanced Query". It features a sidebar on the left with two tabs: "Advanced" (which is selected) and "Basic". The main area contains four input fields, each with a label and an example value:

- Text:** e.g. *introduction OR visualization*
- Title:** e.g. *"information retrieval" AND search*
- Author:** e.g. *"Seth Nickel" AND "Suzanna Smith"*
- Year:** e.g. *>=1970 AND <=1972*

At the bottom of the form are two buttons: "Clear" and "Search".

Image 2 - ezDL bibliographic query form

Image 3 - iPerFedPat patent query form

The implemented fields can now provide a powerful query which is critical for users searching among millions of documents. All the fields are self-explanatory except for the various classifications in which are assigned o patents. Specifically a patent classification is a way the examiners of patent offices or other people arrange documents, such as patent applications and disclose the inventions according to the technical features involved. They patent classification arrangement is done so that they can quickly find a document disclosing the invention identical or similar to the invention for which a patent is claimed. The same document may be classified in several classes to satisfy all of its technical aspects.

#### 5.4. Data Representation

As said before patents consist of various data which are organized in a specific way. Actually patents have fields and these fields are not only used during search but they are also preserved while constructing an internal representation inside iPerFedPat. This is of most importance because it helps a great deal when handling a patent document object (etc. when exposing its content on the interface). For the most part the Java Object has the same fields as the query, with the exception of some extra ones which aren't used in the query and in term, they aren't used directly in the search. These fields are the kind code of the patent and the claims, description, citations and references. As said above these fields aren't used directly in the search by iPerFedPat. What happens is that iPerFedPat, being a federated search

application, isolates itself from the complexity of the search and uses the remote sources' RESTful Web Services provided, to apply a simple query which then is used by the data source internally and by using different algorithms a most representative set of results is constructed and returned.

## 5.5. Integration of Tools

As mentioned above these tools are displayed into an integrated browser using their web interface, which weren't specifically built for iPerFedPat. This fact is what gives the application a pluggable nature and will make it easily scalable with the integration of many tools in the future.

From the beginning it was clear that the user can't easily handle the results that a search can yield. The solution was already available by external tools with varying capabilities like IPC classification suggestions, entity extraction and clustering of the results. iPerFedPat can easily support the integration of many such external tools because many of these tools are web based and thus usually have HTML based interfaces and responses. Having that in mind, a patent tool Java class was created inside the project and parametrically it can be instantiated to host a different tool inside its integrated browser. Following are snapshots of the implemented tools after a patent search in any field that contains the text Titanium.

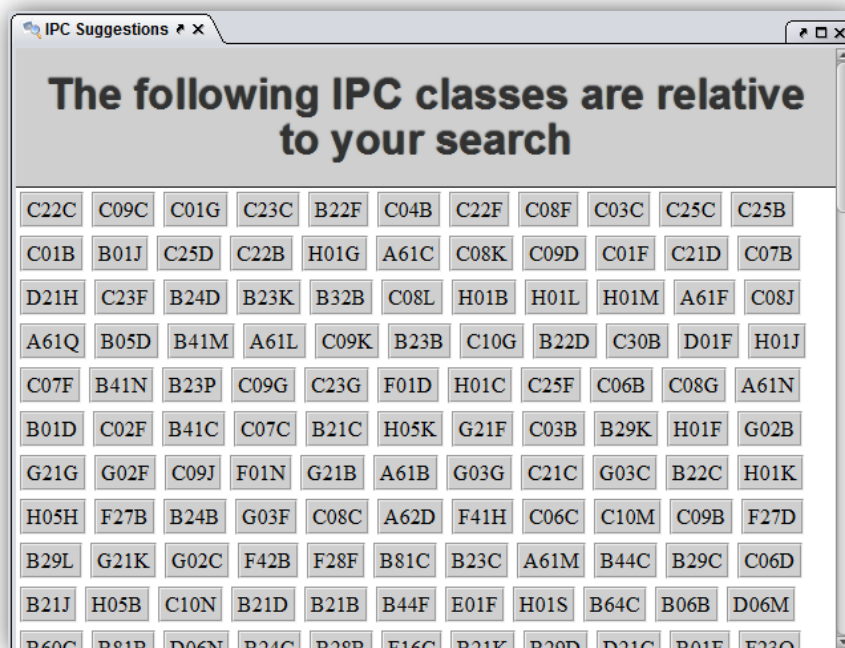


Image 4 - IPC Suggestions for Titanium

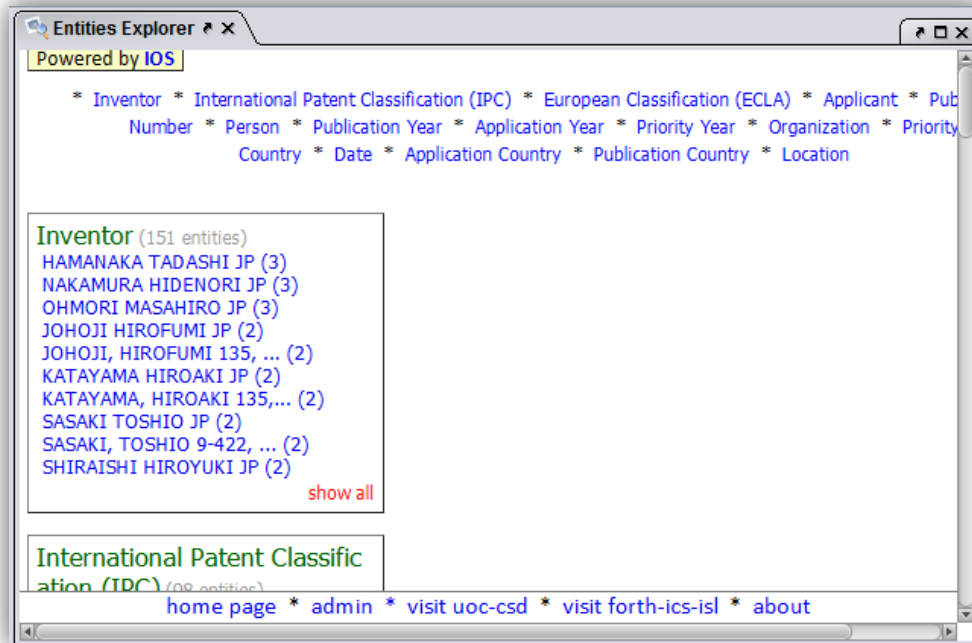


Image 6 - Entities extraction from current search

As mentioned above these tools are displayed into an integrated browser using their web interface, which weren't specifically build for iPerFedPat. This fact is what gives the application a pluggable nature and will make it easily scalable with the integration of many tools in the future.

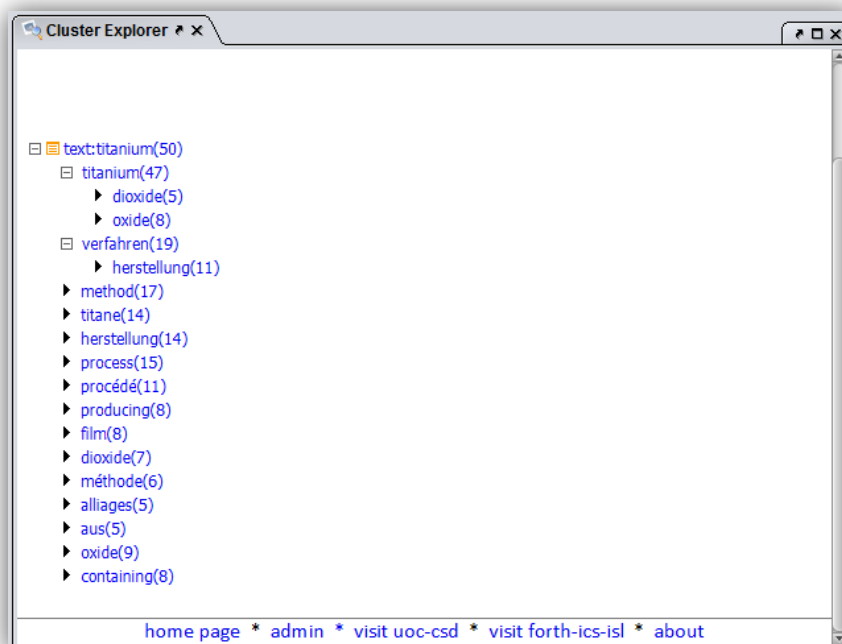


Image 5 - Result clustering for current search

## Chapter 6<sup>th</sup>: Technologies

### 6.1. Introduction

Various technologies were used during development and all of them were crucial in different levels of the application. Development supervision and management was provided with the use of a central project repository powered by Mercurial. Also we needed a private collection of patent documents for test purposes and generally for use as a first phase in testing newly implemented functionalities. For this exact reason Lucene and Solr were used, the first to create an index of the 3.1 million documents and provide powerful searching possibilities and the second to offer an interface from which queries can be sent towards the Lucene implementation. Of course nothing would be possible without the object oriented possibilities given by Java, which helped a great deal when designing the pluggable platform that was requested by the project description. Lastly, big parts of the applications' functionalities are provided using information retrieval techniques. The methods used the most include XML and HTML parsing of documents. These documents are either responses to a search query or a response given after a detail query. In both techniques it was made certain that the maximum amount of information was extracted.

### 6.2. Repositories

Repositories are source control management tools and their functionalities include the power to efficiently handle projects of any size while using intuitive interfaces. For iPerFedPat Mercurial was used because traditional version control systems such as Subversion are typical client-server architectures with a central server to store the revisions of a project. In contrast, Mercurial is truly distributed, giving each developer a local copy of the entire development history. This way it works independent of network access or a central server. Even though Mercurial is a fast and reliable platform, it offers the abilities to increase the functionality with extensions which are written in Python and can change the workings of the basic commands, add new commands and access all the core functions of Mercurial.

### 6.3. Lucene/Solr

As described above, continuously testing newly implemented functionalities of iPerFedPat became an issue because of the fair use policy that most of the remote patent sources adopt. After consideration, a private collection of 3.1 million documents was decided to be set up on a local server with search engine capabilities. For this task Apache Lucene was used which is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform. On the other side, Solr which is a popular and blazing fast open source enterprise search platform from the Apache Lucene project was used as an interface from which queries can be sent to the Lucene implementation. Its major features include powerful full-text search, hit highlighting, faceted search, dynamic clustering, database integration, rich document format handling, and geospatial search. Solr is highly scalable, providing distributed search and index replication, and it powers the search and navigation features of many of the world's largest internet sites. Solr is written in Java and runs as a standalone full-text search server within a servlet container such as Tomcat. Solr uses the Lucene Java search library at its core for full-text indexing and search, and has REST-like HTTP/XML and JSON APIs that make it easy to use from virtually any programming language. Solr's powerful external configuration allows it to be tailored to almost any type of application without Java coding, and it has an extensive plugin architecture when more advanced customization is required.

### 6.4. Java

iPerFedPat is a project with countless aspects and that is why Java was chosen to implement and develop the actual application. JAVA is an object oriented programming language and it is intended to serve as a way to manage software complexity. Java refers to a number of computer software products and specifications from Sun Microsystems that together provide a system for developing application software and deploying it in a cross-platform environment. Java is used in a variety of computing platforms from embedded devices and mobile phones on the low end, to enterprise servers and supercomputers on the high end. Java is nearly everywhere in mobile phones, Web servers and enterprise applications, and while less common on desktop computers; Java applets are often used to provide improved functionality while browsing the World Wide Web.

## 6.5. XML/HTML Parsing

Connecting to a new collection for searching (a digital library, a local IR system, a BibTeX file, etc.) is accomplished by implementing a wrapper agent. These are agents specialized in translating between iPerFedPat and a remote system. Remote systems can be those that provide a stable API like SOAP or SQL but also those that only have a web site and a search form. iPerFedPat has built-in support for most common fields and data types. There are abstract wrappers available to quickly connect to a Solr server and also if required, web pages can be scraped using an elaborate tool kit that is configured by an XML file. Because of this, even digital libraries without a proper API can be connected. Using these features iPerFedPat has the ability to parse and retrieve content of interest from within an HTML file by using its Document Object Model (DOM) to traverse through the nodes so that their content can be accessed. In the case where the document to be parsed is XML the already included in Java mechanisms are used. Specifically SAX parser is used for reading data from an XML document and is an alternative to that provided by the DOM. The difference between the two methods of parsing documents is where the DOM operates on the document as a whole, SAX parsers operate on each piece of the XML document sequentially.

# Chapter 7<sup>th</sup>: Conclusions

## 7.1. Overall Conclusions

In this thesis, iPerFedPat, was researched and developed, which is a framework system for interactive retrieval. Building upon state-of the art interface technology and usability results, iPerFedPat can provide an advanced user interface for patent search applications. The system can also be easily extended, at the functionality level as well as at the presentation level, meaning that with little effort many tools can be integrated and extend the usability of the application.

## 7.2. Applications Of iPerFedPat

iPerFedPat will be used side by side with other patent search tools by patent officers. This applications' goal is to make the search of prior art easier and faster than it currently is. Also, since iPerFedPat is free, anyone can install it and use it making it ideal even for amateur searches throughout the available patent sources. That way someone who may have a new idea can search and see if there are any previous similar applications or references.

## 7.3. Future Work

In the future iPerFedPat will have implemented a very big number of patent data sources and tools. This will make it valuable and reliable in the patent industry. Specifically, all free data sources available today can easily implemented, giving iPerFedPat access to a very big data set of patents. Also having as much results as possible per search isn't enough. That's were integrated tools will be able to assist by providing services to the user that make the task of finding specific patents easier. The next tools to be integrated will support different result visualizations and advanced patent term extraction.



## Publication References

Beckers, Thomas, Dungs, Sebastian, Fuhr, Norbert, Jordan, Matthias and Kriewel, Sascha (2012), ezDL: An Interactive Search and Evaluation System. In: SIGIR 2012 Workshop on Open Source Information Retrieval (OSIR 2012)

N., Fuhr, C.-P. Klas, A. Schaefer, and P. Mutschke. Daffodil: An integrated desktop for supporting high-level search activities in federated digital

Lederman, Sol (2008), Federated Search Primer. In: The Federated Search Blog, Deep Web Technologies