Development of ezDL Search System by Adding Capabilities for Patent Search

Alexander Technological Institute of Thessaloniki
Department of Information Technology

Bachelor Thesis of
Alexandros Bampoulidis

Supervisor Professor
Dr. Michail Salampasis

Thessaloniki, October 2014

ABSTRACT

The preparation of my thesis was a very interesting experience and a great opportunity for me to comprehend 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.

The aim of this thesis is 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.

The first chapter is an introduction to Information Retrieval (IR) and search systems.

The second and third chapter describe what was essential for me to understand in order to move on to developing the application. These chapters are about how the patent industry works. Specifically, they are focused on what a patent actually is, what the application and publication process include and what types of searches are involved.

The fourth and fifth chapters describe the technologies and platforms that were used to develop the application. In these chapters, Federated Search is described and an actual application of it (PerFedPat) is presented.

The sixth chapter describes my experience in developing PerFedPat.

The last chapter describes possible use cases of the application.

Table of Contents

Table of Figures

1. Introduction

Information is the knowledge that gives value to things and events around us. Corporations, government agencies and every individual collect information and make decisions based on it. It is safe to say that information and its retrieval can affect any system and every individual's life.

Information tends to be confused with data, which is not unexpected because there is a close relationship between them. Information is data with some significance and, therefore, information cannot exist without data. Today, data is produced in large amounts and archived information is stored in new spaces. Corporations, such as Google, Facebook and Amazon, collect and store large amounts of data and several other corporations and organizations create collections of data in certain domains (e.g. patent, medical, bibliographic, etc.), which are used to provide services.

The importance of search systems is and will be concerning the field of Computer Science. The exponential growth of the Web and the constant changing of the websites are making information retrieval more difficult. The amount and the validity of the retrieved information are some of the criteria that define a successful information retrieval of a search system.

However, there is a problem when a requested information is not visible by the well-known search engines (e.g. Google). For this kind of information, the development of professional search systems is required.

1.1 Information Retrieval

Information Retrieval (IR) is finding material of unstructured nature that satisfies an information need from within large collections.

Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

Figure 1 presents the basic information retrieval process which is the most widely used model for search systems. A user driven by an information need constructs a query in some query language. The query is submitted to a system that selects from a collection of documents (corpus) which are already indexed, those documents that match the query using certain rules of the retrieval engine. A query refinement process might be used to create new queries and/or to refine the results. More or less traditional search systems were based in this basic model and web search is also based on a modification of this model.

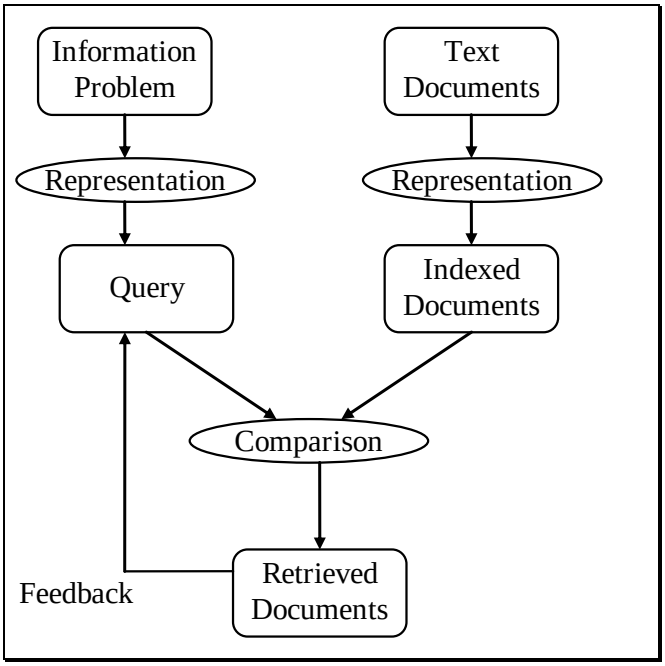

Fig. 1 Basic Information Retrieval Process

Instead of text documents, there can also be images, audio and video. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata.

Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query.

1.2 Web Search

A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as Search Engine Results Pages (SERPs). The information may be a mix of web pages, images, and other types of files. Some search engines also mine data available in databases or open directories. Search engines also maintain real-time information by running an algorithm on a web crawler.

A web crawler is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing. Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly.

Indexer is the software responsible for recording the terms of a web page and creating an index of them for future searches. The index is a result of an editing process of the terms of a web page, which usually are the title, headings and meta-data.

The tremendous power and speed of current web search engines to respond, almost instantaneously to millions of user queries on a daily basis is one of the greatest successes of the past decade.

Web search engines have proved extremely effective and efficient using the "query box" paradigm and ranked lists of search results to find relevant information for general purpose retrieval tasks. To a large extent this has led to the great success and exponential growth of the Web.

1.3 Professional Search

Professional search is the search that is performed in a workplace or for a professional reason or aim. Search technologies are used for professional search (e.g. bibliographic, patent, medical, engineering, scientific literature search) for more than 40 years as an important method for information access.

The current trend in professional search is towards Integrated Professional Search Systems. Although it is relatively easy to differentiate professional search from 'public search' with a number of characteristics, the concept of an *integrated* search system is not clear. Most definitions found in the IR literature converge to use the term "integrated" to define search systems that simultaneously access a number of different data sources providing a single point of search. This view is much more compatible with the Federated Search view that allows the simultaneous search of multiple resources (see Chapter 4).

Integrated search systems incorporate into the design space of next generation professional search systems the importance of the so-called Knowledge Extraction and Organization, e.g. classification schemes, taxonomies, ontologies. These are important prerequisites and resources for developing intelligent search tools and search systems that no longer just do what the professional searcher says but also what he means.

Such systems can manage and store session data as first-class objects and therefore increase the reproducibility of a search process and preserve complete state-full sessions that can be stored and managed at a later stage. This is a very important requirement for professional search systems.

The complexity of the tasks which need to be performed by professional searchers, which usually include not only retrieval but also information analysis and monitoring tasks, require association, pipelining and possibly integration of information as well as coordination of multiple and potentially concurrent search views produced from different datasets, search tools and user interfaces.

1.4 Differences between Professional Search and Web Search

Despite the tremendous success of web search technologies, there is a significant skepticism from professional searchers and a very conservative attitude towards adopting search methods, tools and technologies beyond the ones which dominate their domain.

There are a number of important parameters and characteristics that differentiate professional search from web search such as: lengthy search sessions (even days) which may be suspended and resumed, the notion of relevance can be different, many different sources will be searched separately, and focus is on specific domain knowledge in contrast to public search engines which are not focused on expert knowledge.

The current status of IR and search engine technologies is that they are able to reply to shorter queries (1-3 terms) at the document level and they can also respond to factoid queries ("what is the population of Thessaloniki?") at the sentence level. However, professional information needs are quite different and much more demanding many times. For example, in the patent domain, information needs would include general inquiries such as "how much is my patent worth if I sell it?" or "shall my company invest 10 million EUR in plastic packaging business?"

Despite the fact that many different IR and/or Natural Language Processing (NLP) technologies are used in the various sub-processes depicted in figure 1, and many exciting developments have been achieved that increased the efficiency and the effectiveness of this model, from an architectural point of view, it is important to observe that the relationships and dependencies between the different technologies, the core services which are used and the workflows and interactions which are executed in a search system during an information seeking process are not well defined.

Many search systems today combine a faceted search module based on static or dynamically extracted metadata. The faceted search tool and views can be combined with the "traditional" ranked result list. This simple and very common design of combining multiple search views is not captured in the basic IR model presented in figure 1. This is an important drawback for web search systems. The IR and NLP research communities have achieved tremendous progress in developing new algorithms and tools in various areas of information processing and retrieval, however there was little attention paid on how these results can come together to design next generation search systems. This view is supported by the fact that using and managing information workflows between autonomous (and possibly distributed) IR or NLP tools/services is the main design method used by different groups working in professional search systems.

2. The Patent Domain

An invention can either be a product, a process or an apparatus. To be patentable, it must be new, industrially applicable and involve an inventive step.

Patents protect technical inventions. They are valid in individual countries, for a specified period. Patents confer the right to prevent third parties from exploiting an invention for commercial purposes without authorization. In return for this period of protection, applicants must fully disclose their invention.

Patent applications and granted patents are published, which makes them a prime source of technical information.

2.1 Purpose of Patents

The wide-ranging economic significance of patents derives from the fact that patentees can prevent third parties from commercially exploiting their inventions for up to 20 years from the date of filing of the application. This enables them to recoup their development costs and gives them time to reap the rewards of their investment.

Effective patent protection encourages further investment in Research & Development (R&D), and is a key requirement for raising venture capital. It fosters technical innovation, which is crucial to competitiveness and overall economic growth.

The applicant's obligation to publish a full technical description of the invention contributes greatly to the dissemination of new technical knowledge. Over 80 % of the world's technical knowledge can now be found in patent documents. This inspires further inventions and at the same time prevents the duplication of R & D work.

2.2 Exploitation of Patents

The owner of a patent can exploit the invention himself, or permit someone else to do so. Individual inventors and small and medium-sized companies often lack the technical and financial means to bring their ideas to the market. Nevertheless, they too can derive great benefit from patents. For example, a patent can strengthen an inventor's negotiating position, as it gives the option of granting licenses or selling the protective rights altogether.

In granting a license, the patent holder allows the licensee to use the invention in return for some form of financial reward. This may be a one-off payment or a royalty on sales of a product incorporating the patented technology. A patent does not confer a right to make use of or exploit an invention, but to prevent others from deriving economic gain from the technology without the owner's permission. The use and exploitation of technology remain subject to national laws and regulations.

A patent does not provide a guarantee of commercial success. All it shows is that the idea in question is new, industrially applicable and inventive. It is up to the owner to develop the business side. The purpose of patents is not to establish long-term monopolies. They are granted for a limited period, which can only be extended in the case of medicines and pesticides which have to undergo lengthy clinical trials for safety reasons.

2.3 Preparation of an Application

To obtain patent rights for an inventor, the practitioner typically first drafts an application by interviewing the inventor to understand the nature of the invention and help clarify its novel features. Practitioners need to ascertain what is already known to people familiar with the general field of the invention —such already-known material is termed the prior art— and to obtain drawings and written notes regarding the features of the invention and the background.

During this initial phase, sometimes termed "patent preparation", the practitioner may also seek to determine precisely who contributed to the making of the invention. An incorrect listing of inventors may incurably invalidate any patent that might result from an application.

The practitioner may also seek to find out whether any publications, offers for sale, or other such public disclosures of the invention were made. Under the laws or regulations of some jurisdictions, public disclosures or offers to sell an invention prior to filing an application for a patent may prevent the issuance of the patent.

After drafting an application for a patent, complying with any further rules (such as having the inventor or inventors review the application prior to filing), and obtaining the applicant's permission, the practitioner files the patent application with the patent office. Usually, the practitioner seeks to file the application as soon as possible, because in a majority of jurisdictions including Europe, if two or more applications on the same subject matter are filed, only the party who filed first will be entitled to a patent under the "first-to-file rule".

2.4 Filing an Application

Most patent applications have at least two components, including a general, written description of the invention and at least one "embodiment" thereof, and a set of "claims," written in a special style that defines exactly what the applicant regards as the particular features of his or her invention. These claims are used to distinguish the invention from the existing prior art, and are compared by the patent office to the prior art before issuing a patent.

Patent applications in most jurisdictions also usually include (and may be required to include) a drawing or set of drawings, to facilitate the understanding of the invention. In some jurisdictions, patent models may also be submitted to

demonstrate the operation of the invention. In applications involving genetics, samples of genetic material or DNA sequences may be required.

Specifically a patent application consists of :

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

Applications can be filed in any language. However if an application is not filed in a recognized by the state language, a translation has to be submitted as well.

2.5 Filing and Formalities Examination

The first step in the 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 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.

2.6 Search

The search and examination phases constitute the main part of the prosecution of a patent application leading to a grant or a refusal.

A search is conducted by the patent office for any prior art that is relevant to the application in question and the results of that search are notified to the applicant in a search report.

Generally, the examiner conducting the search indicates in what aspect the documents cited are relevant (novelty, inventive step, background) and to what claims they are relevant. The materials searched vary depending on the patent office conducting the search, but principally cover all published patent applications and technical publications.

The patent office can provide a preliminary, non-binding, opinion on patentability, to indicate to the applicant its views on the patentability and let the applicant decide how to proceed at an early stage.

The search and examination process is principally conducted between the patent office and the applicant. However, in some jurisdictions, it is possible for interested third parties to file opinions on the patentability of an application. Such opinions may take the form of a formal pre-grant opposition procedure or it may simply be an opportunity of filing observations as a third party.

While the formalities examination is being carried out, a search report is drawn up, listing all the documents available to the patent 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 Patent Convention.

The search report is typically published with the patent application, 18 months after the earliest priority date, or if it is not available at that time it is published once it is available.

2.7 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 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.

2.8 Examination

The examination of patent applications may either be conducted at the same time as the search, or at a later date after the applicant has requested examination. Examination is the process by which a patent office determines whether a patent application meets the requirements for granting a patent.

The process involves considering whether the invention is novel and inventive, whether the invention is in an excluded area and whether the application complies with the various formalities of the relevant patent law. After the request for examination has been made, the patent office examines whether the patent

application and the invention meet the requirements of the Patent Convention and whether the 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.

2.8.1 Objections

If the examiner finds that the application does not comply with requirements, an examination report is issued drawing the examiner's objections to the attention of the applicant and requesting that they be addressed. The applicant may respond to the objections by arguing in support of the application, or making amendments to the application to bring it in conformity. Alternatively, if the examiner's objections are valid and cannot be overcome, the application may be abandoned.

The process of objection and response is repeated until the patent is in a form suitable for grant, the applicant abandons the application, or a hearing is arranged to resolve the matter.

In some jurisdictions, substantive examination of patent applications is not routinely carried out. Instead, the validity of invention registrations is dealt with during any infringement action.

2.8.2 Appeals

If the examiner and the applicant cannot reach agreement regarding the patentability of the application, the applicant may file an appeal to either the patent office or a court of law, asserting that his patent application was wrongly rejected.

For such an appeal to be successful, the applicant must prove that the patent office was incorrect in applying the law, interpreting the claims on the patent application, or interpreting and applying of the prior art vis-a-vis the patent application.

If the appeal is successful, the patent office or court may order that a patent be issued based on the application, or that the patent office correct its examination of the application if the patent office is found to have been incorrect. Otherwise, if the applicant is not found convincing, the rejection of the patent application may be upheld.

2.8.3 Abandonment

An applicant is free to abandon an application during the search and examination process. An application may be abandoned if, for example, prior art is revealed which will prevent the grant of a patent and the applicant decides to save cost by terminating the application. An application may be deemed abandoned by the patent office if the applicant fails to meet any of the requirements of the application process, for example replying to an examination report.

2.9 The Grant of a Patent

If the examining division decides that a patent can be granted, it issues a decision to that effect. The decision to grant takes effect on the date of publication.

A mention of the grant is published in the Patent Bulletin once the translations of the claims have been filed and the fees for grant and publication have been paid. The granted patent is a "bundle" of individual national patents.

2.10 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.

2.11 Opposition

After the 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 Patent Bulletin. Oppositions are dealt with by opposition divisions, which are normally made up of three examiners.

2.12 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 patent takes effect on the date on which it is published in the Patent Bulletin and applies from the beginning to all contracting states in respect of which the patent was granted.

2.13 Appeal

Decisions of the 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.

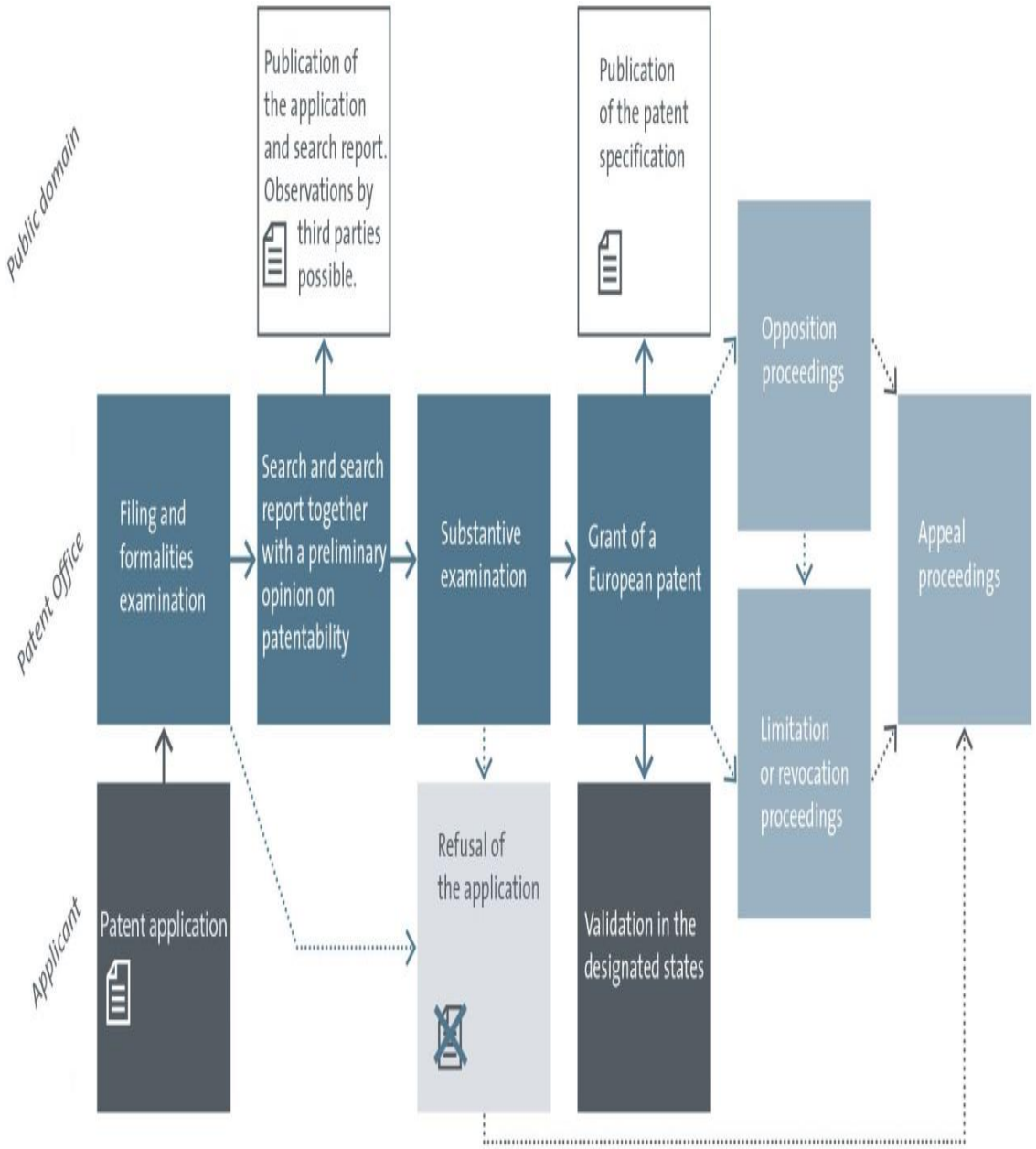

Fig. 2 The Grant Procedure at a Glance

3. The Patent Search Challenge

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 (see chapter 1.4).

In this chapter, the types of search that a professional patent searcher can perform are presented.

3.1 Prior-Art

Prior-art search task in patent retrieval is concerned with finding all prior art patents that are relevant to a patent application. Relevant prior art patents have common technical aspects with a patent application, and include patents that can invalidate the novelty of the invention and patents that describe the state-of-the-art in the field of the invention on which the patent application is building.

Identified relevant patents are cited in a search report which is part of the publication of the patent application. A typical patent application when filed to a patent office will include some initial patent citations describing the state-of-the-art. These citations are considered useful for patent examiners to understand the key aspects of an application and to start a search for relevant existing patents. However, large proportions of these initial citations are ultimately not found to be relevant, and are not included by patent examiners in the search report. Moreover, patent examiners usually identify a large amount of additional relevant patents.

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 novelty 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 R&D activities, it 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 the initiation of some work, a prior-art search needs to be performed in order to detect all existing similar developments or inventions.

3.2 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 date, if the patent requires priority, it will lose novelty.

In some countries, such as China, U.S.A. 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 European of countries, any invention that 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.

3.3 Freedom to Operate

Freedom to Operate (FTO) (also known as Right to Use, or Clearance search), includes a comprehensive infringement search of unexpired patents.

These searches also include a limited validity search of expired and unexpired patents, publications, and non-patent literature. These searches also help to locate expired patents and provide relevant proof of an invention that is already in public domain.

One of the primary tasks of FTO searches, therefore, is to determine if a particular act (method or process), such as testing or commercializing a product, has freedom to operate in any particular country and can be done without infringing valid intellectual property rights of others.

Freedom to Operate from a patent perspective means that it has been established – with a reasonable certainty – that a product does not infringe the intellectual property rights of others. Although, "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 a product would infringe or not.

Below are some reasons why the matter claimed in a patent could still be obtainable:

- Similar patents may still be available in other countries. Any claimed matter, in countries where no related patents have been issued, can be used.
- Laws about patentability vary from country to country, so even if a patent application was made, it need not have been approved.
- Some of the patents may have lapsed due to defaulting on due payments.
- Patents have limited shelf life, so it makes sense to verify the expiration dates.
- Generally, a particular patent claim can be rendered invalid due to the existence of some kind of prior art, like a publication or a presentation, about the matter claimed in the patent that the patent examination process missed. In fact, a patent can be challenged in some countries just because an inventor wasn't correctly named.

3.4 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.

One example of patent invalidity would be when 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 non-infringement, because the patentee is given a presumption of validity on the patent once it is issued.

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, validity search is also known as invalidity search. It is different from a patentability search which is conducted before granting 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.

3.5 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.

Finally, a patent landscape map is produced which analyses a collection of patents and groups patents relating to the same technology sub-areas into clusters. Those clusters which have a large number of patents are represented as peaks or mountains on the landscape map, whereas technology areas where there are few closely related patents are represented as deserts or islands in an ocean. Figure 3 below is a patent landscape map called a ThemeScape™ map generated using the Thomson Innovation™ software for the solar energy field of technology.

Fig. 3 ThemeScape™ Map for Solar Energy

Collections of patents for generating patent landscape maps may be obtained in different ways, e.g. by collating the patents of known competitors in a particular technology, by conducting subject matter searches in patent databases using various combinations of keywords and/or international patent classifications, and/or from citation trees based on key patents in a particular technology.

Each dot on a patent landscape map represents an individual patent, and patents of different owners can be shown in different colors to distinguish them. This helps to identify particular technology sub-areas in which different competitors are concentrating their R&D and patenting activity.

The patent landscape maps can also be time-sliced, e.g. to show how a technology area has developed over time and to show how some businesses have changed their patenting focus over time. Further advantages of analyzing patent landscape maps can include identification of trending technologies, opportunities in adjacent or related markets, discovery of new players in the field and potential partners or acquisition targets.

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.

3.6 Patent Search Systems

Patent search is an example of professional search where professional search experts typically use the Boolean search syntax and quite complex intellectual classification schemes. Of course there are good reasons for this.

A patent search professional often carries out search tasks for which high recall is important. Additionally s/he would like to be able to reason about how the results have been produced, the effect of any query re-formulation action in getting a new set of results, or how the results of a set of query submission actions can be easily and accurately reproduced on a different occasion (the latter is particularly important if the patent searcher is required to prove the sufficiency of the search in court at a later stage).

Classification schemes and metadata are heavily used because it is widely recognized that once the work of assigning patent documents into classification schemes is done, the search can be more efficient and language independent.

Users working in complex information workplaces (such as the patent domain) use multiple tools, interfaces, and engage in rich and complex interactions to achieve their goals. This view expresses a user-centered and highly interactive approach to information seeking. To address this view better the model of Integrated Search Systems is implemented in patent search systems.

The key objective of a patent search system is to integrate a set of tools and to enable effective support of the different tasks, stages and the cognitive states of the user during the patent search process.

The tools that a designer will decide to integrate into a patent search system, do not only have to do with existing IR technologies, but probably more with the context in which a patent search is conducted and the professional searcher's attitude. Furthermore, it is also very important to understand a search process and how a specific tool can attain a specific objective of this process and therefore increase its efficiency.

From an information seeking process perspective, the integration of different search tools in addition to the basic ranked list of patent documents returned from

the Distributed IR engine (see Chapter 4), allows different views of patent information to coexist.

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 shown in figures 4 and 5.

PerFedPat, a patent search system, is presented in chapter 5.

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

Fig. 4 Patent Search Systems

System Data	EAST	PatBase	PatBase Express	QPat	SureChem	WIPS Global
Owner Name	United States Patent and Trademark Office (USPTO)	Minesoft Ltd; RWS Group	Minesoft Ltd; RWS Group	Questel-Orbit	Macmillan Publishers Ltd.	WIPS Global
Full Text: Patent Authority Coverage	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
Current US Class	Yes	Yes	Yes	Yes	Yes	Yes
Original US Class	Yes	No	No	Yes	No	Yes
IPC - R	Yes	Yes	Yes	Yes	Yes	Yes
Original IPC data (v1-v7)	Yes	Yes	Yes	Yes	No	No
ECLA	Yes	Yes	Yes	Yes	No	Yes
Japanese File Index Terms	No	Yes	No	Yes	No	Yes
Japanese F-Terms	No	Yes	No	Yes	No	Yes
Other National Classification Systems	Yes	Yes	No	Yes	No	No

Fig. 5 Patent Search Systems

4. Federated Search

Federated search, also known as Distributed Information Retrieval (DIR), is a technique for searching multiple collections/information resources simultaneously. Each resource which is part of the federation must provide a function (accessible over a URL, a web service or any other remote procedure call method) for searching and retrieving results from its own index. The searcher can manually select the resources s/he wants to search, or all available resources can be part of a federated search.

However, when applying this technique usually queries are submitted to a subset of available remote resources which are most likely to return relevant answers. Particularly, when many resources are available automatic resource selection is necessary and it is based on creating pre-processed representations of the existing resources.

The results returned for each query by selected resources are integrated and merged into a single list. Using this process, federated search systems offer users the capability of simultaneously searching multiple online remote information sources through a single point of search.

From a user perspective, the defining feature of federated search is that the user interacts solely with the federated search system, without any requirement to know the intricacies of the underlying information sources, the query syntax and the methods which are internally used to index or retrieve documents. In effect, a federated search system functions as an intermediary between the user and multiple information resources.

Finally, the experience of using a DIR system is similar to that of using any other centralized IR system, as the DIR system in principle acts as a complete interface to the underlying information sources providing to its users a holistic, unified view of the available retrieval space comprising of the federated resources.

If the federated search process is decomposed it can be perceived as three separate but interleaved sub-processes (Figure 6):

- Source representation, in which surrogates of the available remote collections are created.
- Source selection, in which a subset of the available information resources is chosen to process the query.
- Results merging, in which the separate results are combined into a single merged result list which is returned to the user.

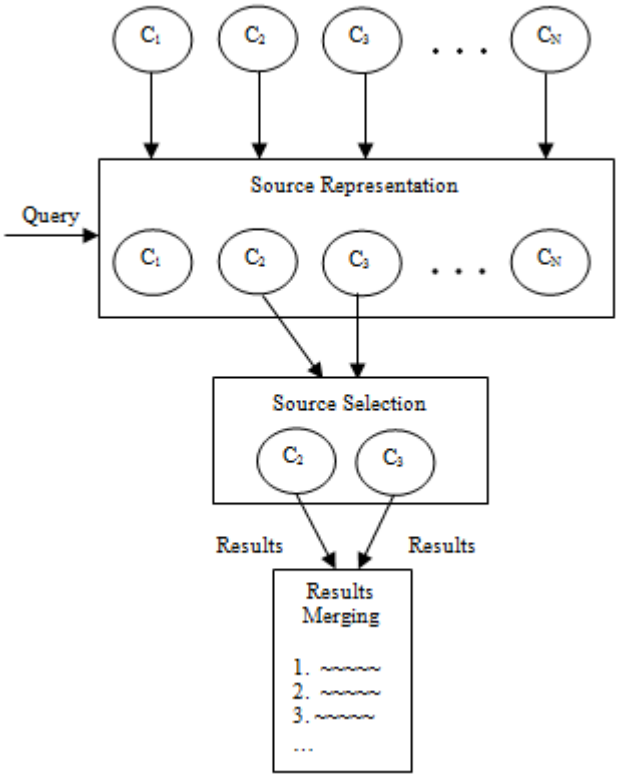

Fig. 6 Federated Search Process

4.1 Use and Advantages

DIR and federated search have been explored for about 20 years now. One recent application of DIR methods is the aggregated or vertical web search. Also many enterprise search applications rely on forms of DIR. Additionally, it is known that big web search companies use DIR techniques in maintaining distributed indexes mainly for scalability reasons. Federated search as a topic is also closely related to searching in peer-to-peer networks and metasearch engines. To understand the design and application space of DIR and federated search, one must understand that both can be selected as the basis for developing a search tool or solution either by inevitability or driven by an effort to engineer a more efficient or sometimes effective solution.

For example, DIR has been explored in the last decade mostly as a potential response to technical challenges such as the prohibitive size and exploding rate of growth of the web which make it impossible to be indexed completely. Big commercial search engines use programs called crawlers (or spiders) to locate and download documents when creating their indexes. Unfortunately, for a number of different reasons (e.g. pages are not linked therefore cannot be discovered, robot exclusion commands, download process is too slow, dynamic pages with content generated on the fly might be ignored) search engines cannot easily crawl documents located in what is collectively known as the hidden or invisible web.

Studies have indicated that the size of the invisible web may be 2–50 times the size of the web reachable by search engines.

Also there are many online authoritative resources (web sites), which are not reachable by search engines, offering their own search capabilities. Even publicly available, up-to-date and authoritative government information is often not indexable by search engines. A good example is PubMed which is a very large biomedical library which contains more than 25 million articles published since the 1950s. There are many similar resources which are not indexable by search engines, providing their own access to information such as yellow and white pages, patents, legal information, national statistics, news, catalogs to national libraries, scientific articles.

In the patent domain, for example, nearly all authoritative public online patent resources (e.g. EPO's Espacenet, WIPO's PatentScope) are not indexable and therefore not accessible by general purpose web search engines.

Using a federated search technique an increased coverage can be provided by searching a potentially large number of patent search engines which are wrapped in a federated patent search system. One key advantage, when compared with existing "crawler-based" centralized patent search systems, is that a federated search system does not need to maintain its own dataset and index. As a result, federated searches are inherently as current as the individual information sources, as these which are searched in real time. In other words, instead of expending the tremendous effort and resources which are required to download and index patents documents, something which may not be possible or very expensive in terms of time and costs, federated search techniques directly pass the query to the search interface of existing resource collections and effectively merges their results.

The previous paragraphs presented cases where is deemed necessary or inevitable to apply federated search because the effort to maintain a centralized patent search service is very large. A case where DIR methods, at least in patent search, can be a choice for improving efficiency and effectiveness is when it is applied in a way resembling more the cluster-based approaches to information retrieval. The general expectation is that if the correct sub-collections are selected then it will be easier for relevant documents to be retrieved from the smaller set of available documents and more effective searches can be performed.

4.2 Technical Challenges

4.2.1 Source Representation and Collection Size Estimation

The Source Representation phase takes place before the user submits a query to the federated search system. During this phase, surrogates of the available remote collections are created. The aim of this stage is to provide the DIR system with the best possible approximation about the contents of the federated information resources. Information which is required to create an accurate representation of the

resources typically is their thematic topicality (i.e. news, engineering, medical, sports, etc) and the number of documents that are contained in a collection (the size of the collection). Other information which is utilized in the subsequent resource selection phase are the terms that appear in it (i.e. the vocabulary of the resource), the number of documents that contain each term and potentially the number of times each term appears in each document.

After source representation, the federated search system possesses a representation set for each resource. The representation can be generated manually by providing a short description of the documents found and indexed in each resource. However, manually created representations cannot capture many terms that occur in a large collection. Therefore in practice, collection representation sets are usually generated automatically, and their comprehensiveness depends on the level of cooperation in the federated search environment. Uncooperative environments are these where federated collections do not provide any information about their contents and collection statistics to the federated search system. On the contrary, in cooperative environments the lexicon of the collections is provided to the central broker, therefore complete and accurate information can be used for the phase of collection selection.

However in a typical federated search system the remote collections are uncooperative, external to the "owner" of the federated search system, therefore the collections need to be sampled to establish a representation. This technique is known as query-based sampling or query probing.

Also, very typically source representation is done in advance before the user submits the query. However, when the remote resource is extremely dynamic there are source representation methodologies which can create representations "on-the-fly", during query time.

Besides an estimation of the terms that appear in the remote search engines, the actual number of documents that are available and indexed in each resource is also important. This is reasonable if we consider that source selection algorithms must take into consideration the size of the remote collections in order to determine the number of relevant documents that should be merged from each resource that will be selected in the resource selection phase. A first methodology was based on a simple capture-recapture approach. A second, more economical and yet sufficiently accurate methodology is called sample-resample. Using this method, queries are sent to the remote resource to estimate the document frequency of a term in a collection and with some simple calculations calculate the size of the remote collection taking into account the document frequency of a term in the representation of a collection (sample) and the size of the sample.

4.2.2 Automatic Resource Selection

There are a number of source selection approaches including CORI, gGlOSS, and others, that consider documents collections as document surrogates, consisting of

the concatenation of the collection's documents (the so-called big-document approach). These methods characterize different collections using collection wide statistics like term frequencies. These statistics, which are used to select or rank the available collections' relevance to a query, are usually assumed to be available from cooperative search providers. Alternatively, statistics can be approximated by sampling uncooperative providers with a set of queries as briefly discussed in the previous paragraph and extensively reported in.

The collection retrieval inference network (CORI) algorithm is probably the most widely used source selection algorithms from those following the big-document approach. The algorithm creates a hyper-document for each sub-collection, containing all the documents that are members of the sub-collection. When a query Q is submitted, the sub-collections are ranked based on the belief $p(Q|C_i)$ that the collection C_i can satisfy the information need of the query Q. The belief $p(r_k|C_i)$ that a term r_k -part of f the query Q-, is observed given collection C_i is estimated based on calculations using the number of documents in collection C_i that contain term r_k, the number of collections that contain term r_k, the number of terms in C_i, the average number of documents between all remote resources, the number of available collections. The overall belief $p(Q|C_i)$ in collection C_i for query Q is estimated as the average of the individual beliefs of the query terms $p(r_k|C_i)$.

The Decision-Theoretic framework (DTF) presented by Fuhr is one of the first attempts to approach the problem of source selection from a theoretical point of view. The Decision-Theoretic framework produces a ranking of collections with the goal of minimizing the occurring costs, under the assumption that retrieving irrelevant documents is more expensive than retrieving relevant ones. It is likely that DTF can provide a solid basis for source selection when developing industry-level federated search systems.

In more recent years, there has been a shift of focus in research on source selection, from estimating the relevancy of each remote collection to explicitly estimating the number of relevant documents in each resource. ReDDE focuses at exactly that purpose. It is based on utilizing a centralized sample index, comprised of all the documents that are sampled in the query-sampling phase and ranks the collections based on the number of documents that appear in the top ranks when querying the centralized sample index. Its performance has been shown to be similar to CORI at testbeds with collections of similar size and better when the sizes vary significantly. Two similar approaches named CRCS(l) and CRCS(e) were presented by Shokouhi, assigning different weights to the returned documents depending on their rank, in a linear or exponential fashion. Other methods see source selection as a voting method where the available collections are candidates and the documents that are retrieved from the set of sampled documents -retrieved from the centralized sample index- are voters. Different voting mechanism can be used (e.g. BordaFuse, ReciRank, Compsum) mainly inspired by data fusion techniques.

There is a major difference between CORI and the source selection algorithms that utilize the centralized index. CORI builds a hyper-document for each sub-collection while the other collection selection methods are based on the retrieval of individual documents from the centralized sample index. Due to its main characteristic CORI has been repeatedly reported in the literature not performing consistently well in environments containing a mix of “small” and “very large” document collections. However, in the patent domain where similar inventions contain to a large extent very different terminology in some settings the idea of building hyper-documents centred around a specific technical concept such as IPCs (International Patent Classifications) may be very well suited. The homogenous collections containing patent documents of the same IPC as the hyper-documents in CORI should normally encompass a strong discriminating power, something very useful for effective and robust resource selection.

4.2.3 Results Merging

Merging the result lists from remote resources is a complex problem not only because of the variety of retrieval engines that may be used by the individual collections, but also because of the diversity of collection statistics.

In environments where the remote collections return not only ranked lists of documents but also relevancy scores Raw Score Merging merges the results as they are returned from the remote collections in a descending order. However, this approach does not produce good results because of the problem of different statistics which eventually makes the scores from different remote resources incomparable. For example, in a collection that is mainly about sports a document containing the term “computer” will rank very high if that term appears in the query, while the same document would rank lower in a computer science related collection. The Weighted Scores Merging algorithm overcomes the above issue by assigning each document a score which is based both on the relevancy of the document itself and the relevancy of the collection where it belongs. This way, high scoring documents from low scoring collections (as in the above example) rank lower than highly relevant scores from highly relevant collections.

The CORI results merging algorithm is a weighted scores merging algorithm and has proved effective. The final score of each document coming from different remote resources is calculated using two simple equations that are used to normalize the collection and document scores to a range of 0 to 1.

Another set of results merging algorithms (e.g. SSL, MRRM), make use of a centralized index, comprised of all the sampled documents from the remote collections. The algorithm takes advantage of the common returned documents and their corresponding relevancy scores between the centralized index and the remote collections to estimate a linear regression model between the two scores. In case when a collection does not return scores and only ranked lists, factitious scores are calculated and assigned to the documents in a linear fashion.

5. PerFedPat

5.1 Overview

There is an abundance of systems today to search for patents. Some of them are free and have become available from patents offices and Intellectual Property (IP) organizations the last ten years (e.g. Espacenet and Patentscope), as the growth of the internet and the development of search technologies facilitated the provision of powerful web-based search systems of patent databases. Other systems are free – but developed by search technology providers (Google Patents)-, or are based on subscription and are provided from other independent producers (e.g. Delphion).

All web-based patent search systems allow searches using the simple "search box" paradigm. Other free or commercial systems may have better capabilities, for example for structural searching in particular fields, term proximity operations or to leverage domain semantics, but essentially they all operate on the same centralized index paradigm. According to this paradigm, patent documents need to be periodically crawled or otherwise collected, afterwards they are analyzed and eventually become part of the centralized index.

PerFedPat is an interactive patent search system that follows a different approach based on Federated Search. Federated Search represents a DIR scenario and allows the simultaneous search of multiple searchable, remote and physically distributed resources. PerFedPat provides core services and operations for being able to search, using a federated method, multiple online patent resources (currently Espacenet, Google patents, Patentscope and the CLEF collection), thus providing unified single-point access to multiple patent sources while hiding complexity from the end user who uses a common query tool for querying all patent datasets at the same time.

PerFedPat is developed upon ezDL therefore, in addition to the patent resources which are provided in PerFedPat, there are other resources already provided by ezDL, most of them offering access to online bibliographic search services (e.g. ACM DL, DBLP, Springer, PubMed) for non-patent literature.

The searcher can manually select the patent resources s/he wants to search, or all resources can be part of the federated search. The federated search system then aggregates the results that are received from the search engines for presentation to the user. Using this federated search process PerFedPat can provide increased coverage using a large set of patent search engines which are wrapped in the PerFedPat federation. One key advantage, when compared with existing "crawler-based" search systems, is that PerFedPat does not need to maintain its own dataset. In effect, federated searches are inherently as current as the individual information sources, as these are searched in real time.

Wrappers are used which convert the PerFedPat internal query model into the queries that each remote system can process. "Translated" queries are routed to

remote search systems and their returned results are internally re-ranked and merged as a single list presented to the patent searcher.

Also, PerFedPat not only searches multiple datasets in parallel, but it also offers more sophisticated services such as removing duplicates, merging and re-ranking the results. There are also additional features like filtering or grouping and sorting the results according to existing features or patent metadata (e.g. per patent resource, per year, IPC, inventors etc). Using the grouping function a searcher can very quickly get an overview of the full set of results returned from the different federated patent systems. The basic objective is to improve the accuracy and relevance of individual searches as well as reduce the amount of time required to search the multiple resources which are available. For some tasks, for example prior-art patent search, these are key objectives.

PerFedPat uses a pluggable and extensible architecture that provides multiple patent search tools and User Interfaces (UIs). Consequently, in PerFedPat federated search is used beyond the way that it is used in traditional Distributed IR, i.e. to provide a single merged list of multiple ranked results. Hence, an innovative feature of PerFedPat is that it enables the use of multiple search tools which are integrated in PerFedPat to assist professional searchers.

In that way different search tools can become part of the PerFedPat search system exploiting several existing IR and Natural Language Processing (NLP) technologies. The way these tools could be actually used depends on the context, e.g. the task and the experience and the persona of the user conducting the search.

Currently the search tools which are integrated are: a) an International Patent Classification (IPC) selection tool, b) a tool for faceted navigation of the results retrieved based on existing metadata in patents, c) a tool producing clustered views of patent search results d) a Machine Translation (MT) tool for translating queries for cross lingual information retrieval.

Furthermore, it is also very important to understand the search process and how a tool can attain specific objectives and generally increase the efficiency of the process which it is supposed to support. PerFedPat can deliver parallel views from the patent resources which can be opened in different tools on the user's workbench. Using this idea the PerFedPat system implements the strong UI metaphor of the workbench based on the following general schema:

```
For each action in user's Workbench in PerFedPat(e.g. submission of a
query)
Repeat
     Retrieve data from N data source(s)
     Transform data appropriately (e.g. translate), select, filter
     Merge data if required
     Present final results, group, visualize etc.
     Notify other search tools and adapt if possible and necessary
Until goal is achieved or search is terminated or saved (user
decision)
```

Based on this architecture PerFedPat is a pluggable system which puts together the following components: retrieval, selection, integration, presentation and adaptation (Figure 7).

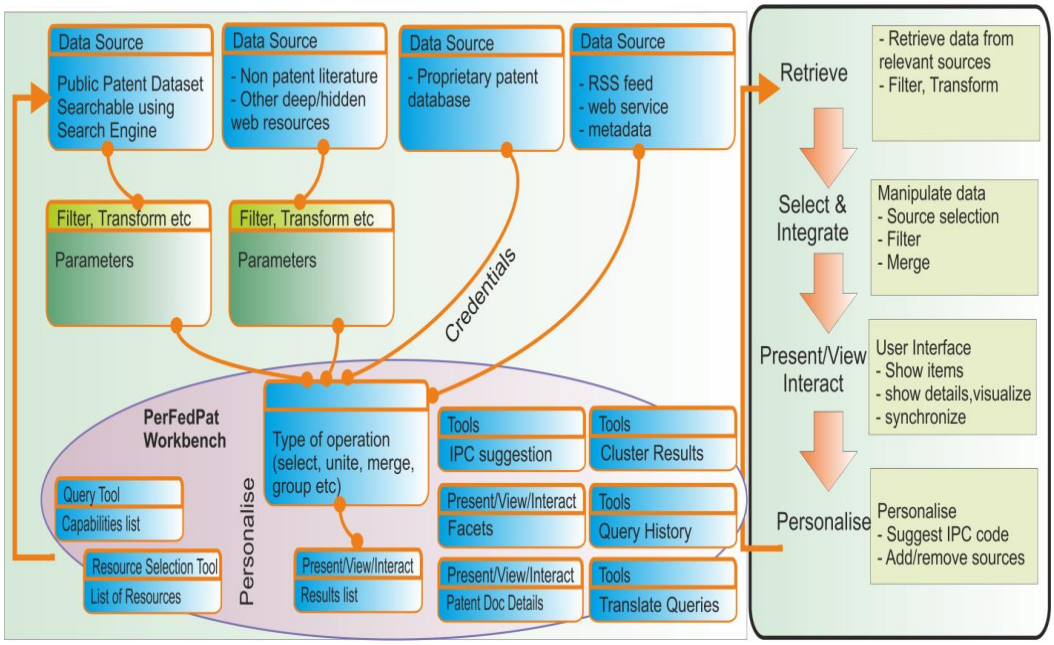

Fig. 7 PerFedPat Architecture and Component Overview

The basic principles explained above define PerFedPat as an Integrated Professional Federated Search System. In PerFedPat the meaning of the term integrated is expanded to define search system designs where multiple search tools can be used (in parallel or in a pipeline) by the professional searcher. As a result the definition of integrated professional search systems, in the case of PerFedPat, primarily describes a rich information seeking environment for different types of searches, utilizing multiple search tools and exploiting a diverse set of integrated IR and Natural Language Processing (NLP) technologies.

Although PerFedPat relies on existing patent search systems to execute the core retrieval task, from an architectural point of view PerFedPat is innovative using the Federated Search approach and goes beyond the state of the art in patent search systems in terms of scale, heterogeneity as well as extensibility as it is based on a service-oriented, message-centric architecture able to integrate data sources into new, more useful ways.

From that perspective, the PerFedPat system is the first open architecture data aggregator for patent information, and its contribution is to show that the sum of the utilities provided by each search tool could be really bigger than the single utilities and enabling possibilities lie in an integrated approach for patent data delivery and intelligent processing and presentation.

- Scale*:* The PerFedPat patent search system is in principle more scalable than other systems since it is based on a highly distributed method for accessing

patent information sources, each using its own storage, processing, and searching capabilities.

- Heterogeneity*:* Different data sources, search tools and UIs can be combined in a non-predetermined way, non over-engineered method, as far as they abide by the PerFedPat framework.
- Extensibility*:* The PerFedPat framework is not developed be a single turnkey one-size-fits-all solution, but instead is designed as a pluggable architecture in which it is easy to develop and deploy new components. The ezDL framework on which PerFedPat is based is easy to extend and is based on a service-oriented architecture.

5.2 Architecture

PerFedPat follows the client-server component-based architecture.

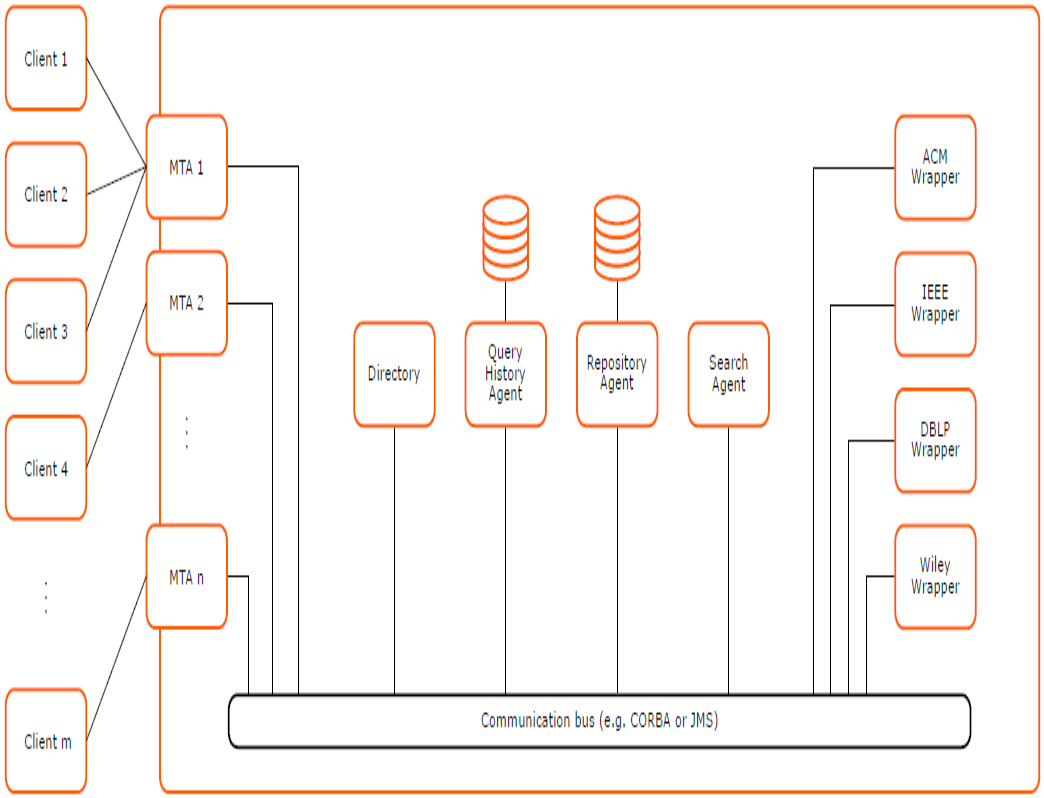

Fig. 8 High-level Overview of ezDL

5.2.1 The Backend

The server (Backend in ezDL terms) provides a large part of the core functionality such as the meta-search facility, user authorization, a knowledge base (repository) about previously retrieved documents, as well as wrappers that connect to external services. The system architecture makes extensive separation of components to keep interdependencies to a minimum and make the system more stable.

Within the backend individual processes operating as “software agents” handle specific parts of the functionality. Software agents are autonomous software

components that communicate with their peers, by exchanging messages in an agent communication language.

The Directory is a special agent that keeps a list of agents and the services they provide. Upon start, each agent registers with the Directory and announces the services it provides.

In PerFedPat wrapper-agents are implemented for the Espacenet, Google Patents, Patentscope and CLEF patent resources. There are two different types of wrappers.

When interfaces (APIs) exist (such as for example in the case of CLEF in PerFedPat), in which full control and access are possible, then it is easier to write a wrapper which sends a query or other requests and receives back information from the fully controlled search system, usually in XML or other structural format (e.g. JSON).

In case of web-based systems completely external to PerFedPat (for example Google Patents, Patentscope), an analysis of the search results web page is required and is programmed in the wrapper and conducted in the backend. Usually this is facilitated by web page analysis tools using the XPath Language, a query language for selecting nodes from an HTML/XML document.

Also, multiple sections ("pages") from search results can be obtained; by default PerFedPat retrieves 200 results from CLEF, 100 from Espacenet and 50 from Google Patents and Patentscope.

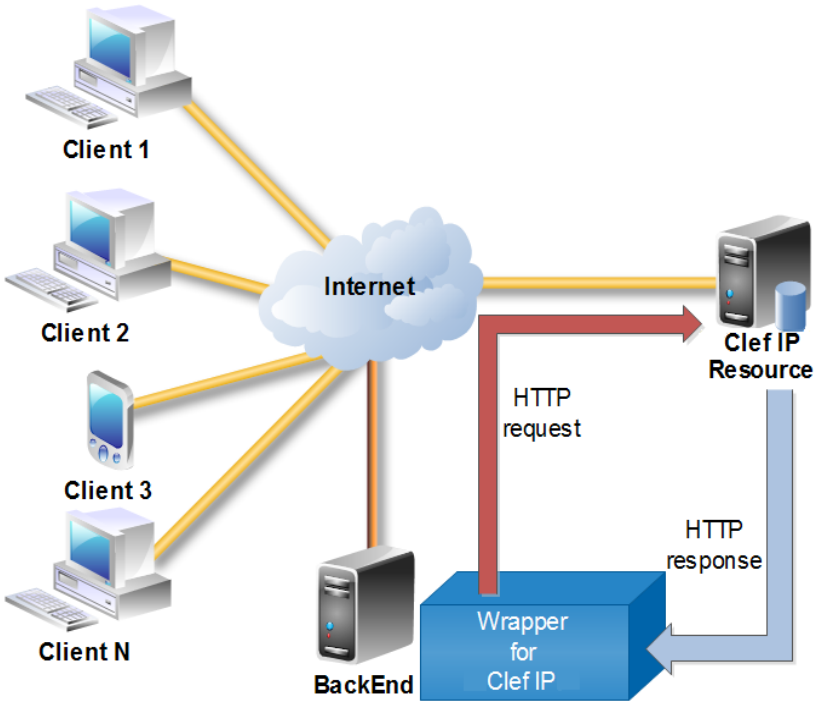

Fig. 9 A Wrapper's Functionality as Part of the Architecture

MTAs (Message Transfer Agent) are the links between the client and the backend. Each MTA is responsible for user authorization and transferring messages between the backend and the client. For example, the client's submitted query is transferred to a MTA and then, the MTA transfers it to the Search Agent using a special type of message.

5.2.2 The Frontend

The desktop client (Frontend in ezDL terms), like the backend, is separated into multiple independent components/agents called “tools” (Figure 10). A tool comprises a set of logically connected functionalities. Each tool has one or more tool views, interactive display components that can be placed somewhere on the desktop/workbench.

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 their own custom perspectives and load them later when needed.

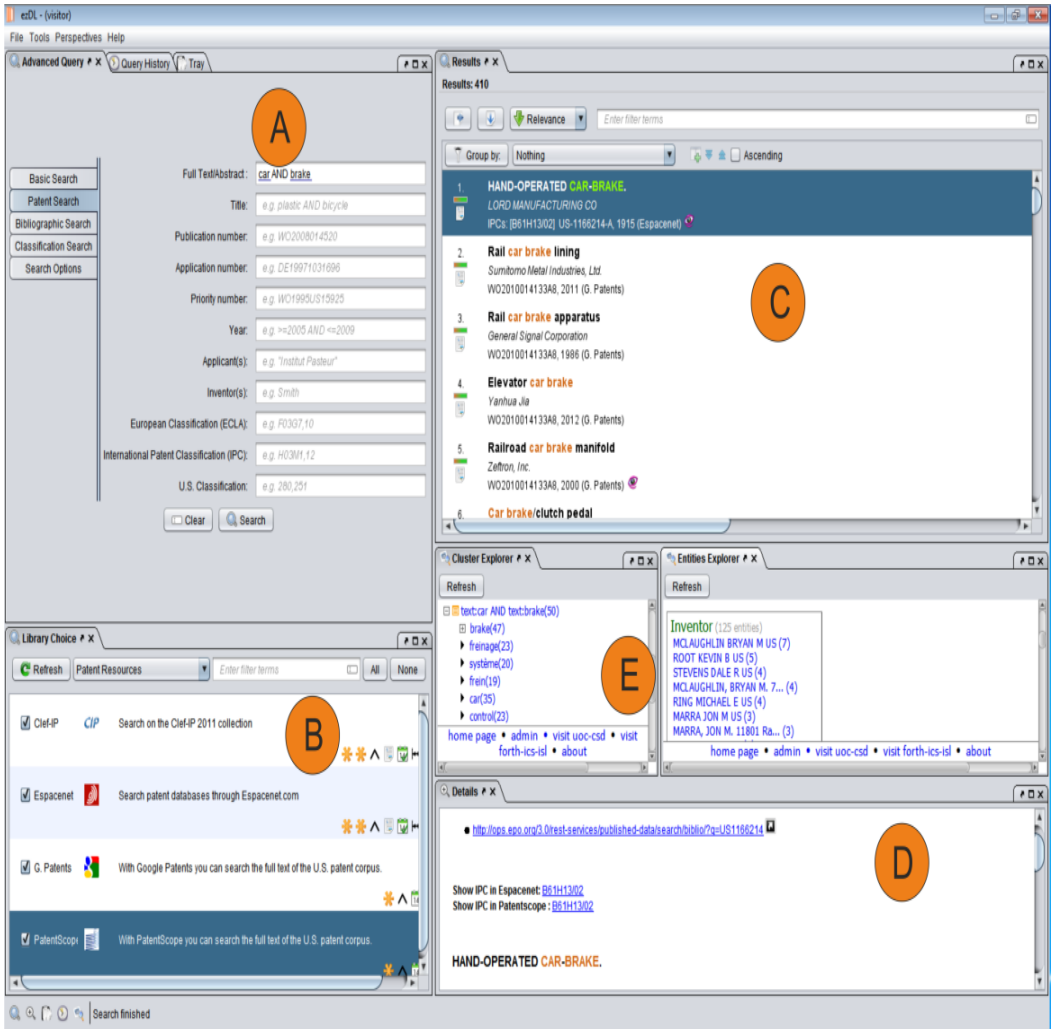

Fig. 10 PerFedPat Workbench Overview with Some Core and Patent Search Tools Open

5.3 Core Tools

The PerFedPat desktop client already has many built-in tools and functionalities inherited directly from ezDL. The functionality of some of the existing tools was extended and new tools were added in PerFedPat, specifically designed and implemented to address the needs of patent search.

The query tool offers a variety of query forms for different purposes. In PerFedPat this tool was extended to address the need for more advanced fielded search which is necessary in patent search. Each patent resource from the four available supports a different set of fields in the fielded search it implements locally. Fields that are supported from all or from most of the patent federated resources are implemented in the advanced search view in PerFedPat (colored circle area A in Figure 10).

The queries that users enter are expressed in a grammar specific to PerFedPat that is flexible and allows simple free text queries like “term1 term2” as well as more complex ones like “term1 AND (term2 NEAR/2 term3)”. Wildcards and phrases are also supported by the internal query tool. Fields can be combined so more complex queries can be constructed. A query is internally represented in a tree structure. Obviously in each wrapper the query which is received in the internal tree structure is transformed to the form that each patent resource is able to process. Note that each patent resource is marked with a number of symbols (area B in Figure 10) which show to the user which capabilities of the internal query structure are supported in the remote patent resource. When full support is not available queries are partially translated in a way to include the capabilities which are supported.

Other “standard” tools include:

- the Library Choice for selecting information sources,
- the Query History which lists past queries for re-use and allows grouping by date and filtering,
- the Tray tool can be used to temporarily collect relevant documents within a search session,
- The Results tool which shows the merged and re-ranked results returned from the patent resources. Results can be grouped, sorted, filtered and exported (C).
- The Details View tool (D) shows additional details on individual documents, such as thumbnails or short summaries where available, or additional metadata not included in the surrogate that is shown in the result list. A detail link can be provided to retrieve the full text of a patent document if available. Since patents can be very long documents, this tool was extended to provide quick reference links to parts of the patent (e.g. citations). Also some shortcuts were built to link the classification codes of a patent shown in the details or results view directly to online services presenting the classification hierarchy.

5.4 Patent Tools

The new tools that are integrated in PerFedPat are presented here. Their use is presented in detail in chapter 7.

5.4.1 IPC Suggestions Tool

The IPC suggestion tool aims, given a query, to select a number of IPC codes, at different levels of the classification hierarchy if requested, which include patents related to this query. The algorithm and the method which was used to implement this tool is based on DIR techniques for collection selection which was extended for patent search. The essence of the method is that it identifies relevant IPC codes not by searching the textual description of IPC classes, groups, subgroups etc., but by using an indirect method. First it retrieves patents which are already allocated to IPC codes, and then indirectly builds a probability estimation of the relevance of the allocated IPC codes to the query.

This tool was integrated to analyze the improvement it could provide for real users conducting prior-art patent searches. The improvement is related to the very fundamental step in professional patent search which is “defining a text query, potentially by Boolean operators and specific field filters”.

In prior art search probably the most important filter is based on the IPC classification. Selecting the most promising/relevant IPC codes depends of course on the prior knowledge of a patent professional in the technical area under examination, but sometimes the area of a patent application may not be easily distinguishable or usually a patent uses various technical concepts represented by multiple IPC codes. To identify all these relevant IPC codes could be a difficult, error prone and time-consuming task, especially for a not very knowledgeable patent professional in some technical area.

The IPC suggestion tool supports this step automatically; this is, given a query, it selects the most appropriate IPC codes and can copy the top 5 relevant IPC codes to the clipboard and then the user can paste them to the query tool. The query tool then initiates a filtered search based on the automatically selected IPC codes. This process naturally resembles the way patent professionals conduct various types of patent search.

Also, the patent searcher may use the tool not only to produce IPC-based filters automatically to narrow his/her search, but also as a classification search which will be used as a starting point to identify and closer examine technical concepts as these are expressed in IPC codes and to which a patent could be related and should be examined more vigorously. This ground understanding step helps soon after in formulating better queries with higher precision which will usually include expansion with noun-phrases from the IPC codes which deemed relevant. Of course the patent searcher has the flexibility to add the IPC codes that he assumes relevant in addition to the ones suggested by the IPC suggestion tool.

5.4.2 Entities and Cluster Explorer

The Entities Explorer tool supports an exploratory strategy for patent search that exploits the metadata already available in patents in addition to the results of clustering and entity mining that can be performed at query time. The results (metadata, clusters and entities grouped in categories) can complement the ranked lists of patents produced from the core patent search engine with information useful for the user (e.g. providing a concise overview of the search results) which are further exploited in a faceted and session-based interaction scheme that allows the users to focus their searches gradually and to change between search methods as their information need is better defined and their understanding of the technical topic evolves in response to found information.

The Cluster Explorer tool provides patent searchers with an overview of the results shown in the Results tool. It aims at grouping the results into topics (called clusters), with predictive names (labels), aiding the user to locate quickly one or more documents (patents in our case) that otherwise would be difficult to find, especially if they are low ranked. The Suffix Tree Clustering algorithm is used that derives hierarchically organized labels and is able to favour occurrences in a specific part of the result (e.g. in the title). The last feature is very useful for clustering the results of a patent search, because the invention title usually is the most descriptive part of a patent.

These two tools are basically meant for more exploratory types of patent search and can be also used as patent analysis tools. In several types of patent search and analysis, one must look beyond keywords to find and most importantly analyze patents based on a more sophisticated understanding of the patent's content and meaning. Technologies such as entity identification and analysis could become a significant tool for such searches and, together with other text analysis technologies, could become the cutting edge of information retrieval science.

5.4.3 Query Translator

Cross-Language Information Retrieval (CLIR) is a subfield of IR dealing with retrieving information written in a language different from the language of the user's query. For example, a user may give his/her query in English but wants to retrieve relevant documents written in Chinese. Multilingual IR (MLIR) addresses the problem of multilingual access to text databases and can be seen as an extension of the general IR problem corresponding to paraphrase. It aims for retrieval of documents in several languages from a query.

Machine Translation (MT) is an essential tool for CLIR and MLIR (if the translation quality is high) and the challenge of accessing patent document written in different languages from all around the world using MT methods has been addressed in several evaluation campaigns.

The Query Translator tool uses third party MT services (Microsoft Bing & Patentscope) in order to translate queries into different languages so that some

types of CLIR and MLIR can be conducted in PerFedPat. Depending on the languages which are selected from the information searcher to use from the MT tool, and the availability of patent documents in different languages in PerFePat's federated patent resources, the Query Translator tool in PerFedPat can assist the information searcher to retrieve documents in several languages from a query posed in one language.

To initiate an on-off CLIR process the user needs to press "translate and query" or alternatively s/he can activate the Query Translator tool from the Search Options Panel and keep it active for a complete search session. If the MT tool is activated, for every query which is submitted the query tool sends a message to the MT tool which then it sends the appropriate requests to the selected MT service. There are two different MT services currently integrated into PerFedPat. Standard HTTP requests are used to communicate and receive information from the translation services. Only one MT service can be selected for each query submission process. This option (i.e. using one instead both MT services) probably attains better query homogeneity and accuracy something which could be useful for getting a coherent set of results in multiple languages. The user selects the MT service and the source/destination languages. The query is translated into different languages and the translations from each selected language are combined (using the OR operator) and passed as a single query to the Query Tool.

The translated query is subsequently sent from the Query Tool to the selected patent resources. Note that the language of the documents which are returned from the patent resources cannot be always fully controlled. For example in the CLEF dataset, the user can fully select which lingual subset of the patents to search (for example search only patent documents written in French or German), while the same is not possible in Espacenet for example.

5.4.4 URL Logger

The URL Logger tool shows the final query which is transmitted for execution to the remote patent resource. This tool is used to validate the search process more easily. In this way the federated search process becomes more transparent for the end users and the designers of the application.

Also, this tool has two more functionalities : Pause/Resume Session and Level of Logging. Both of them are used for logging reasons (for more details, see Chapter 6.8).

5.5 Technologies

In this section, the various technologies that were used during development, all of which were crucial in different levels of the application, are presented.

5.5.1 Java

PerFedPat 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.

5.5.2 Mercurial SCM

Development supervision and management was provided with the use of a central project repository powered by Mercurial.

Mercurial is a source control management (SCM) tool and its functionalities include the power to efficiently handle projects of any size while using intuitive interfaces.

In PerFedPat 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, change the workings of the basic commands, add new commands and access all the core functions of Mercurial.

5.5.3 MySQL

MySQL is a relational database management system (RDBMS), and ships with no GUI tools to administer MySQL databases or manage data contained within the databases. Users may use the included command line tools, or use MySQL "front-ends", desktop software and web applications that create and manage MySQL databases, build database structures, back up data, inspect status, and work with data records. The official set of MySQL front-end tools, MySQL Workbench is actively developed by Oracle, and is freely available for use.

In PerFedPat, MySQL offers the following functionalities:

- Storing of users.
- Caching of results.
- User logging.
- Event logging.
- Users' personal library.
- Users' query history.

5.5.4 Maven

Maven is a build automation tool used primarily for Java projects. Maven addresses two aspects of building software: First, it describes how software is built, and second, it describes its dependencies.

It uses conventions for the build procedure, and only exceptions need to be written down. An XML file describes the software project being built, its dependencies on other external modules and components, the build order, directories, and required plug-ins. It comes with pre-defined targets for performing certain well-defined tasks such as compilation of code and its packaging.

Maven dynamically downloads Java libraries and Maven plug-ins from one or more repositories such as the Maven 2 Central Repository, and stores them in a local cache. This local cache of downloaded artifacts can also be updated with artifacts created by local projects. Public repositories can also be updated.

The use of Maven in PerFedPat is more than a coincidence because this technology offers a handling of the projects aspects in such a way, that development is more productive and fuss free.

Specifically, all the dependencies were automatically added and used during the development of the various modules, while building specific modules or the whole project became as easy as running one command because of the Maven-IDE integration and the various build phases and goals that are supported.

6. My Work

This chapter is a summary of my work on PerFedPat as a developer during my internship in Vienna University of Technology. It includes bug fixes, extensions and updates, all of which are briefly analyzed below.

6.1 Details View

- Rearrangement of the Details View: As seen in the figures below, an aesthetic change was made to the Details View to make it more user friendly.
  The Details View is a HTML page, so this rearrangement required HTML/CSS editing.
- Claims and description are not displayed anymore. That is because they were deemed unnecessary, because a link to the original patent document is provided to the user.
- Removed or updated the detail links: Some of the links to the original patent document were removed or updated, either because a shorter version of them exists or became unified.

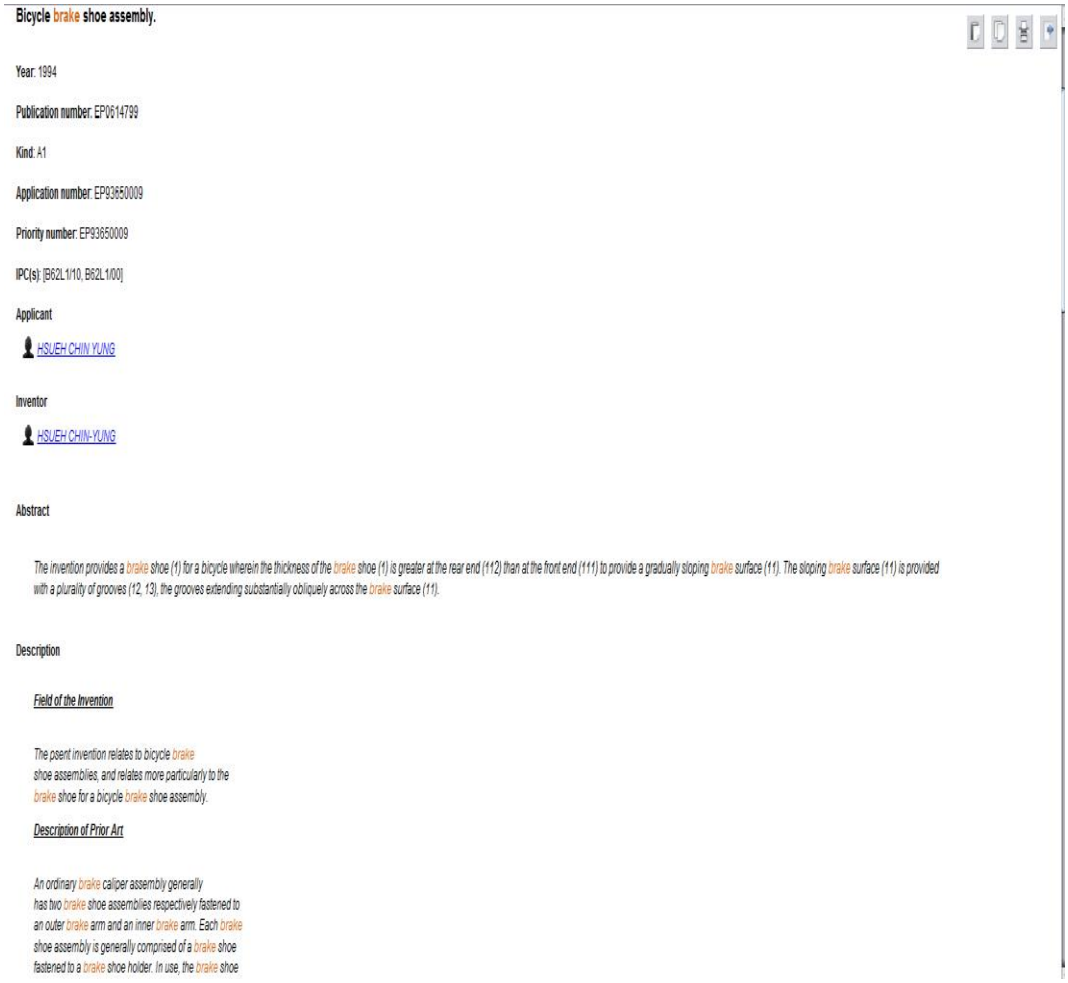

Fig. 11 Details View - Before

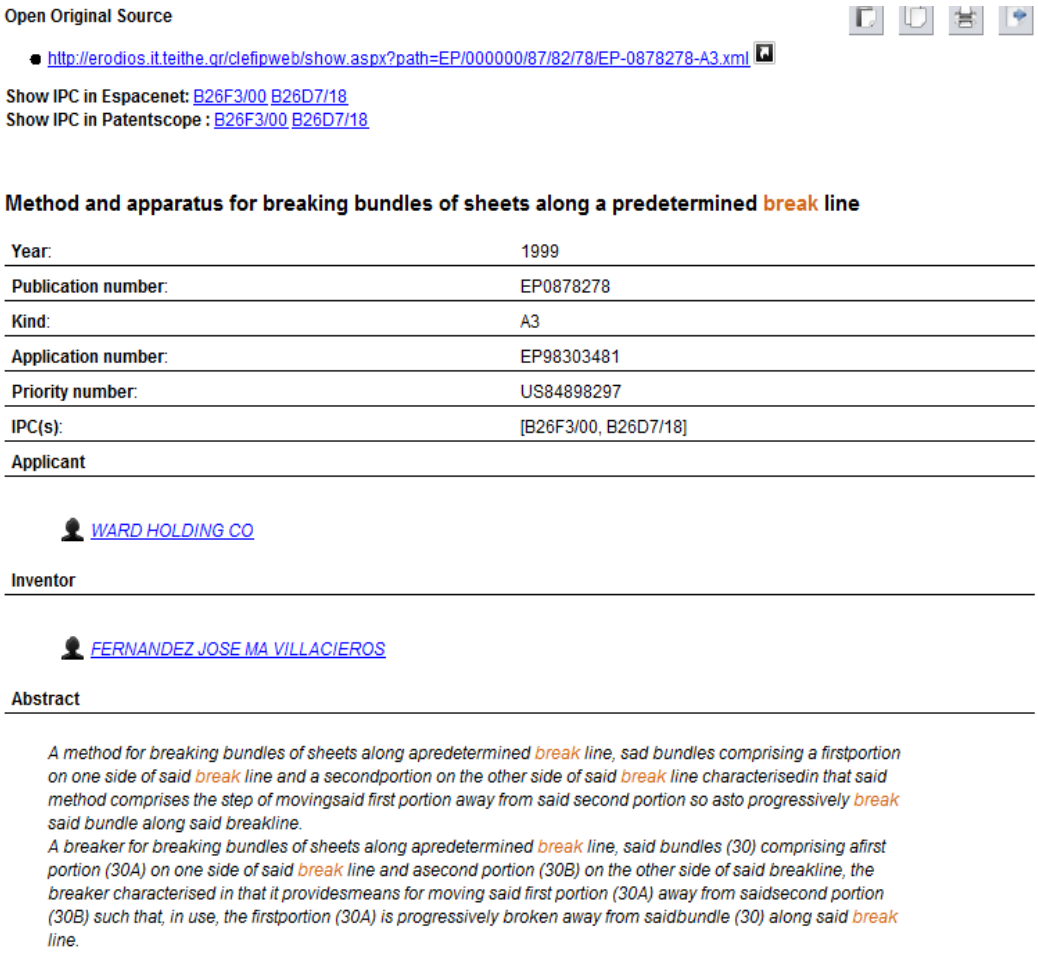

Fig. 12 Details View - After

6.2 Results View

- Handling not available information: As seen in figure 13, not available information (title or applicant) was represented by a question mark. This was replaced by "*Applicant/Title not available*".
- Automatic grouping: Before, when a user chose a feature to group by the results, s/he needed to press the Group By button.
  Now, the grouping is done once the user chooses a feature to group by and a waiting cursor is displayed while the grouping process takes place.
- CPCs in the snippet: As CPC was introduced to PerFedPat, it was needed to be displayed alongside the IPCs in the snippet of each result.
- Term filtering extension: Before, the term filtering did not support Boolean operators. This was extended to support AND, OR and NOT operators.
- Removed the sort by group size check box. That is because it was deemed unnecessary, because this functionality is provided by the Entities Explorer tool.

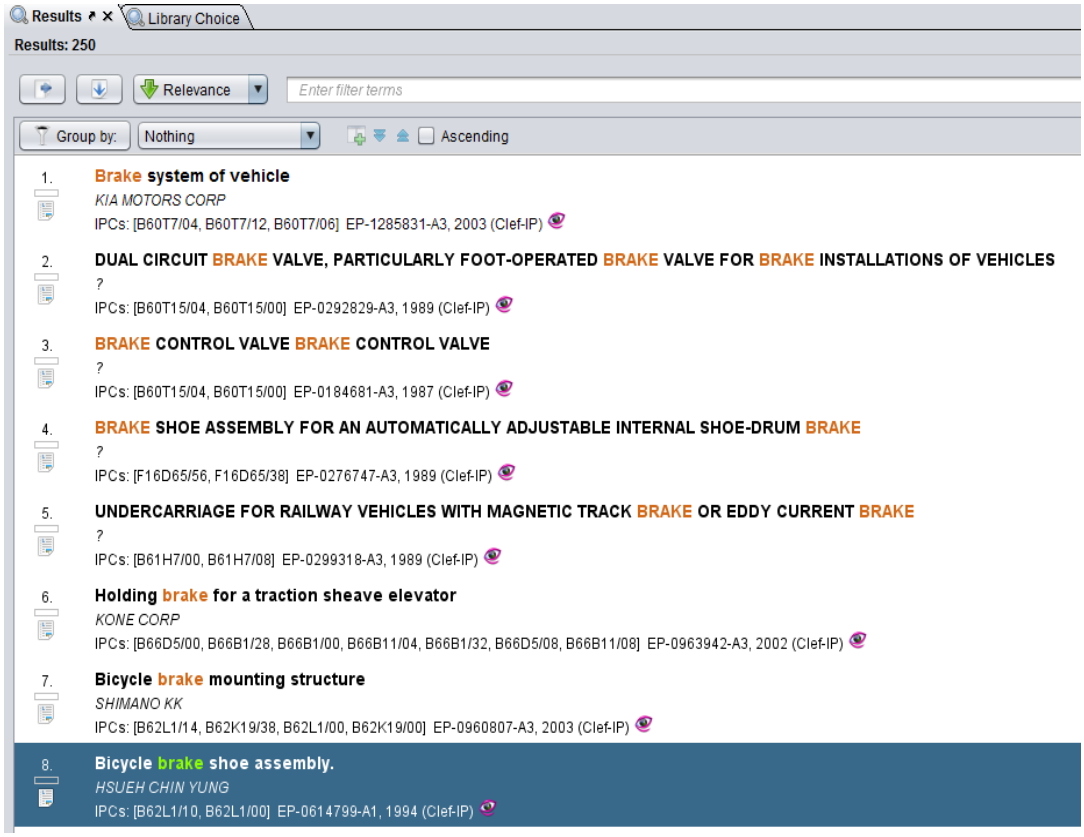

Fig. 13 Results View - Before

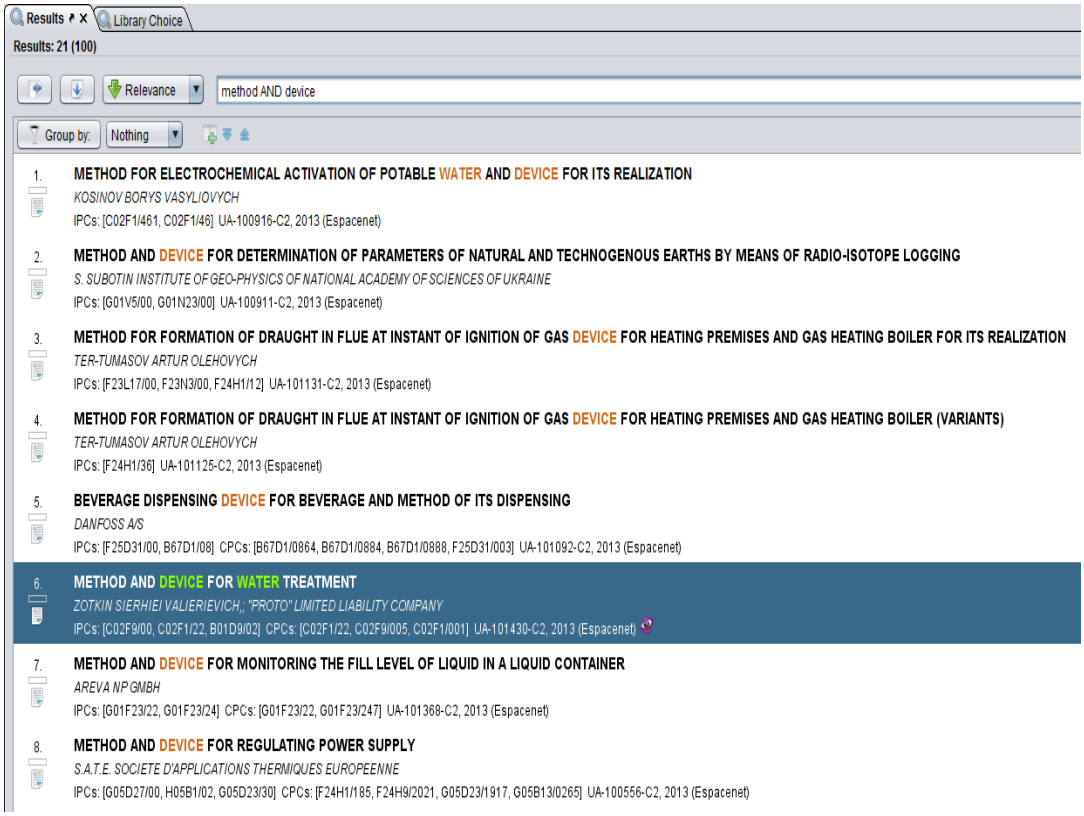

Fig. 14 Results View - After

6.3 *Query View*

- Classifications input: Before, the classifications were inputted without the symbol '/', which is normally part of the classification.
  This symbol is now supported but the classification must be enclosed in quotation marks (see Figures 15 and 16).
- Field examples: The field examples were updated for the reason above.
- Info icons: Information icons were implemented to provide more information about the support of each field from the resources, when the field is moused-over.

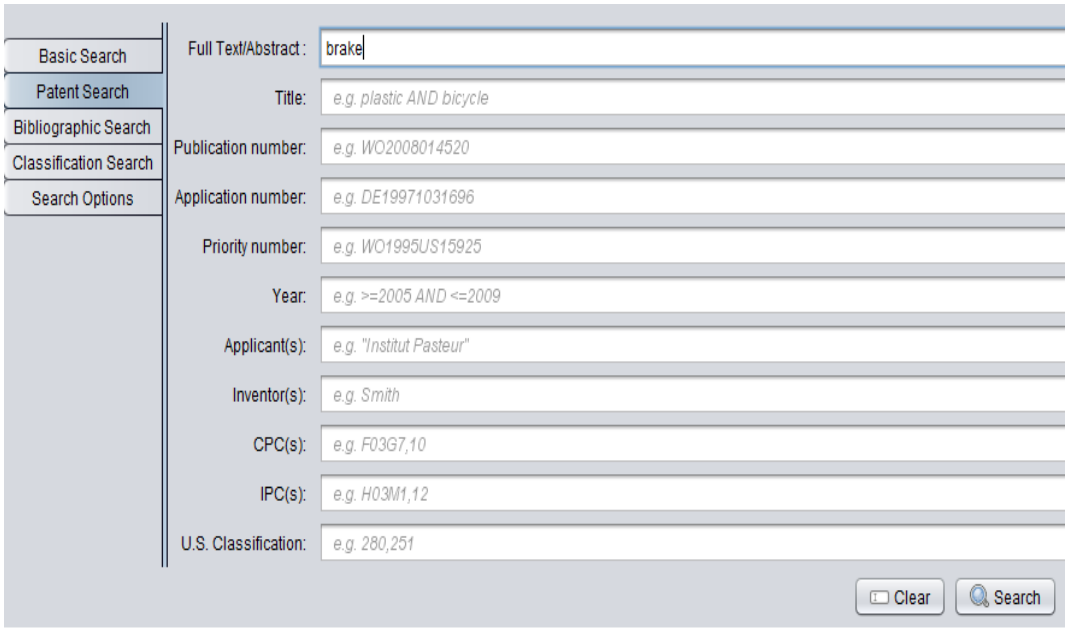

Fig. 15 Query View - Before

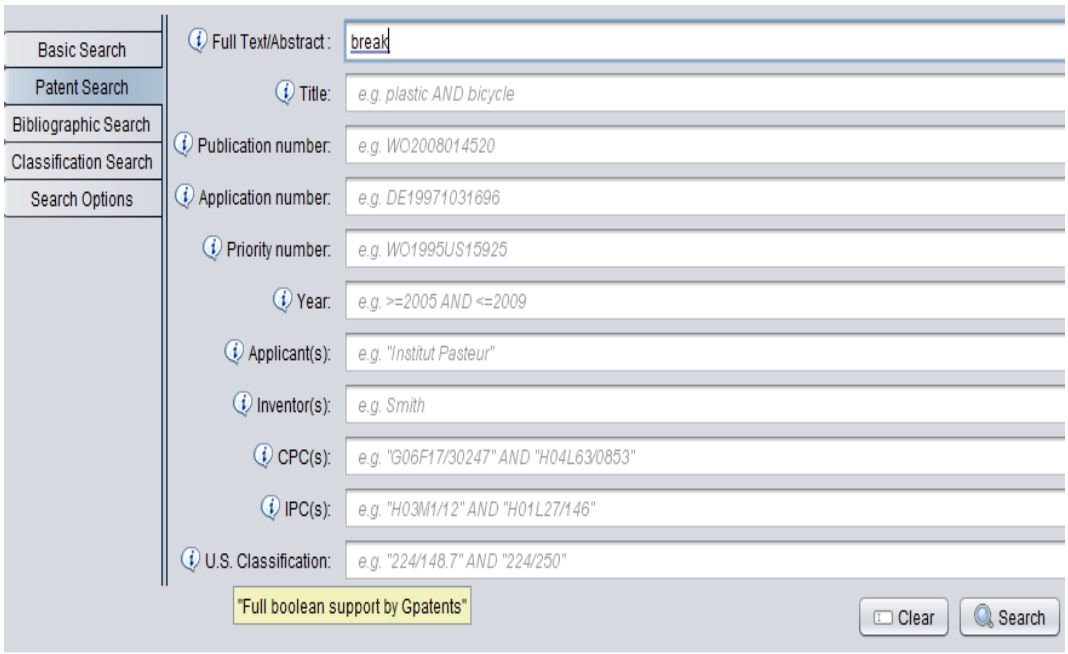

Fig. 16 Query View - After

6.4 *Patent Tool View*

- Browser error: There was a long standing error in the browser of the patent tools, that caused the web page to stop loading. This happened because the browser had a mechanism to parse any web page whose URL included the word "xml". This mechanism was removed, as the web pages of the tools do not load xml files, although they have the word "xml" in their URL.
- History buttons: History buttons were added to navigate, which is especially useful for the Cluster and Entities Explorer tools.
- Rearrangement: As seen in the figures below, an aesthetic change was made to make the Patent Tool View more user friendly.

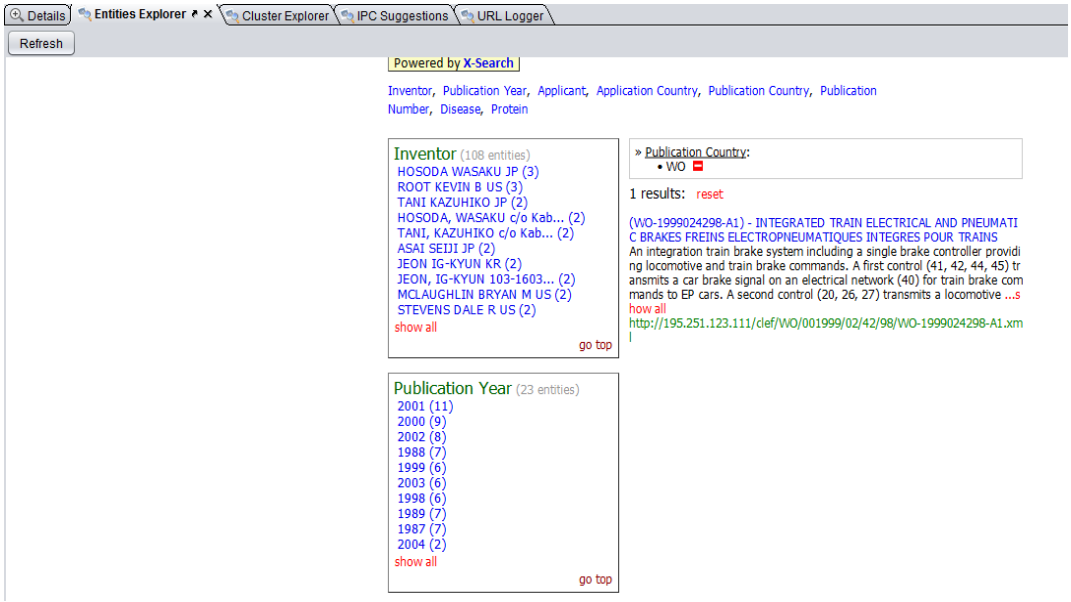

Fig. 17 Patent Tool View - Before

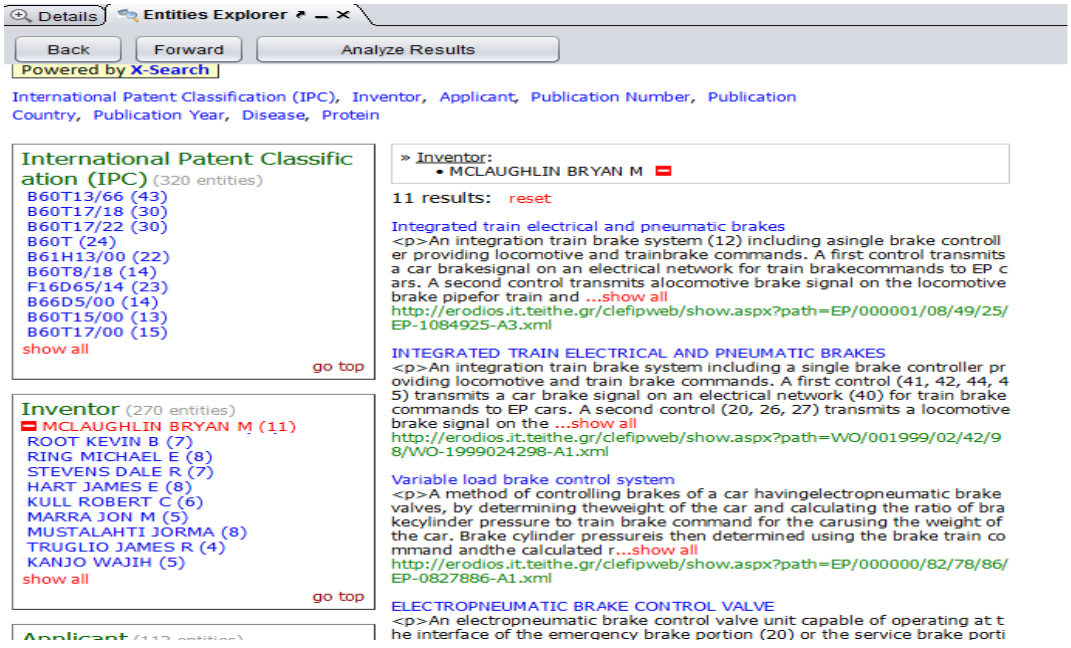

Fig. 18 Patent Tool View - After

6.5 Library Choice View

- Update to the capabilities: Every patent resource was tested and some changes to their capabilities were made.
- In addition to the above, the text of the single and multi wildcard capabilities were registered wrong.

6.6 IPC Suggestions Tool

- Default level 4: Set the default level to 4, even if the user has not chosen a level.
- Extract Top 5 IPCs: The functionality that is described in chapter 5.4.1, paragraph 4 was implemented.

6.7 Entities and Cluster Explorer Tools

Before, these tools worked only with the CLEF resource. The client would send a URL that described the query that was used in the search process to the third party (X-Search). Then, the third party would execute the same query to the CLEF resource and analyze the results. This method was slow and not very efficient.

Now, these tools work with all the resources and, additionally, more metadata (such as classifications, publication year, etc.) are analyzed by the third party. All the search results are written in an XML file on the backend and then the backend sends the URL of this XML file to the client. The client sends this URL to the third party and the third party downloads the XML file and analyzes all the results.

6.8 URL Logger

The following changes were made for experimental reasons only:

- Pause/Resume Session: A pause/resume session button was added. When the session is paused, all the logged events that take place in this period are ignored.
- Level of Logging: Four radio buttons were added to choose the level of logging (for more details, see Chapter 6.10, Extension of the logging system).

6.9 Wrappers

- Patentscope: Before, Patentscope wrapper was returning only 10 results because the web page of the results could not be analyzed thoroughly using XPath (see Chapter 5.2.1).

  Now, the wrapper can return more than 10 (50 by default) results. The analysis of the web page is still done using XPath, except for the URL of the next page of the results, which is done programmatically.

- Fix analysis: All the wrappers needed a fix in their analysis process, because their resources' APIs (CLEF, Espacenet) and web pages (Google Patents, Patentscope) were updated.
- Bug: CLEF and Espacenet were returning some results twice. This happened because they were executing some queries twice.
- Change of the URL: CLEF and Espacenet changed their base URL.
- Change in the query: Espacenet wrapper required a change in the query that was to be executed when there was a year range included.

6.10 General

- Results bug: There was a bug that caused the Results View to be reset. This happened because the results would be sent twice from the Search Agent to the client as a result of the Search Agent not being halted when it should have been halted.
- Extension of logging system: More user events are now logged. The new logged events are organized in 3 levels, depending on their importance.
- ECLA with CPC: ECLA (European Classification) was replaced by CPC (Cooperative Patent Classification) in 2013.
  There were all the necessary changes made to the client and the backend.
- Unit tests: All the necessary unit tests were written.
- A warning message before the splash screen is shown, if the user's Java version is not up-to-date.

6.11 External

A dynamic ASPX web page was developed which transforms an XML patent document of CLEF resource to a HTML web page. The URL of this HTML web page is displayed in the Details View of every CLEF document.

7. PerFedPat Use Cases

In this chapter, possible use cases of the application are presented.

7.1 Perspective

The default perspective of PerFedPat is depicted in figure 19.

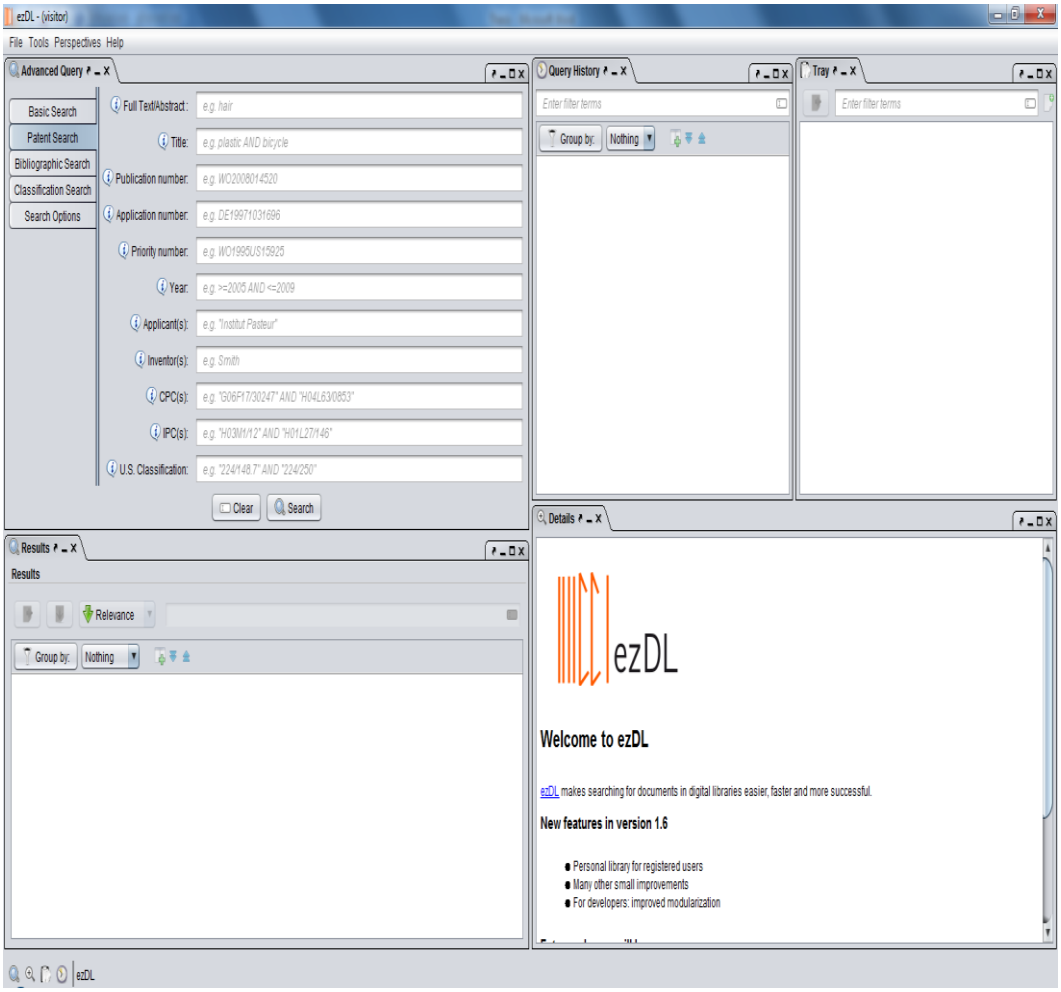

Fig. 19 Default Perspective of PerFedPat

Perspectives can be modified, saved and reset (Figure 20).

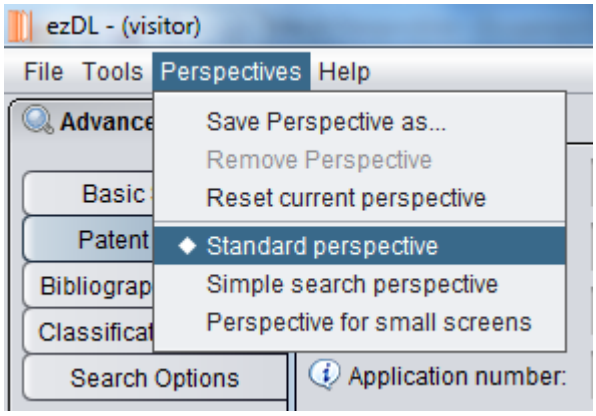

Fig. 20 Perspective Options

7.2 Merging/Reranking

To enable Merging and/or Reranking, go to Search Options tab and check Merging and/or Reranking.

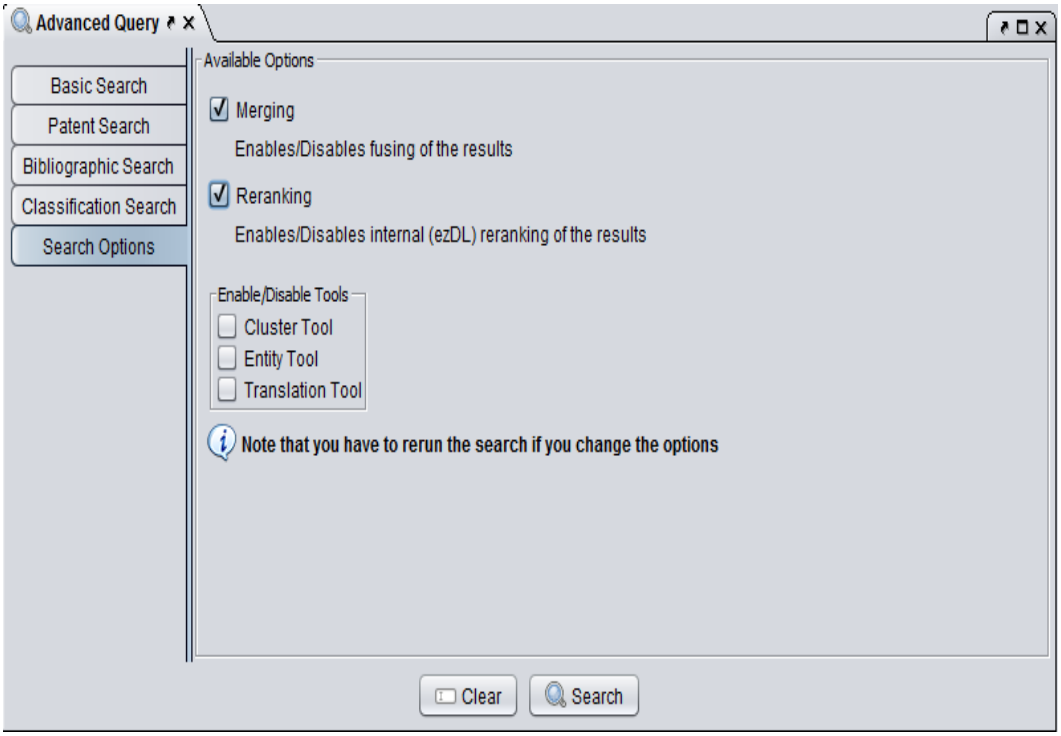

Fig. 21 Merging/Reranking

7.3 Patent Search

Step 1: Open the Library Choice tool.

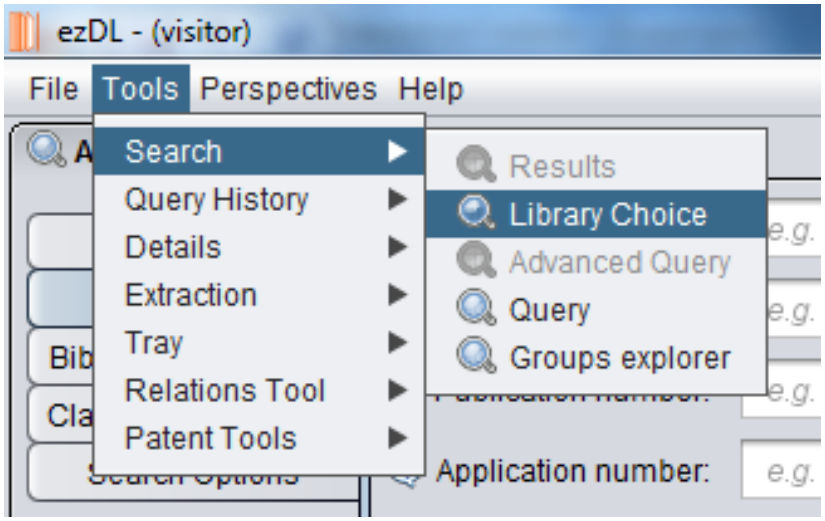

Fig. 22 Patent Search - Step 1

Step 2: Filter the resources by choosing "Patent Resources".

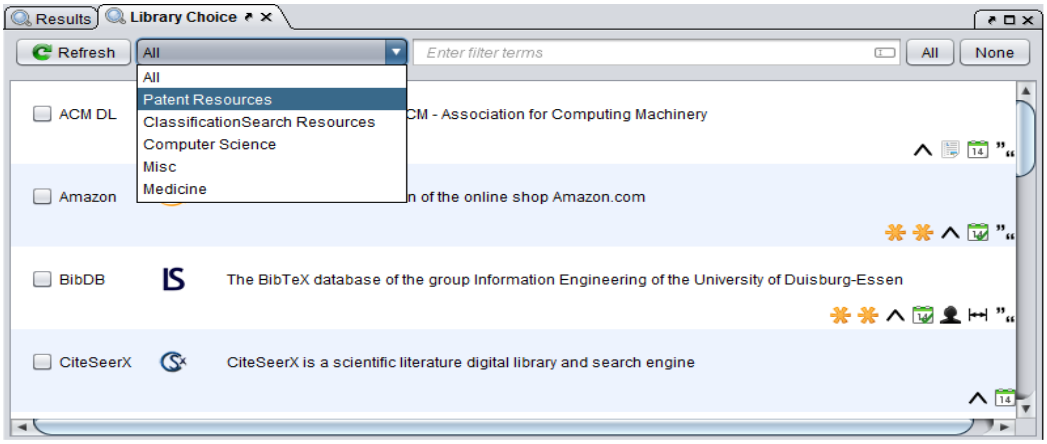

Fig. 23 Patent Search - Step 2

Step 3: Choose the resources.

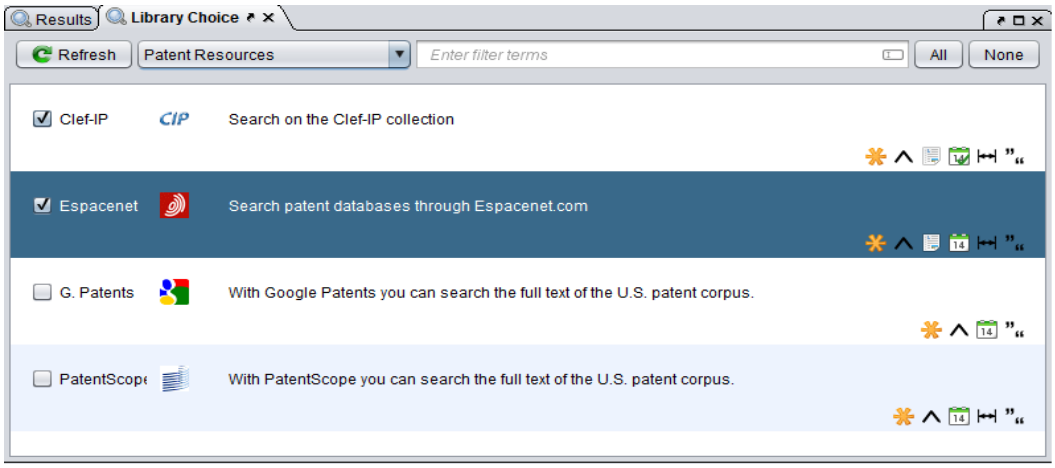

Fig. 24 Patent Search - Step 3

Step 4: Construct the query in the Patent Search tab and press "Search".

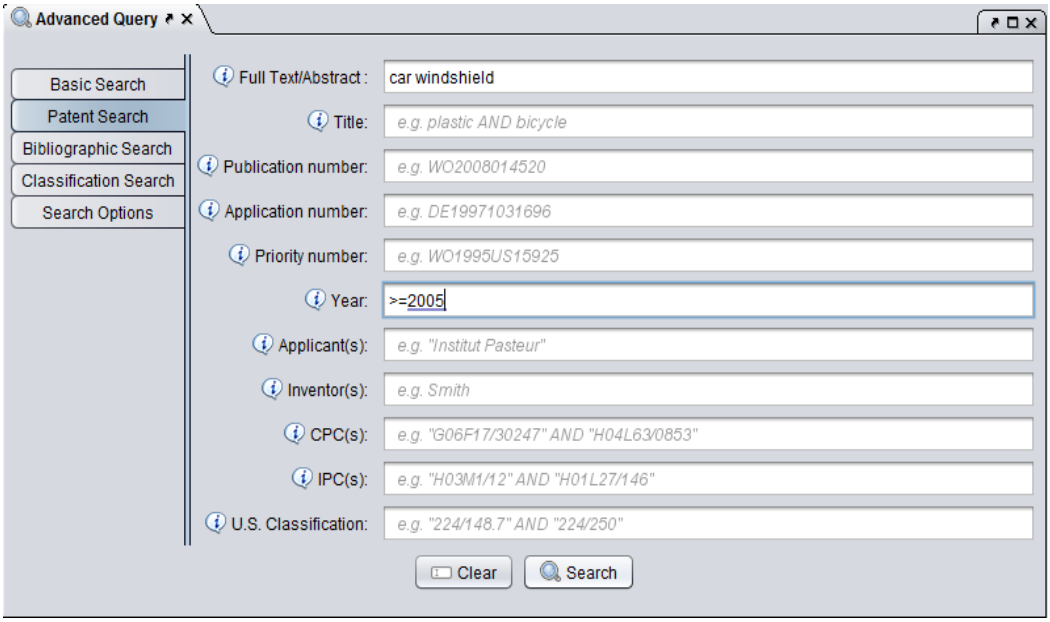

Fig. 25 Patent Search - Step 4

Step 5: Click a result in the Results View and view the details in the Details View.

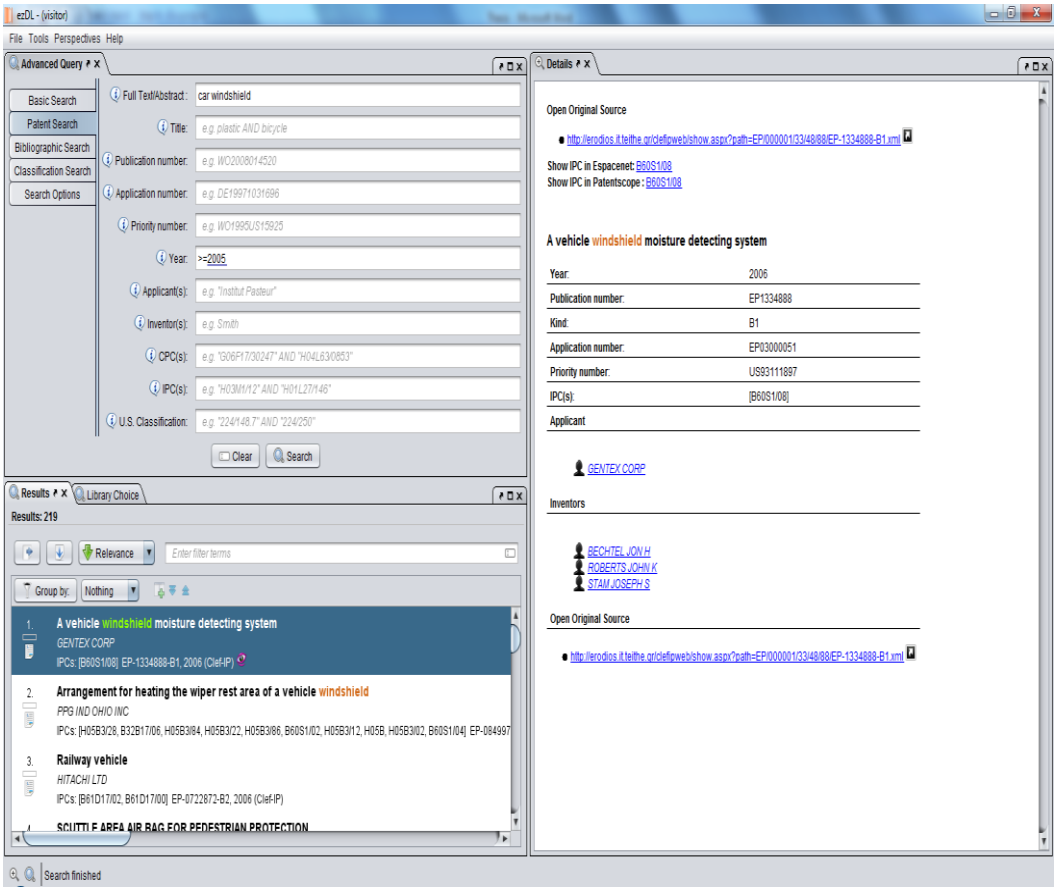

Fig. 26 Patent Search - Step 5

7.4 Group By Feature

Select a feature to group by from the combo box in the Results View.

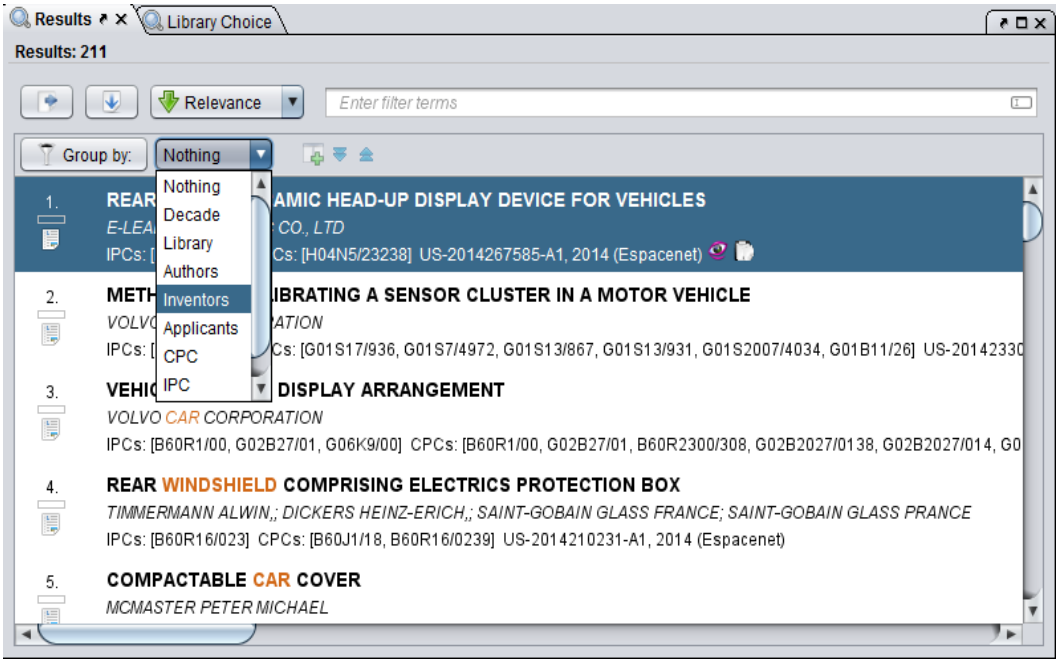

Fig. 27 Group By Feature

7.5 Filter Terms

Enter filter terms in the field in the Results View. AND, OR and NOT operators are supported.

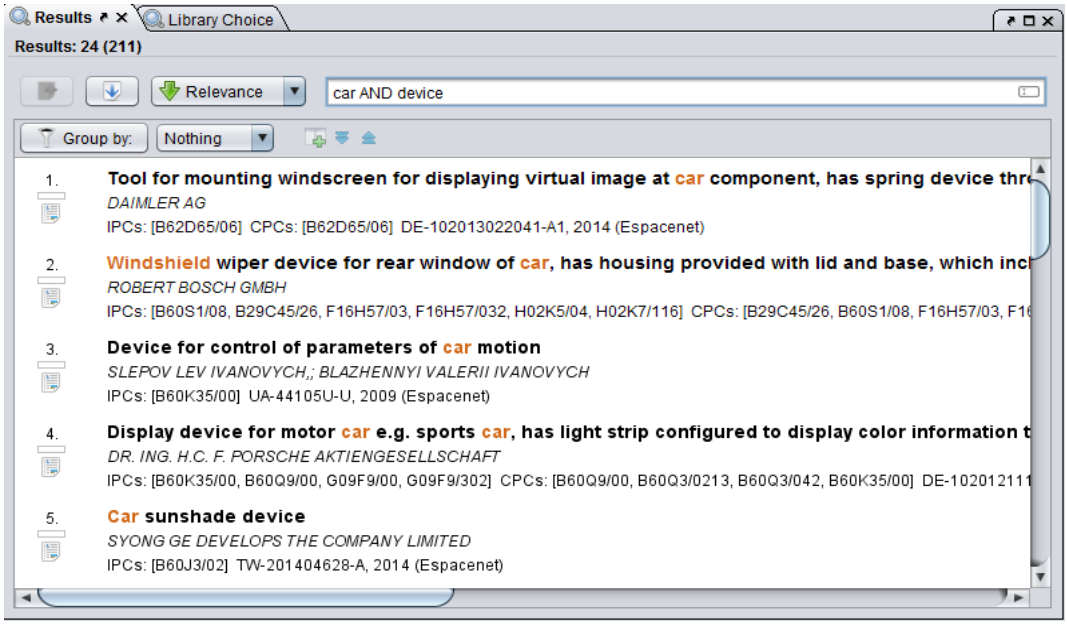

Fig. 28 Filter Terms

7.6 Add to Tray

Step 1: Open the Tray tool.

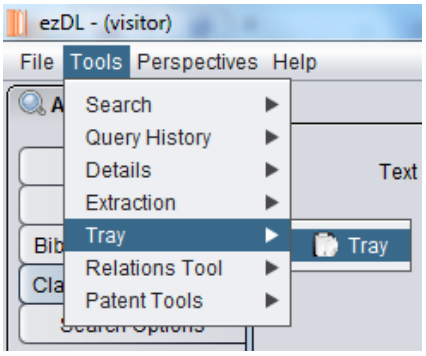

Fig. 29 Add to Tray - Step 1

Step 2: Drag and drop a patent from the Results View to the Tray.

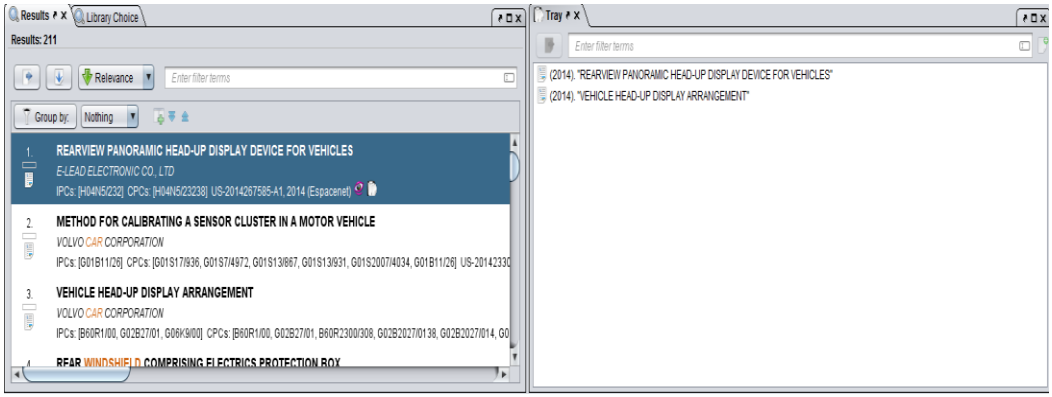

Fig. 30 Add to Tray - Step 2

7.7 *IPC Suggestions Tool*

Step 1: Open the IPC Suggestions tool.

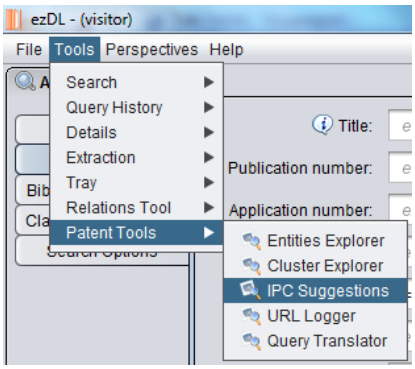

Fig. 31 IPC Suggestions Tool - Step 1

Step 2: Enter the terms in the Classification Search tab, choose the IPC level and click IPC Suggestions button.

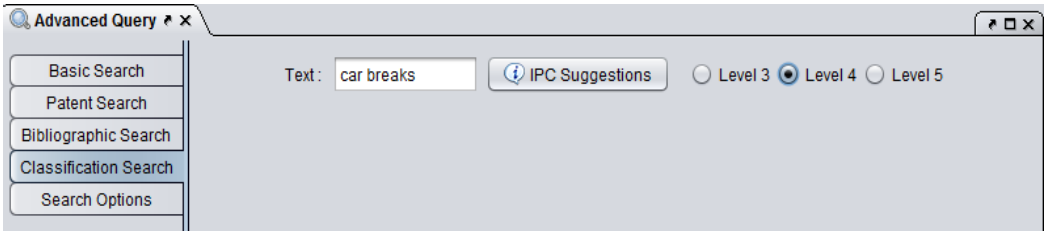

Fig. 32 IPC Suggestions Tool - Step 2

Step 3: Click the "Copy Top 5 IPCs to Clipboard" button in the IPC Suggestion tool.

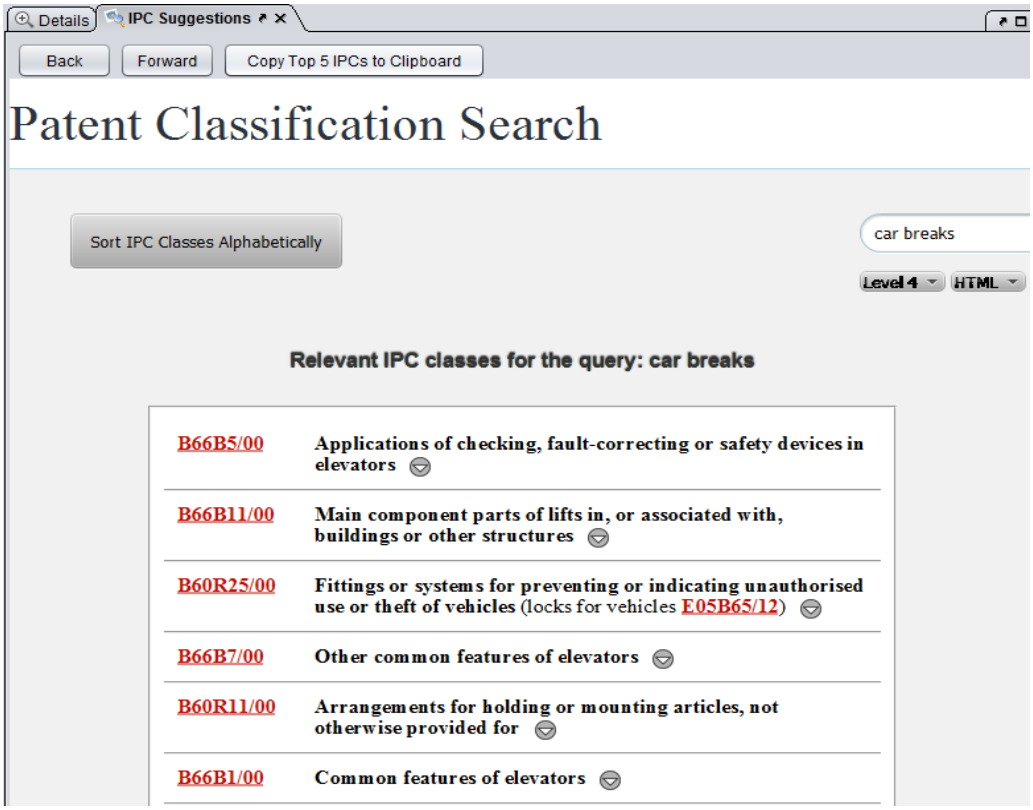

Fig. 33 IPC Suggestions Tool - Step 3

Step 4: Go to Patent Search tab, right click on the IPC(s) field text box and choose Paste -> As characters.

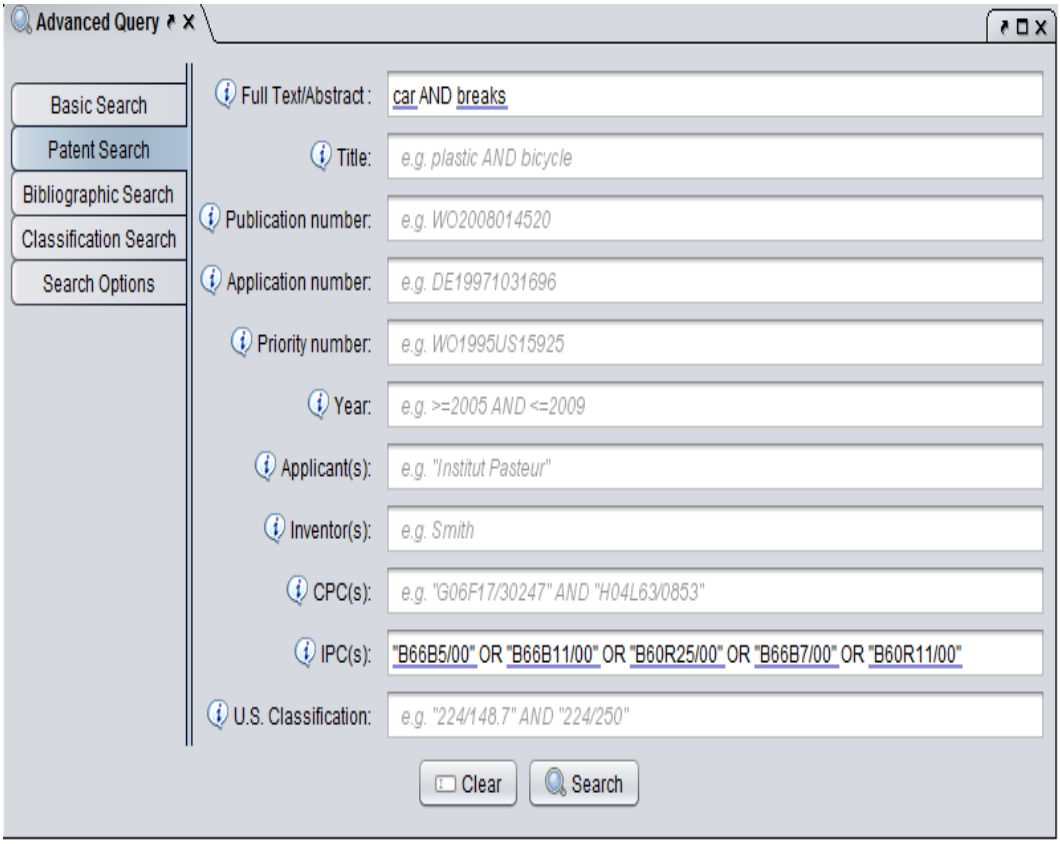

Fig. 34 IPC Suggestions Tool - Step 4

7.8 Entities and Cluster Explorer Tools

In order to use Entities and Cluster Explorer tools, a search must be completed.

Step 1: Open the Entities and Cluster Explorer tools.

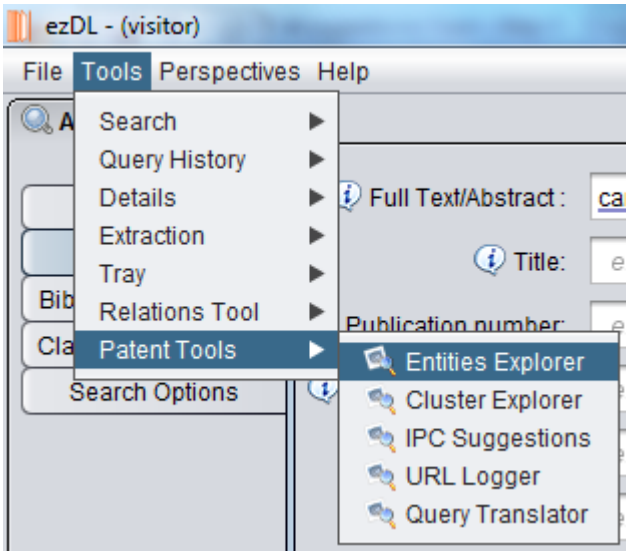

Fig. 35 Entities and Cluster Explorer Tools - Step 1

Step 2: Click the "Analyze Results" and "Cluster Results" buttons in the Entities Explorer and Cluster Explorer tools respectively.

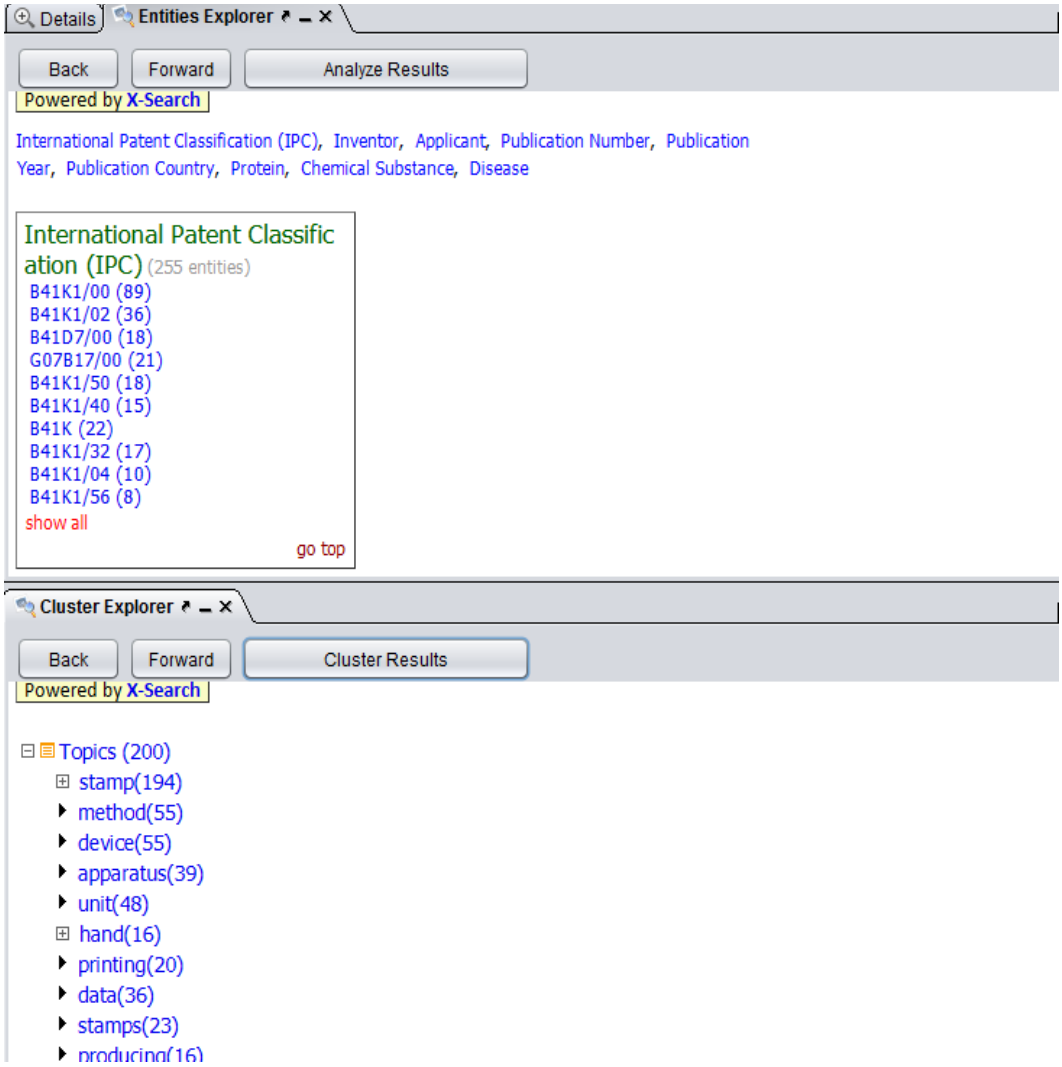

Fig. 36 Entities and Cluster Explorer Tools - Step 2

7.9 Query Translator Tool

Step 1: Open the Query Translator tool.

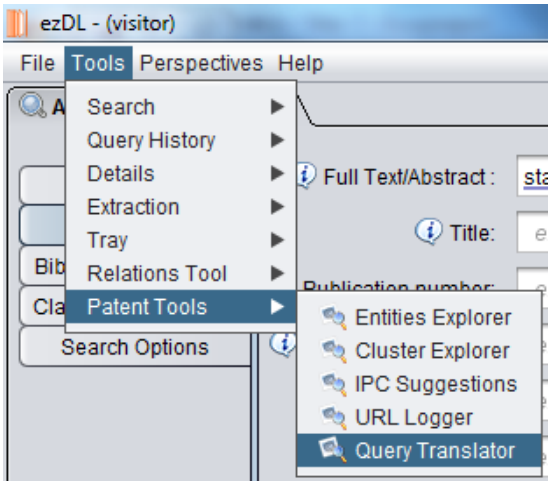

Fig. 37 Query Translator Tool - Step 1

Step 2: Choose MT Service, Source Language, Target Language(s) and enter the query terms.

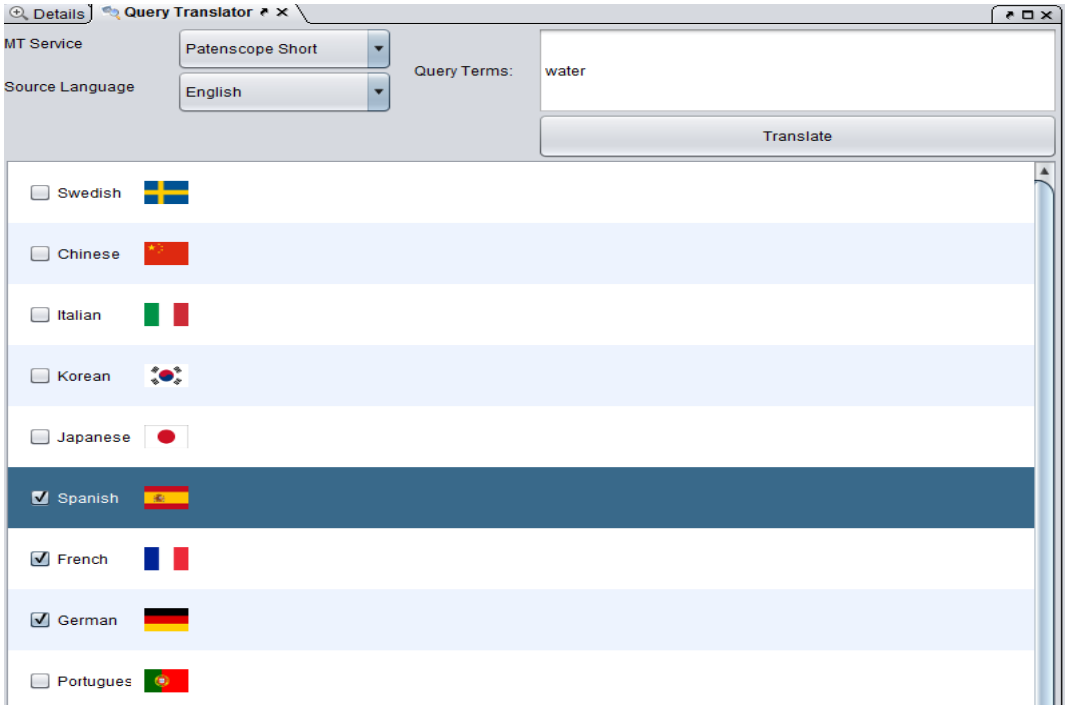

Fig. 38 Query Translator Tool - Step 2

Step 3: Click the Translate button. The translated query will be automatically passed to the Query View.

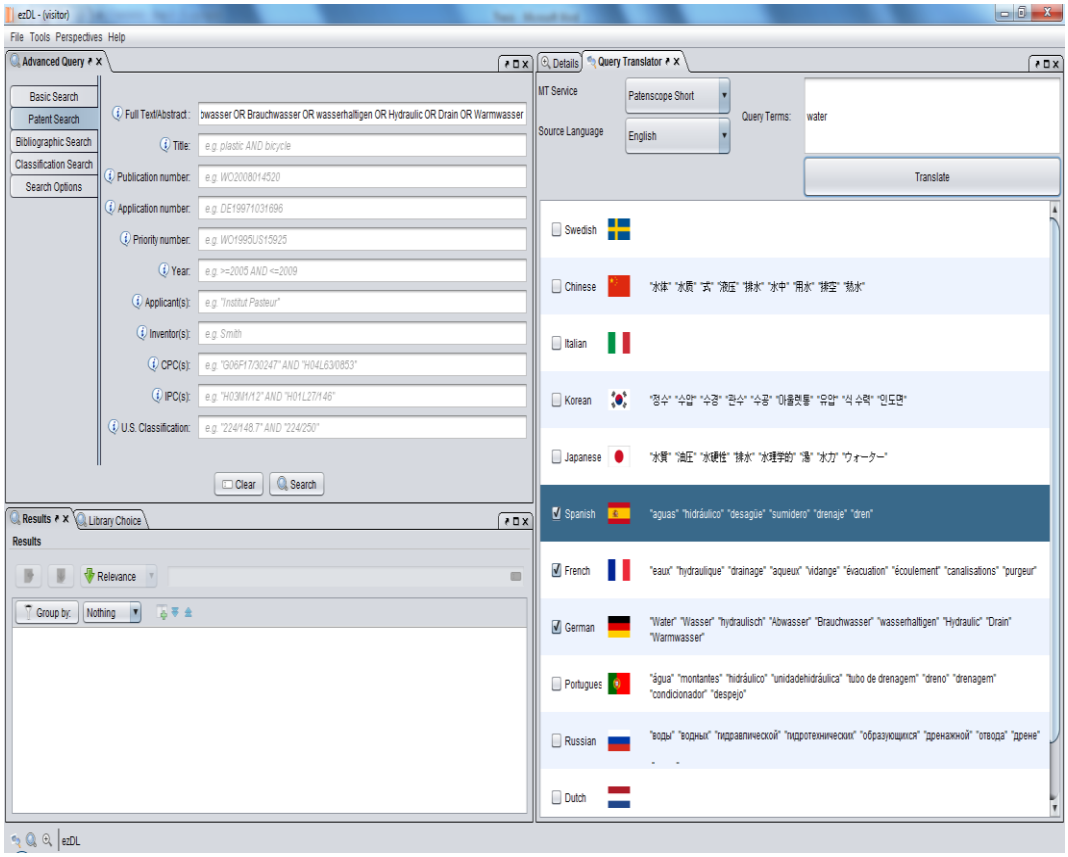

Fig. 39 Query Translator Tool - Step 3

8. Conclusion

Professional search in the patent domain usually needs both an analytical and an exploratory type of search which is characterized more often, in comparison to fact finding and question answering web search, by recall-oriented information needs and sometimes by uncertainty and evolution or change of the information need.

Federated search can become an important technology for developing patent search systems that could potentially play a useful role in some particular settings, when crawling and maintenance of a centralized index is not possible.

The PerFedPat system was inspired by the design idea of providing an integrated patent search system which will be able to provide a rich information seeking experience for different types of patent searches, potentially exploiting techniques from different IR/NLP technologies.

I believe that PerFedPat demonstrates the feasibility and the applicability of federated search for patent searching. PerFedPat provides core services and operations for being able to search multiple online patent resources, thus providing a unified single-point access to multiple patent sources while hiding complexity from the end user. More patent resources can easily be made part of the PerFedPat federation to increase coverage or for reasons of specialized searches that some patent systems may provide.

The other important aspect of PerFedPat is the tight integration of search tools which can be utilized during an information seeking process by the professional searcher, depending on the task type, the stage of the task, the experiences and objectives of the end-user. Based on the ezDL framework, core tools were extended and tools specifically for patent search were developed and in this way the feasibility of the proposed PerFedPat architecture was demonstrated.

During the development, I realized how intelligent the system's architecture is because it allows extendibility and parameterization.

The system can be easily extended, at the functionality level as well as at the presentation level by integrating tools in the application's frontend or by adding resources in the backend.

Parameterization is possible because of the agent-wrappers. They run as autonomous programs/services and, therefore, can be activated according to each user's needs. Also, their functionality can be easily edited making it easy for the system to adapt to the changes that were made to each wrapper.

PerFedPat can be used side by side with other patent search tools by patent officers. Also, since PerFedPat 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.

In the future PerFedPat 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 be implemented, giving PerFedPat access to a very big data set of patents.

Also, having as much results as possible per search is not enough. That is where 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.

In conclusion, I think that federated search and systems such as PerFedPat represent a promising approach for patent retrieval and therefore could play an important role in the development of next generation patent search systems.

9. References

[1] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, "Introduction to Information Retrieval", *Cambridge University Press*, 2008

[2] Michail Salampasis, Allan Hanbury, "A Generalized Framework for Integrated Professional Search Systems", *6th Information Retrieval Facility Conference*, 2013

[3] Michail Salampasis, "Rethinking the Search Experience: What professional search systems could do better?", *7th Metadata and Semantics Research Conference, MTSR*, 2013

[4] Milad Shokouhi, Luo Si, "Federated Search"

[5] Michail Salampasis, "Federated Patent Search"

[6] Thomas Beckers, Sebastian Dungs, Norbert Fuhr, Matthias Jordan, Sascha Kriewel, "ezDL: An Interactive Search and Evaluation System"

[7] Thomas Beckers, Sebastian Dungs, Norbert Fuhr, Matthias Jordan, Georgios Kontokotsios, Sascha Kriewel, Yiannis Paraskeuopoulos, Michail Salampasis, "ezDL: An Interactive IR Framework, Search Tool, and Evaluation System", *Professional search in the modern world, Springer LNCS Series*, 2014

[8] Michail Salampasis, Allan Hanbury, "PerFedPat: An Integrated Federated System for Patent Search", *World Patent Information, Elsevier*, 2014

9.1 Websites

[1] http://www.perfedpat.eu/

[2] http://www.ezdl.de/