# RESEARCH AND INNOVATION THROUGH THE PROINOVA PATENT MANAGEMENT SYSTEM

**Ágnes Sebestyén-Pál**

**Gheorghe Sebestyén-Pál**

Technical University of Cluj-Napoca

Computer Science Department

Cluj-Napoca, România

Agnes.Sebestyen.Pal@gmail.com

**Marius Bulgaru**

Technical University of Cluj-Napoca

TCM Department

Cluj-Napoca, România

Marius.Bulgaru@tcm.utcluj.ro

**Abstract**: Rezultatele precum şi succesul activităţilor de cercetare şi inovare sunt în mare măsură dependente de un proces de documentare bine fondat şi de disponibilitatea şi utilizarea unor surse de informare de calitate. După cum este susţinut în această lucrare, brevetele au un mare potenţial şi sunt o sursă valoroasă de informare cu condiţia ca disponibilitatea şi utilizarea lor să fie activată şi promovată printr-un sistem de gestionare eficientă a brevetelor. Deşi există mai multe baze de date de patente oficiale care oferă acces online la documentele brevetelor se constată o lipsă considerabilă de servicii de administrare a acestora şi a unui sistem integrat de stocare pentru a încorpora aceste servicii. Această lucrare prezintă un model pentru un depozit de brevete şi descrie colecţia sa de servicii care se adresează prelucrării brevetelor. Ca atare, acest depozit tratează problema disponibilităţii patentului şi îşi asistă utilizatorii în procesul de regăsire a informaţiei prin servicii precum procesarea brevetelor, căutarea brevetelor şi generarea de rapoarte.

**Cuvinte cheie:** depozit de brevete; căutare de documente; regăsirea de informaţii; căutare semantică; clasificarea documentelor; inovare.

**Abstract:** The results and ultimately the success of innovation and research activities are highly dependent on a well-founded documentation process and inherently on the availability and the use of quality information sources. As argued in this paper, patents have a great potential as such valuable information source, with the condition that their availability and use is enabled and promoted by an efficient patent management system. While there are several official patent databases offering online access to patent documents, there is a considerable lack of patent managing services and an integrated repository system to incorporate these services. This paper presents the model for a patent repository and describes its collection of services that address specific patent processing tasks. As such, the repository addresses the issue of patent availability through patent harvesting, and assists its users in the process of information retrieval through services like patent processing, patent searching and report generation.

**Keywords:** patent repository; document harvesting; information retrieval; semantic search; document classification; innovation.

## 1. Introduction

The results and achievements of the activities involved in the field of research are known to be challenging to plan and control, making the process of innovation difficult to predict. However, there are contexts and conditions that stimulate and favor the development of innovative solutions, one of these being the availability of relevant information sources. Consequently, an important requirement in the field of research and innovation is to have the right means to support the documentation activities, involving searching for and classifying information.

Moreover, considering the great amount of information and the variety of information sources available (in the form of on-line books, journals, scientific articles, web pages etc.), there is an urgent need for tools that enable efficient and time-effective searches, resulting in meaningful documents. Simply put, this need refers not only to *finding* information, but rather to finding *relevant* information and *reducing* the obtained information amount to one manageable by a single person.

The need for such a solution is reflected by the amount of research done in the field of information processing and document retrieval. An overview of the relevant directions, strategies and fields of research is provided in [1], presenting approaches like vector-space models[2], natural language processing approaches, probabilistic models[3], latent semantic analysis[4] or neural networks.

An important issue that still remains open refers to the development of an integrated environment that combines the results of these different research areas into one composite solution and offers a platform that supports all representative processes, starting from information and document harvesting to document classification and result-synthesis through report generation.

The purpose and objectives of our work closely relate to [5] in that both approaches consider patents an efficient information source and promote the idea of using these documents for providing assistance in the innovation process. However, while Heusch focuses on a specific area (namely gearing systems) and developing a framework that automatically generates new ideas based on existing solutions, the current work addresses the more general problem of defining a set of services that provides a more efficient access to patents and a considerable support for document processing.

The rest of the paper is organized as follows: Section II discusses the potential of patents as relevant information sources and the issues related to ensuring patent access and availability. Section III introduces the proposed Pro-Inova Platform for Innovation, describing the purposes and functionality of the core set of services supported by the solution. This perspective is complemented by an architectural view that refers to the modules that implement the solution and details their requirements, responsibilities and interaction. The experimental results are discussed in Section IV and the paper concludes with the ideas presented in Section V.

## 2. Patents as Information Sources for Innovation

One relevant decision in developing an integrated innovation platform refers to the selection of the appropriate information sources. In order for the offered services to effectively provide assistance for researchers, they must be combined with an acknowledged document type. Patents represent a strong candidate for information sources for both analysis and synthesis purposes. On the one hand, patents can provide information about the progress and the historical evolution of a given field that can lead to a productive exploration and idea generation activity. On the other hand, they record state of the art innovations, delimiting the existing approaches from unsolved issues. Identifying this limit represents a considerable step in recognizing the area where innovation is possible and in converging toward a proposed approach. Consequently, patents can be efficient documentation in both the divergent and convergent iterations discussed in [6], involved in the innovation process.

According to [7] on the role of patents, these solutions determine inventors "to publicly disclose information on their invention in order to enrich the total body of technical knowledge in the world. Such an ever-increasing body of public knowledge promotes further creativity and innovation in others. In this way, patents provide not only protection for the owner but valuable information and inspiration for future generations of researchers and inventors". This statement confirms the benefits of using patents as information sources and establishes the bidirectional connection with the innovation process: patents inspire innovation and innovation produces new patents.

One advantage of patents over other potential information sources (like scientific articles, inproceedings etc.) is determined by the fact that they represent well-organized documents that provide technical information in a strictly structured, standardized form. The main sections of a patent can be seen in Fig. 1, along with the specific numbering of each type of information.

The first page of a patent presents the *bibliographic data* section of the patent, containing meta-data information about the title, the inventors, the date or the classification index of the patent. The *abstract* summarizes the subject of the patent, which is detailed in the *description* section. Some patents further divide the description section into a *background* and a *detailed description* subsection. The *claims* address the legal issues related to patents, by pointing out explicitly what

constitutes the innovation and what methods or approaches are covered and protected by the patent. The description and claims sections are complemented by representative drawings of the proposed model and present sketches of the solution.

From a researcher's point of view, relevant information can be extracted from patents considering the title and the abstract that indicate the subject of the patent, the background subsection that helps identify existing and unsolved issues, and finally the detailed description (and generally the first claim) that present a potentially inspiring approach.

Additionally, the bibliographic data section also offers information that contributes to linking similar patents and offering related references. Patents are grouped into patent families (having the same priority number), indicating that they address related issues. A higher level of organization is introduced by indexing patents according to different patent classification systems (IPC[8], ECLA[9]) that indicate the different fields and disciplines they belong to. Knowing the classification indices of a patent can help gain semantic information about its content.

The standardized structure and these classification systems make patent search a more efficient and precise operation. Additionally, a great percentage of patents are publicly available via online databases and a relevant amount of technical information is (only) accessible through patents. Hence, the need to include them as documentation is obvious.

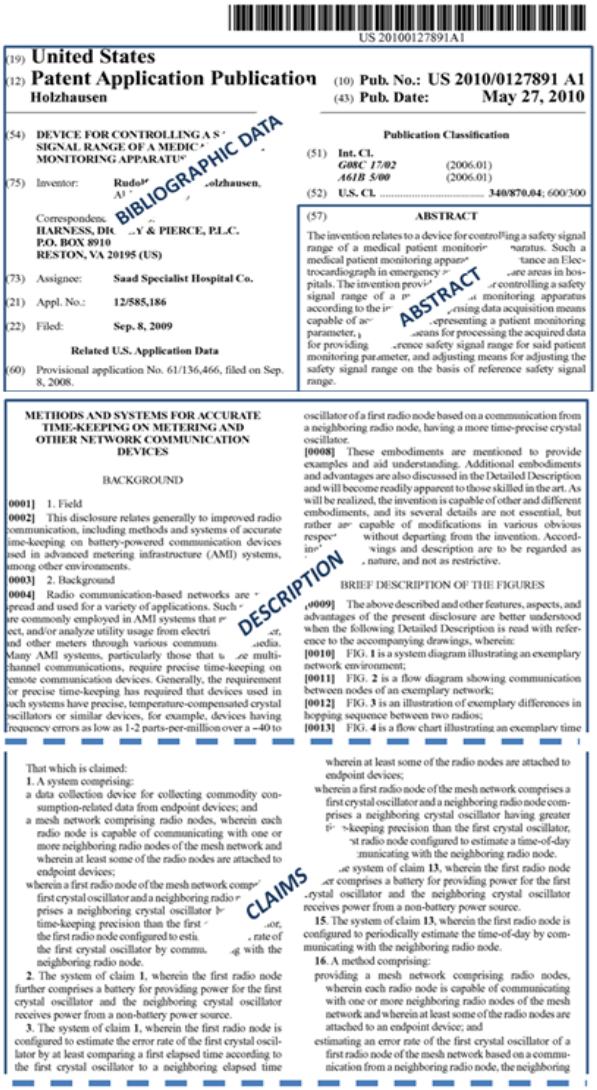


**Figure 1. Sample patent demonstrating the main sections of a patent**

The representative organizations and patent offices that manage online databases and make full-text patents publicly available include the World Intellectual Property Organization[10], the United States Patent and Trademark Office[11] and the European Patent Office[12]. The services offered by these organizations allow for users to search for patents based on keywords or bibliographic data and result in full-text patent documents that the user can view and manually download.

An important progress in accessing online patent databases is represented by the introduction of the Open Patent Services (OPS)[13] solution, developed by the European Patent Office. This solution provides a more efficient alternative for the manual patent downloading operation, in the form of a set of Web services that "deliver production stable and high quality raw patent data"[13]. Using these services, the patent downloading operation can be automated, leading to a more efficient patent harvesting component in future patent management system.

# 3. The Pro-Inova Platform for Innovation

Based on the evaluation of existing solutions in the field of information processing and document retrieval, on the one hand, and the availability of patents as information sources, on the other hand, we propose the introduction of the Pro-Inova Platform for Innovation that combines these results and opportunities into one integrated solution. The goal of our system is to provide assistance for PhD students involved in research activities and stimulate the innovation process by offering quality information about their domain of interest. As the main source of information, we consider the use of patents to obtain valuable technical information. The following subsections present the services proposed by the Pro-Inova platform and detail the modules that implement the specific set of functionalities.

## 3.1 Services offered by the Pro-Inova Platform for Innovation

The Pro-Inova Platform for Innovation covers a diverse set of services, addressing aspects like patent access and harvesting, patent processing, interactive search in the patent database and report generation. These services are presented in detail in the following.

*Patent access and harvesting* One of the main objectives of the Pro-Inova Platform for Innovation is to ensure a higher availability of patent documents for its users. The services that address this objective provide access to online patent databases, via the OPS Web services and allow for users to access remote patent databases through the interface offered by the platform. Additionally to the transparent access, the system increases patent availability by creating a local patent database through the patent harvesting service. The harvesting process consists mainly of the automated patent download operation via the OPS Web services, but it can be augmented by manual patent downloads from additional patent databases. The existence of such a database results in increased performance while searching and supports complex patent processing operations.

*Document indexing, classification and cataloging* Once the availability of information is guaranteed by the patent accessing and harvesting services, a further issue is to offer services that extract relevant sources of information and reduce the result-set of a search. Our approach considers the development of information processing algorithms, including indexing, semantic analysis and document classification, in order to implement data structures (like index DB and patent catalogs) that increase the performance of patent searches.

*Interactive patent search* The service that combines the results of patent harvesting and patent processing is the patent searching service. The input for this service is a user-defined query specifying the information need, and the service responds by returning a list of relevant patents answering this information need. The performance of this module is highly dependent on the selected search method used for retrieving patents and the precision expected from the system. In case of index-based searches, the search operation is made more time-effective by relying on

pre-computed auxiliary data structures, namely the index databases. In cases like full-text searches, however, there is no preprocessing step that can be executed offline to speed up the search and therefore, these searches take longer to execute, but in the meantime offer more relevant results.

Selecting the appropriate approach depends on the specific needs of the user and the quality expectation with regard to the results. In order to provide a flexible, full-featured system, several searching methods are provided by the platform, allowing users to decide on their approach.

*Report generation* In order for the system to offer a structured and customizable result as a response to a search, a service is provided for generating a synthesis report. This functionality presumes that the user is only interested in certain sections of the resulting patents (for example the title and abstracts) and therefore provides a mean to define report templates, that identify the sections of interest and their order in the report. Based on such a report template, the search results can be presented to the user in a summarized format, eliminating the irrelevant sections.

## 3.2 The general architecture of the Pro-Inova Platform for Innovation

At the basis of the development and integration of the services listed in Subsection A, stands a modularized architecture, decomposed into subsystems responsible for implementing core services of the system. In the following, these subsystems are presented in detail, illustrating the building blocks of the Pro-Inova platform.

*The harvesting module* The harvesting module (presented in Fig. 2) implements the patent access and harvesting services provided by the platform. It takes into consideration a harvesting policy, through which the user defines and delimits the field of interest in terms of the topic and period of relevant patents. Based on these specifications, patents from online patent databases, such as the EPO, USPTO and WIPO are downloaded and added to the domain specific patent database.

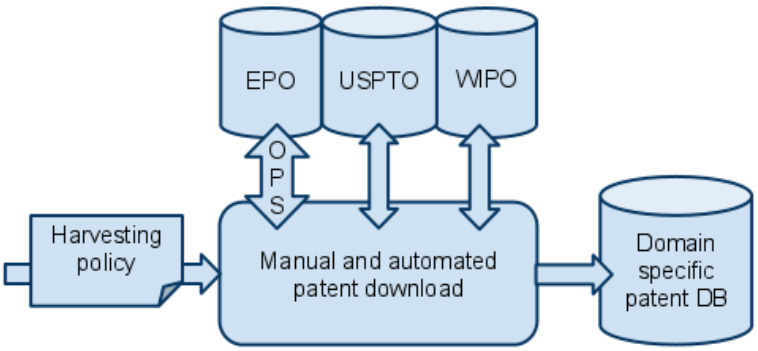


**Figure 2. The harvesting module**

The automation of the patent downloading process is mainly implemented by using the OPS v2 Web services that establish a connection with the EPO online patent database. The specifications provided in the harvesting policy are translated into the Common Query Language (CQL)[14] supported by the OPS services and a search query is created. A SOAP message request is sent containing this query, invoking the bibliographic search service of the OPS.

The response to the request lists the patents satisfying the harvesting policy available in the patent database. Using the identification information obtained from this list, the content of each patent can be obtained individually. The sections of a patent are provided through specialized OPS services, including the bibliographic data retrieval service (offering the bibliographic information, the title and the abstract of the patent), the claims retrieval service and the description retrieval service. Once these sections are obtained, they are compiled into a single

patent text file and this text file can be added to the domain specific patent database.

Since the method presented above only guarantees the download of textual information, we have also integrated an alternative tool for automated patent downloading, using a full-document patent downloader component. This component connects to the European publication server(EPS)[15] via the Web service described in [16] and provides the same facilities as the online interface offered by the EPO site. The request sent to the EPS server contains the identification information of a patent and a specification regarding the preferred format of the resulting patent. The available choices are PDF files or XML files, with the option of requesting the drawings belonging to the patent, as well.

Having access to patent documents in PDF format is a preferred alternative for users over the text files compiled from the OPS services, since they also contain suggestive drawings of the patents. However, the possibilities of using this component are restricted by the limited flexibility offered by the EPS Web service (EPS-WS). Firstly, the EPS-WS only provides patents available in the European publication server (having the "EP" index), similarly to the online interface, which considerably reduces the number of downloadable patents (unlike the OPS services). Secondly, in order to request a patent through this service, the identification information (namely the patent publication number) must be known prior to defining the request.

Since the EPS-WS does not offer the means to obtain the required information, this component is used in combination with the OPS services, that include the bibliographic search service. The list of identification information obtained through the OPS services can then be used in requests to both types of services (OPS and EPS-WS), obtaining copies of the same "EP"-type patent in text and PDF format.

*The patent processing module* The patents obtained through the patent harvesting process can be used in a consecutive step as the input data for the patent processing module. Considering the fact that the methods applied for patent processing deal exclusively with the textual content of patents, this module uses the text files obtained through the OPS services and leaves the PDF versions to be processed by the user.

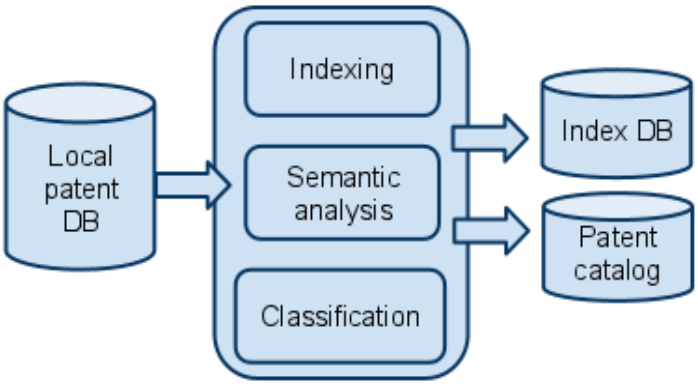


**Figure 3. The patent processing module**

The patent processing subsystem is composed of several modules, as shown in Fig. 3, each responsible for implementing a patent processing algorithm. Before the patents are processed by these modules, they pass through a preprocessing phase that parses their content to eliminate stop words and extract the relevant terms. In case of semantic analysis, words that appear in all documents or in a single document are also disregarded, since they provide no additional information.

For the patent *indexing module*, we have applied the Apache Lucene[17] text search engine library implemented in Java. During the indexing process, the different sections of the patents were considered different attributes of the patent documents, and were indexed individually. Based on the generated indices, search queries could be formulated to focus on specific attributes of patents, similarly to the advanced searches offered by the online patent databases. Additionally to the online facilities, our module extends these searches with the possibility to

search exclusively in the description or the claims sections. After the user-defined query is processed by the same analyzer as the patents, the Lucene search engine scores the indexed patents according to their similarity to the search query. This score indicates which of the patents matched the user query with regard to the search attributes.

For an alternative algorithm that supports semantic processing, we introduced the *semantic analysis module* that implements the Latent semantic indexing [4] algorithm. The development and testing of this module is still in progress, and based on the results we hope to obtain, we plan to introduce a *classification module* that defines document clusters according to the semantic similarity computed by the algorithm. However, for the time being the role of the *patent catalog* is assigned to the *index database.*

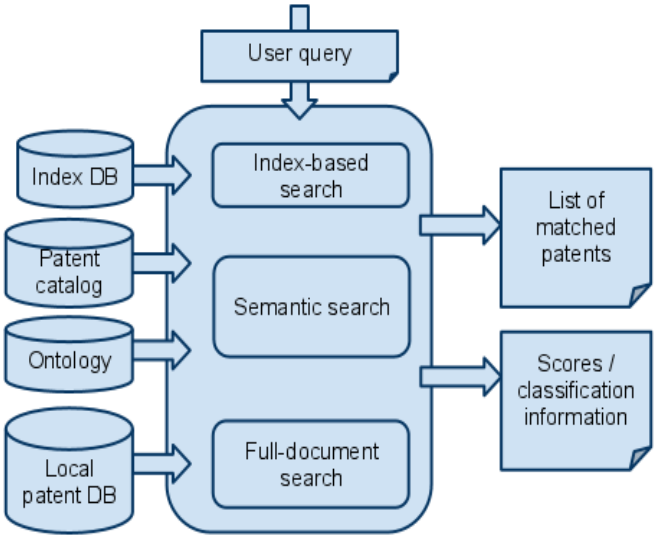


**Figure 4. The searching module**

The searching module The services supported by the searching module (Fig. 4) rely on the data structures created and managed by the processing module. Consequently, the searching module is designed to support index- or keyword- based searches, semantic searches and it is planned to be extended with a full-document searching facility. As a response to a user query, this module returns a list of relevant patents and evaluation information in the form of scores obtained from the Lucene search engine or classification information resulting from the semantic clustering of the patents.

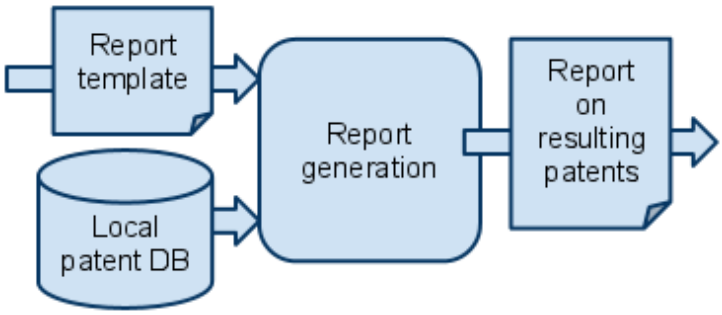


**Figure 5. The report generation module**

*The report generation module* The problem addressed by the report generation module – detailed in Fig. 5 – is to extract relevant information for the user from a set of predefined patents. The user's interest is specified in a report template that indicates which sections and what type of patent information should be compiled into the resulting report. To facilitate this selection of specific sections, we maintain a special format of patent documents, where each section represents a document attribute that can be accessed individually with no time overhead. The same format is used by the indexing module, when specific sections are indexed individually. As a result, patent sections can be extracted efficiently and the final report contains only

information requested by and relevant for the user.

A promising future extension of this module involves combining it with an *archiving module* that offers services to store and reload already defined user input. The data structures that can be reused in such a manner are the report templates and the user-defined search queries. When these structures become persistent and reloadable, users can store their intentions (formulated in user queries or report templates) and they can repeat previous actions on a new set of data.

## 4. Experimental Results

An important part of the system relies on the successful operation of the patent downloading service, therefore a considerable time has been dedicated to test and evaluate different aspects of this module. Firstly, our experiments targeted the evaluation of the interaction between the automated downloading module and the OPS services. The limitations discovered during this phase referred to the fact that the OPS services provide patent documents broken into sections, requiring subsequent requests from the downloading module. The response to these requests contains only textual information, leaving the issue of downloading patent drawing unsolved. An additional observation, identifying a problem, referred to the missing and incomplete sections, caused by the unavailability of information at the remote database level.

In order to determine the cause of these incomplete responses, the results obtained from the downloading module have been compared with the results of manual online searches (to evaluate the online databases) and with the responses obtained by using a SOAP client (to test the OPS services). The incomplete responses of our module were identical with the ones obtained by the SOAP client, demonstrating that the missing sections could not be found on the OPS server and the cause of the incomplete result was not the malfunctioning of the harvesting module. The difference in the results of the manual online searches and the results of our harvesting module showed that the OPS services have limited access to certain online databases, ensuring higher quality for European patents and providing incomplete results for patents only availably in American or world-wide databases.

During the development of our project, we have experienced an unforeseen change in the interface of the OPS services. As the OPS v1 version was replaced by the OPS v2 services, we were forced to redesign the download module, since the new version was not backward compatible and it no longer accepted the previously defined requests from our system.

Considering the fact, that after adapting to the new version, we have also experienced periods of unavailability of the OPS services, we reached the conclusion that an important requirement of our system is to attain a level of independence from the Web services used for downloading patents, an independence ensured through a customizable and flexible downloading module.

Since patent availability is an important precondition for both the patent processing and searching module, the local database creation and patent harvesting operations proved to be highly motivated. The existence of such a patent collection managed by the platform maintained the system operational even during the unavailability of the OPS services, allowing the user to continue working on the locally available patents.

## 5. Conclusions

As stated in Section I, the main goal identified and addressed by the current work refers to increasing the availability of relevant information sources and stimulating the process of innovation by providing inspiring reference material. The presented approach did not set out to introduce new document processing or information retrieval methods, but rather focuses on integrating existing solutions from the fields of document harvesting, information retrieval and patent processing into one consistent platform for innovation.

Additionally to the core set of features offered by the system in terms of functionality, the use of

patents adds to the value provided by the platform in terms of relevant content. The strict structure of these documents also contributed to the introduction of structure in the results returned by the system. The idea of using customizable report templates defined by the user represents an element of originality inspired by the use of patents and the possibility of organizing information into sections.

As a concluding idea, the proposed solution can still benefit from further refinements and extensions at the level of its components, but the need addressed by it and the utility of its services motivate the additional effort to be invested in its development.

## Acknowledgment

The results presented in this paper are part of a national project called Pro-Inova, which aims to develop an innovation environment for PhD students.

## REFERENCES

1. **GROSSMAN, D. A.; FRIEDER, O.:** Information Retrieval: Algorithms and Heuristics. The Information Retrieval Series.

2. **SALTON, G.; WONG, A.; YANG, C. S.:** A vector space model for automatic indexing.

3. **FUHR, N.:** Probabilistic Models in Information Retrieval.

4. **DEERWESTER, S; DUMAIS, S.; FURNAS, G.; HARSHMAN, R.; LANDAUER, TH.; LOCHBAUM, K.; STREETER, L.:** Computer information retrieval using latent semantic structure, 1988 (US Patent 4,839,853).

5. **HEUSCH, CH.:** Contributions regarding the processing of the structure and content of patents and development of special system and program, with application for gears. PhD thesis.

6. **LEONARD, D.; SENSIPER, S.:** Divergence and convergence in thinking – „The Role of Tacit Knowledge in Group Innovation", 1998.

7. PATENTSCOPE® WIPO's Gateway to Patent Services and Activities, http://www.wipo.int/patentscope/en/

8. International Patent Classification (IPC)

9. http://www.wipo.int/classifications/ipc/en/

10. European Patent Classification (ECLA)

11. http://library.queensu.ca/webeng/patents/ECLA_Class_Guide.pdf

12. World Intellectual Property Organization, http://www.wipo.int

13. United States Patent and Trademark Office, http://www.uspto.gov/

14. European Patent Office, http://www.epo.org/

15. Open Patent Services (OPS), http://www.epo.org/patents/patent-information/free/open-patent-services.html

16. CQL: Common Query Language, Version 1.1, http://www.loc.gov/standards/sru/sru1-1archive/cql.html

17. European publication server, https://data.epo.org/publication-server/

18. European publication server Web service

19. https://data.epo.org/publication-server/soap/Server_1_1.jws?wsdl

20. Apache Lucene, http://lucene.apache.org/java/docs/