World Patent Information 63 (2020) 101989

Contents lists available at ScienceDirect

World Patent Information

journal homepage: http://www.elsevier.com/locate/worpatin

Search strategies at the European Patent Office

Yan Tang Demey [a,*], Domenico Golzio [b]

[a] European Patent Office, Department of Business Analysis, Patentlaan 2, 2288 EE Rijswijk, the Netherlands
[b] European Patent Office, the CTO office, Patentlaan 2, 2288 EE Rijswijk, the Netherlands

ARTICLE INFO

Keywords:
Ontologies
Search strategies
Patent search
Prior-art search
BPMN
CMMN

ABSTRACT

Prior-art search is a critical step towards determining whether a patent can be granted or not. In 2016, an internal project called Search Workflow Modelling (SWM) was launched at the European Patent Office (EPO) for building a search knowledgebase, which contains a set of models that record not only the current situation on how patent examiners deal with prior-art search (i.e. the as-is models), but also their requirements of being able to do a more efficient and effective search (i.e. the to-be models). We use the Fact-based Modelling (FBM) approach for formalizing search ontologies, which cover a common vocabulary, relations between concepts related to search, and constraints applicable to these relations. We use a hybrid modelling approach of Business Process Modelling Notations and Case Management Model and Notations (BPMN/CMMN) to model search work flows. A patent search strategy typically involves at least one FBM model and at least one BPMN/CMMN model. In this paper, we will illustrate 5 types of existing search strategies (including recursive flow patterns and FBM models for future search features), and future search strategies. The SWM empirical studies in this paper are being put into practice in the ongoing projects concerning search tools at the EPO.

1. Introduction

One requirement of obtaining a patent from the European Patent Office (EPO) is that the idea behind the patent application has to be 'new' (or 'novel') if "it does not form part of state of the art" (Article 54 European Patent Convention (EPC)). Another requirement is, as defined in Article 56 EPC, an invention involves an inventive step if "having regard to the state of the art, it is not obvious to a person skilled in the art". Both requirements, namely 'novelty' and 'inventive step', are verified by examiners in a reasonably objective manner by carrying out a search in the patent and non-patent literature.

At the EPO, a search strategy is a list of search queries in a chronological order. In this paper, we broaden its definition into: a search strategy is an approach chosen by examiners to achieve the goal of successful search.

Examiners from different technical fields, i.e. Electrical Engineering, Chemistry, Mechanical Engineering, Biotechnology, etc. have different needs for the in-house developed search tools. How an examiner from a field does a search is often different from another examiner in another filed. We consider a search strategy is a set of representative search workflows from examiners in similar fields. In 2006, an internal project called Search Workflow Modelling (SWM) was launched in the IT department (now Business Information Technology, BIT, Directorate General 4, Corporate Services) in order to have working methods to capture search strategies from examiners (in Directorate General 1, Operations) and their requirements for search tools. The detailed search strategies and requirements have been further modeled in examiners' search knowledge, with which we can systematically improve the efficiency and effectiveness of search tools in the future.

This paper is a review of the SWM project. It records the following SWM results: 1) search ontologies in form of Fact-based Modelling language (FBM); 2) search workflows in a hybrid language proposed in this paper: business process modelling notations and case management model and notations (BPMN/CMMN); 3) examiners' search strategies.

The paper is organized as follows. In Sec 2, we present the paper background and related work. In SWM, We start with establishing a method of building search knowledgebase, which is illustrated in Sec 3. Then, we apply the SWM method to construct the knowledgebase, which are examiners' search strategies as presented in Sec 4. In Sec 5, we discuss the paper ideas. We conclude the paper and illustrate our future work in Sec 6.

2. Background and related work

Most research challenges of information retrieval or text mining,

* Corresponding author.
E-mail addresses: ytang@epo.org (Y.T. Demey), dgolzio@epo.org (D. Golzio).

https://doi.org/10.1016/j.wpi.2020.101989
Received 23 September 2019; Received in revised form 11 April 2020; Accepted 4 August 2020
Available online 1 October 2020

Abbreviations

ANSERA	A New Search ERA, BIT, Business Information Technology (department)
BPMN	Business Process Modelling Notations
CMMN	Case Management Model and Notations
COMBI	COMBInation of citing and cited documents
DOSYS	DOssier SYStem
EPO	European Patent Office
EPOQUE	EPO QUery Engine
FBM	Fact-based Modelling (language)
IP5	Five Intellectual Property offices
NORMA	Natural Architect for Object-Role Modelling (tool)
Rexx	Restructured Extended Executor
SWM	Search Workflow Modelling (project)
UX	User Experiences
WIPO	World Intellctual Property Office
XFull	Cross-Full (preparation)

such as search benchmarking, ranking and filtering, and information visualization, are also applicable to the research field of search in the literature in the patent world. We refer to Ref. [1] for the new research topics in the field of information retrieval.

For search in patents, the related work includes automatic search and search algorithms (e.g. Ref. [2–4]) and search methods in general (e.g. Ref. [5,6]). The main purpose of this paper is not to propose a new automatic search algorithm or some machine aided means. Instead, we want to study how EPO examiners execute a search and automatic search is one of many existing search tools used by them.

Applicants may file patent applications using a convolute language, which may make them difficult to understand the invention at the first glance. To be able to cope with this situation, patent search has additional research challenges. Another challenge is the search efficiency and effectiveness. Our literature corpus increases exponentially over years. There are currently 110 million patent documents and trillions of other types of data (e.g., meta-data of patent, class information, and non-patent literature) in our internal databases. It is still requested for an examiner to take about two days to do a search of high quality.

There are many patent search tools provided by private companies and patent offices, such as EPO Espacenet,[1] Google Patents,[2] PatentScope[3] provided by World Intellectual Property Office (WIPO) and DEPATISnet[4] by the German Patent Office. The authors from Ref. [7] have presented a comparison between Espacenet, PatentScope and DEPATISnet in some aspects, such as data coverage, search functionalities and view functions. These tools are open to the public and the targeted users are patent applicants, patent attorneys and researchers. The targeted users in the SWM project, which we will discuss in this paper, are the examiners working at the EPO.

Our colleagues have presented a case of applying examiners' search strategy in Espacenet in Ref. [8]. The search strategies are limited to the technical field of pharmaceutics. We will illustrate the search strategies that cover almost all technical fields. In Ref. [9], a structured search strategy is introduced to obtain search efficiency. Compared to the related work, which focuses on a search strategy at a high level, we focus on methods of how to systematically discover search strategies and build a knowledgebase.

EPO Examiners have been using internal search tools for decades. The main legacy internal search tool, as illustrated in Ref. [10], is called EPO QUEry engine (EPOQUE [11]). Other legacy search tools are EPOQUE preparations, which are executable queries created by examiners using a specific script language called Rexx,[5] and which are shared amongst examiners. One preparation is called Cross-Full (XFull), which is used to search in patent and non-patent literature. EPOQUE 2.0 is a new internal search tool, which is built based on the modules from an internal searching tool called A New Search ERA (ANSERA). In the SWM project, we start with analyzing why a search feature exists, why an examiner needs a particular function/feature and how he/she works with existing tools. Consequently, we also record the examiner's new requirements by observing his/her search workflow or interviewing him/her.

We use the Fact-Based Modelling approach (FBM [12]) to model search ontologies, which are a part of the SWM knowledgebase. In particular, we use the two FBM dialects – Object Role Modelling [13,14]

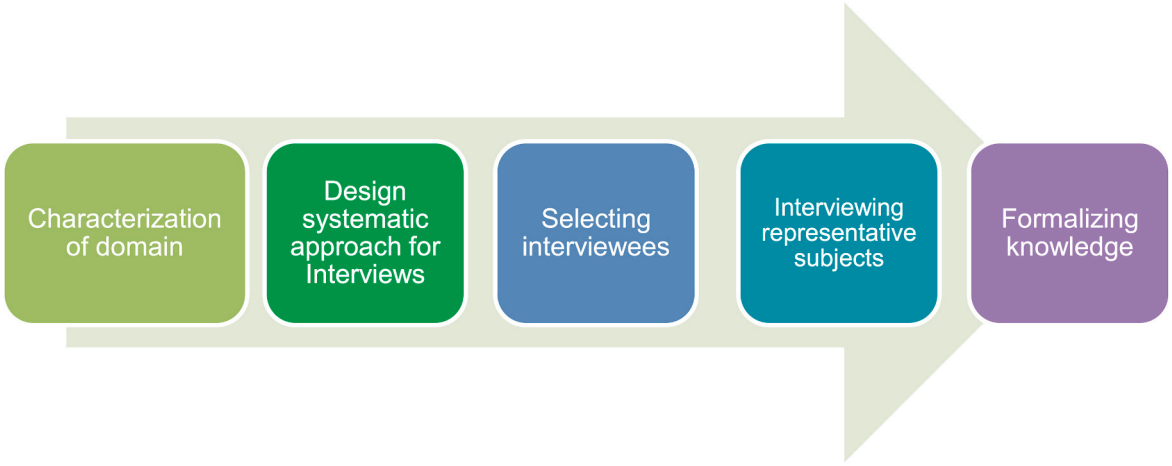


Fig. 1. The SWM method.

Step 1: Characterization of domain

[1] https://worldwide.espacenet.com/.
[2] https://patents.google.com/.
[3] https://patentscope.wipo.int/search/en/search.jsf.
[4] https://www.dpma.de/english/search/depatisnet/index.html.
[5] Restructured Extended Executor (https://en.wikipedia.org/wiki/Rexx).

How many years of search experience do you have?	○ 1-2 ○ 2-5 ○ 5-10 ○ more	
Which of the following apply to you (more than one option possible) **If none applies, but you feel you fall in a specific, relevant category, please tick the last option and fill in a name for that category in the space next to it.**	☐	**Academy teacher**
	☐	Tutor or coach
	☐	**Author of simple personal preparations**
	☐	Author of more complex personal REXX or Java preparations
	☐	**Author of / contributor to public preparations**
	☐	Developer of other personal or public tools supporting the search process
	Please name or briefly describe the nature of these preparations:	
	space for brief description of the nature of the preparations	
	☐	Involved in search related project(s) of IM
	Please indicate which project(s):	
	space for project name(s)	
	☐	Classification expert
	☐	space for name of other category

Fig. 2. Questions that derive examiner's expertise level.

Step 2: Design systematic approach for interviews

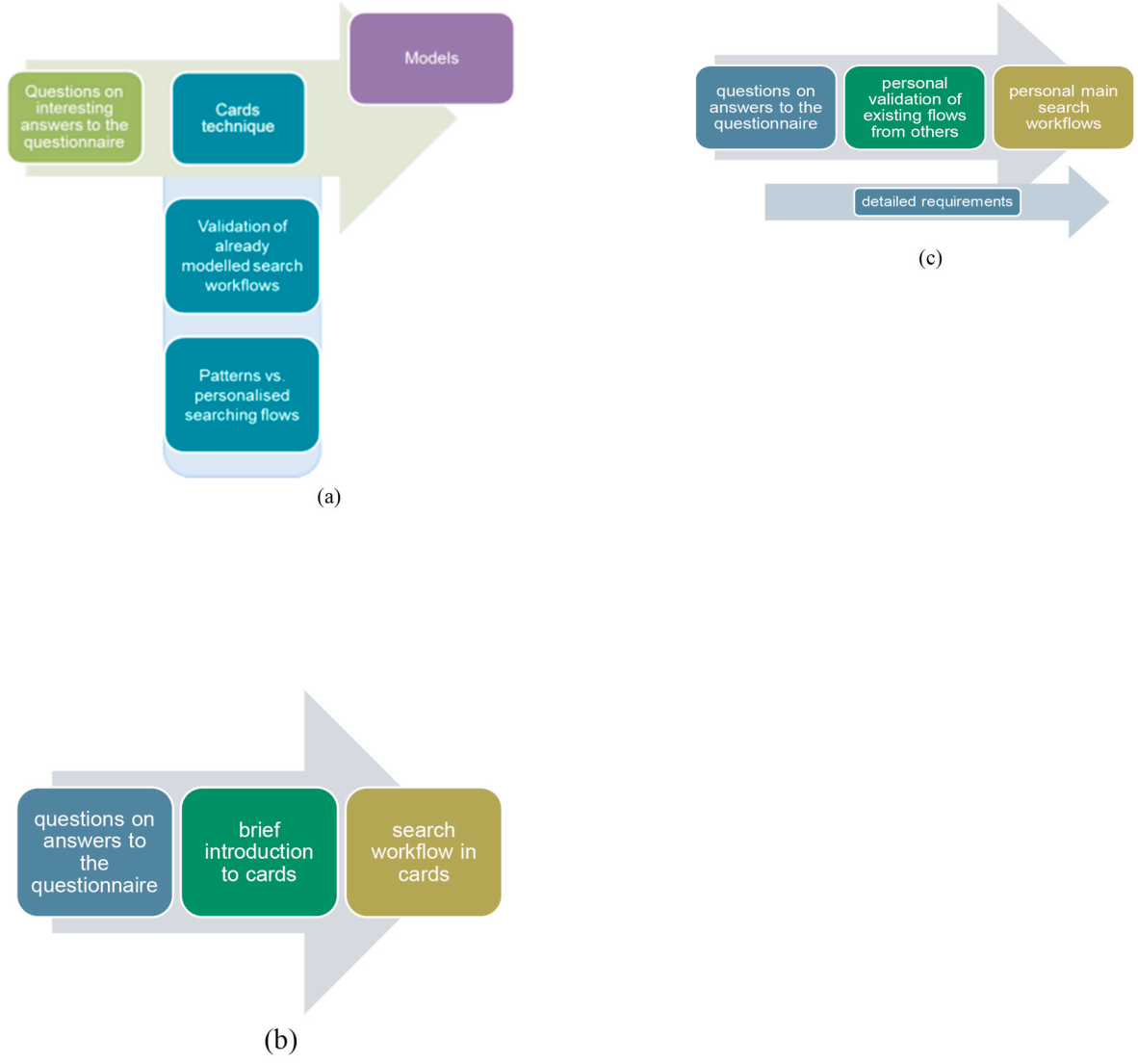
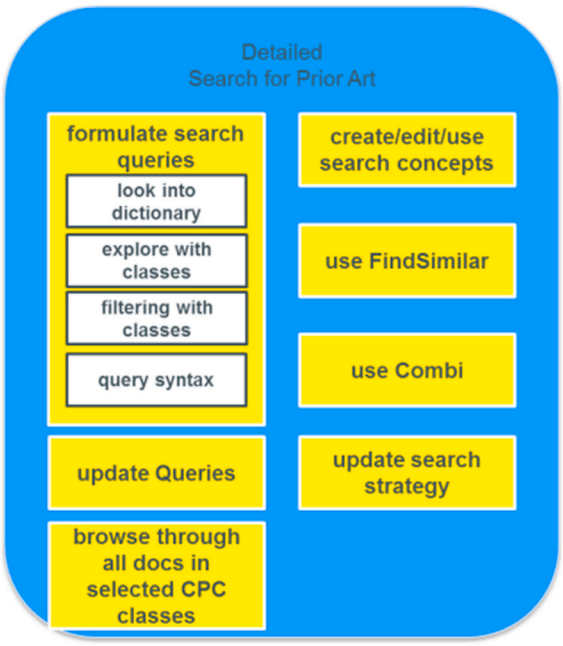
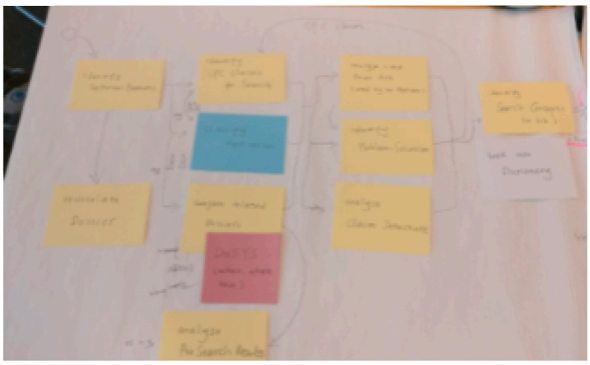


Fig. 3. Interview procedures.

Step 3: Selecting interviewees



(a)



(c)

(b)

Fig. 4. Card game in an SWM interview meeting.

and Developing Ontology Grounded Methods and Applications [15] – to create FBM models. An FBM model typically consists of a common vocabulary, relations between concepts related to search, and constraints applicable to these relations.

When a new requirement is proposed, we ask the examiner (i.e. the requirement owner) to provide facts, as in the name of FBM (fact-based). If a proposal is interesting but has no immediate supporting facts, then, if time allows, we run an analysis to gather facts. We include it in the model when the result is positive. The nature of 'basing on facts' is the main reason why we use FBM as the modelling approach.

We use a hybrid language of business process modelling notations and case management model and notations (BPMN/CMMN) to model examiners' search work flows, which are also included in the SWM knowledgebase. The flow models are annotated with the resultant FBM models.

3. The SWM methodology

The method is illustrated as shown in Fig. 1.

In this phase, we collect materials of the following types to have a rough idea of examiners' working background in general.

- Deliverables from previous projects
- Existing factsheets of search tools
- Videos recorded by other IT teams, especially the User Experiences (UX) team
- Ideas collected in sharing tools, such as EPOQUE suggestion box and internal WIKI pages

After we have made an initial analysis of these materials, we design a questionnaire, which contains 100 search-related questions and sent it to examiners. The questionnaire takes into account a number of key dimensions or categories, which exist in search: 1) examiners' technical field, 2) search type (e.g. text, class-based and figure-based), 3) examiners' experiences and personalities, 4) available search information or data, 5) available search tools, 6) properties of relevant patent applications.

Each question (also called 'competency question' in Ref. [16]) in the questionnaire is designed with a purpose from the above key dimensions. For example, the purpose of the questions illustrated in Fig. 2 is to derive examiner's expertise level.

On this step, we design how we can proceed based on the answered questionnaires and existing search requirements from the IT projects in the past. Afterwards, we design the procedures of interview meetings for

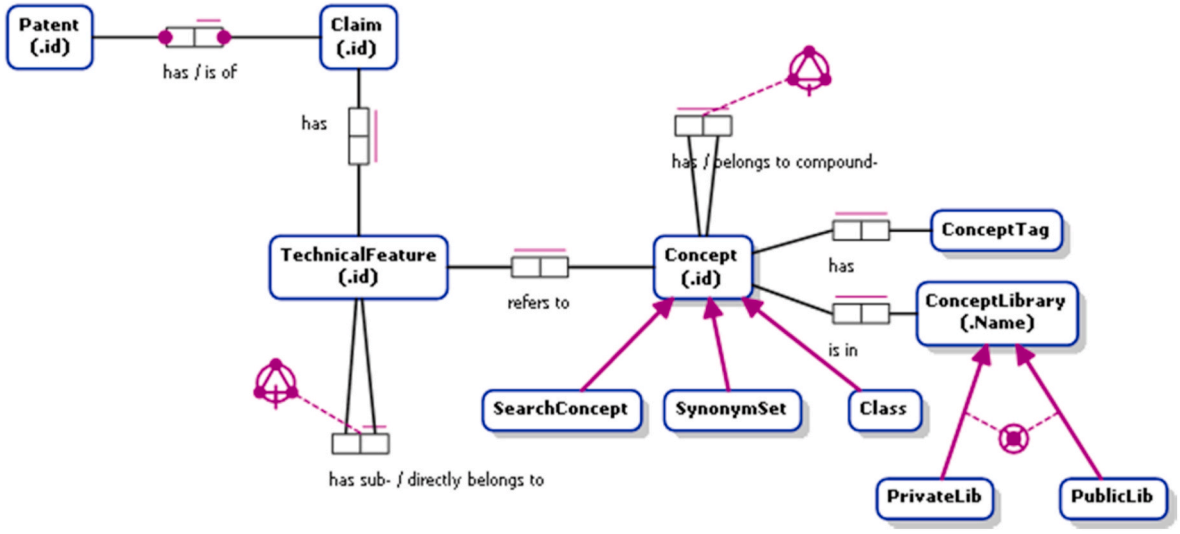


Fig. 5. Partial view of the FBM model of Search Concept.

For each ConceptLibrary, exactly one of the following holds:
that ConceptLibrary is some PrivateLib;
that ConceptLibrary is some PublicLib.

the following steps. This step is to ensure that we will have efficient meetings with the examiners who will be interviewed.

Fig. 3 (a) shows the main procedure of interview meetings. Each meeting takes a derived procedure based on our analysis. For example, we take the procedure as shown in Fig. 3 (b) for interviewing an examiner from the technical field, which has few recorded search knowledge. Fig. 3 (c) is applied when a technical field is well understood.

Based on the collected answers from the questionnaires from the previous step, we identify 'good' candidates for an in-depth interview. We also use the correlations between technical fields, search types, strategic preferences, and experience levels to prepare for interview meetings.

Step 4: Interviewing representative subjects

On this step, we select and follow a procedure designed on step 2 based on our analysis from step 3, e.g. the background of the interviewee.

We use the techniques of card game for the modelling purpose. Each card represents a task, which uses the vocabulary defined in the search ontologies. The card color indicates different levels as shown in Fig. 4 (a). There are also cards in green and red. The green cards are representative and highly repeated editorial activities, such as commenting, tagging and highlighting. The red cards are planned new features of our search tools, which will be delivered in the future (e.g. Concept Management).

We play the cards with an interviewee as shown in Fig. 4 (b) and (c). We ask the interviewee to select cards and place them in a chronological order. During a meeting, we also encourage an interviewee to create new cards, which help us to gradually build the models. For the purpose of knowing the relevance between the tasks and the interviewee's technical field, we also suggest the interviewee not to use all the cards; instead, only important ones should be selected and placed.

We record requirements as informal annotations of the models. When a requirement is complicated, we also use a whiteboard for brainstorming. When necessary, we organize a follow up session of observation at the examiner's office.

Step 5: Formalizing knowledge

We formalize the results gathered from the previous step.

We use FBM to model the ontologies. The FBM modelling principle emphasizes on natural languages as a starting point and an elicitation vehicle of the modelling exercise. By following this principle, we use terms and languages that examiners can understand. An FBM model records many interpretation-independent plausible fact types about search. Fig. 5 shows an example of FBM model in the context of Search Concept.

We use an open source tool called Natural Architect for Object-Role Modelling (NORMA[6]) to model FBM models. Fig. 5 is a NORMA screenshot.

With an FBM model, we can develop specifications for our future tools, e.g. see what follows.

This example can be further formalized in description logic ([17]) and analyzed by any open source ontology reasoners.

We have decided to model the search flows in a hybrid language of BPMN/CMMN, which allows us to group parallel tasks in a more compact way, as shown in Fig. 6. The entity types SearchConcept and ConceptLibrary from Fig. 5 are used to annotate the task called 'identify search concept (in lib)'.

As shown in Fig. 6, the levels of tasks (indicated by different colors as shown in Fig. 4 (a)) are not always respected. For example in Fig. 6, the task 'execution of search strategy' is a parent of 'identify CPC[7] classes for search', whilst yellow cards have a level higher than white cards (as

[6] http://www.ormfoundation.org/files/folders/norma_the_software/default.aspx.

[7] Cooperative Patent Classification (CPC) is a classification system (i.e. an ontology) for patent publications. https://www.cooperativepatentclassification.org.

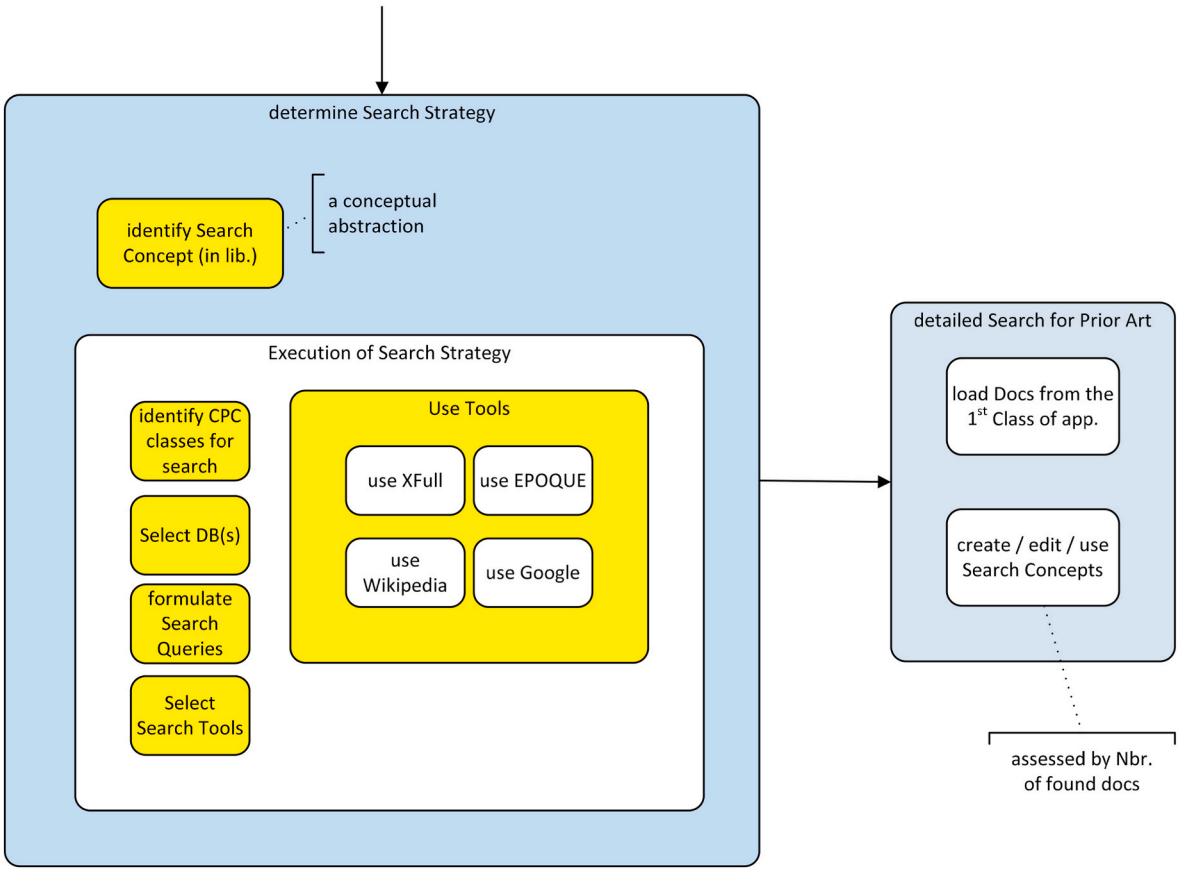
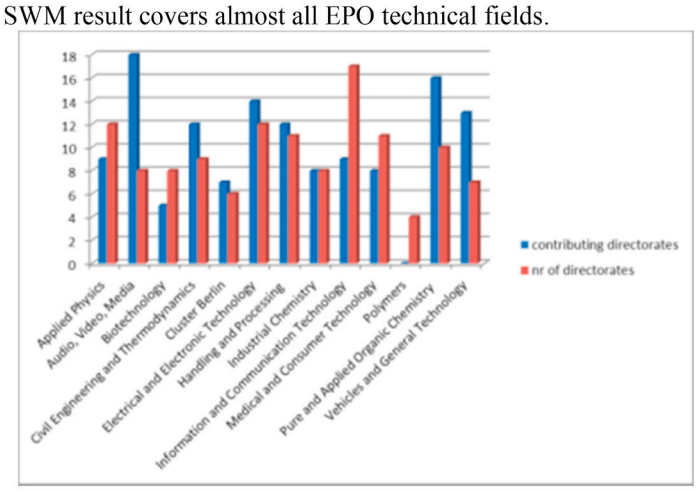


Fig. 6. Partial view of a BPMN/CMMN model.



Cluster	#Directorates in the Cluster	#Answers per Cluster	% in sample	% at EPO
Applied Physics	12	9	7	10
Audio, Video, Media	8	18	14	7
Biotechnology	8	5	4	5
Civil Engineering and Thermodynamics	9	12	9	8
Cluster Berlin	6	7	5	4
Electrical and Electronic Technology	12	14	11	10
Handling and Processing	11	12	9	7
Industrial Chemistry	8	8	6	12
Information and Communication Technology	17	9	7	17
Medical and Consumer Technology	11	8	6	8
Polymers	4	0	0	-
Pure and Applied Organic Chemistry	10	16	12	6
Vehicles and General Technology	7	13	10	7

Fig. 7. Result of characterization of domain.

designed originally in Fig. 4). We allow such flexibilities, seeing that these tasks are not at a system/functional level and we want to encourage examiners to express their needs by interpreting a task in different contexts.

4. Results: examiners' search strategies

We have received answered questionnaires from 131 examiners. 44 examiners were willing to be interviewed.

In Fig. 7, we show the result of step 1 (characterization of field). We have made a statistical analysis in order to understand whether the sample we have gathered is representative or not. In a few fields, such as biotechnology and civil engineering, the coverage of sample almost matches the reality. We can claim that the SWM result covers almost all EPO technical fields.

72% of examiners who have answered the questionnaires have more than 10 years of experience. Only 4% of them have less than two years of experience. 62% claimed to be EPO Academy tutors or coaches. There are 29% of classification experts in the population. A high level of seniority and representativeness gives us a great confidence in having

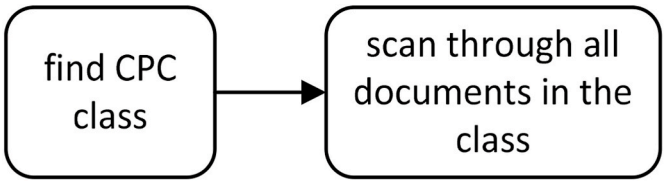


Fig. 8. A pattern in figure-based search.

The result of 'find CPC class' is a set of 1) the classes of the patent application that is being searched, 2) relevant classes in the technical field that are known by the examiner, 3) classes recommended by colleagues, especially trainees and the CPC gérant.[8] It implies that this search strategy counts heavily on the quality of CPC classes.

For most examiners of this search strategy, the task 'scan through all the documents in a class' is further simplified into 'scan through all the figures in the documents from a class'. The average number of scanned

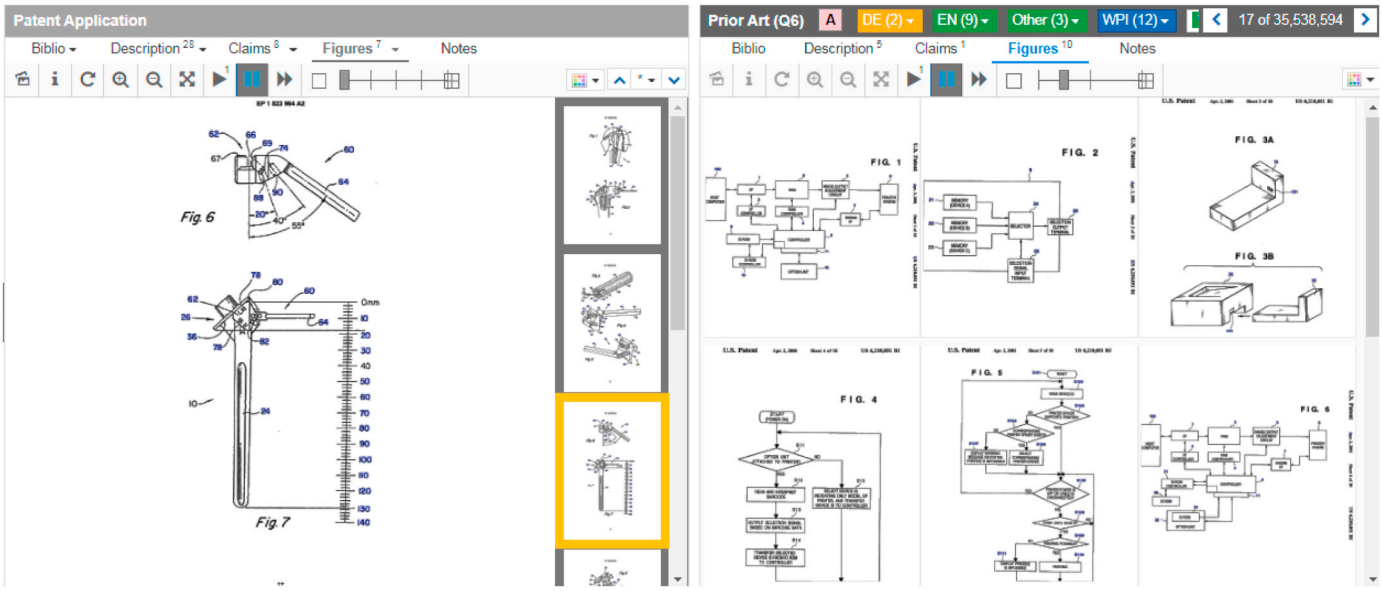


Fig. 9. Comparing figures from application and figures from a prior-art document.

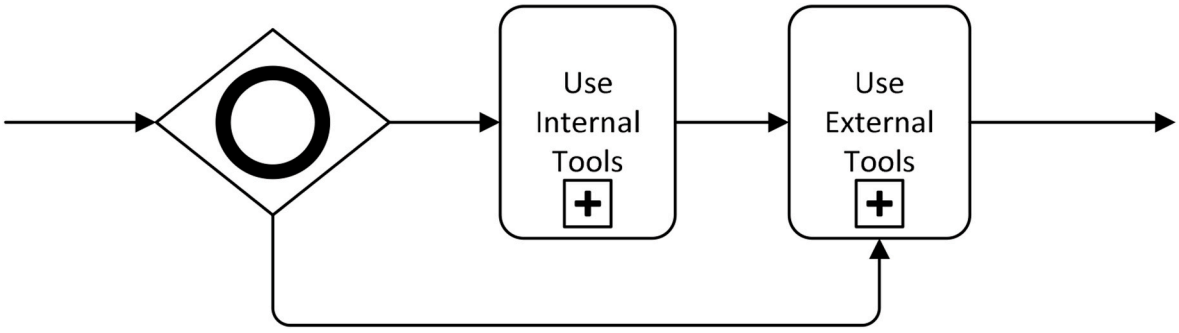


Fig. 10. A pattern in chemical search.

the SWM knowledgebase of high quality.

We have run 26 interviews from February 13th' 2017 until April 24th' 2017, 20 in Rijswijk, and 6 in München. Each interview lasted for 2 h.

We have modeled one BPMN/CMMN model from each interview. The detailed FBM models were fully developed after the interviews. We have produced 26 BPMN/CMMN models, which cover 5 search strategies and 2 future search flows, and 9 FBM models.

4.1. Strategy 1: figure-based search

This search strategy has similar search workflows. We have realized that a few patterns appear recursively after interviewed 4 examiners. They have also claimed that colleagues in the same technical field search in the same way.The most identical pattern in figure-based search strategy is illustrated in Fig. 8.

documents per search is 1000–5000.

Fig. 9 shows how the task of scanning through all the figures in the documents from a class can be executed using the EPOQUE2 prototype.

Note that this search strategy is to compensate the technical limit of our current figure search function, which is based on labels or annotations of figures. Note also that shape reorganization cannot tackle the challenge of figure-based search, seeing that a shape in one CPC class at the lowest level is not identical to the other.

[8] A CPC gérant at the EPO is a patent examiner nominated for managing the CPC scheme and definitions in his technical field. Normally, there is one gérant per field.

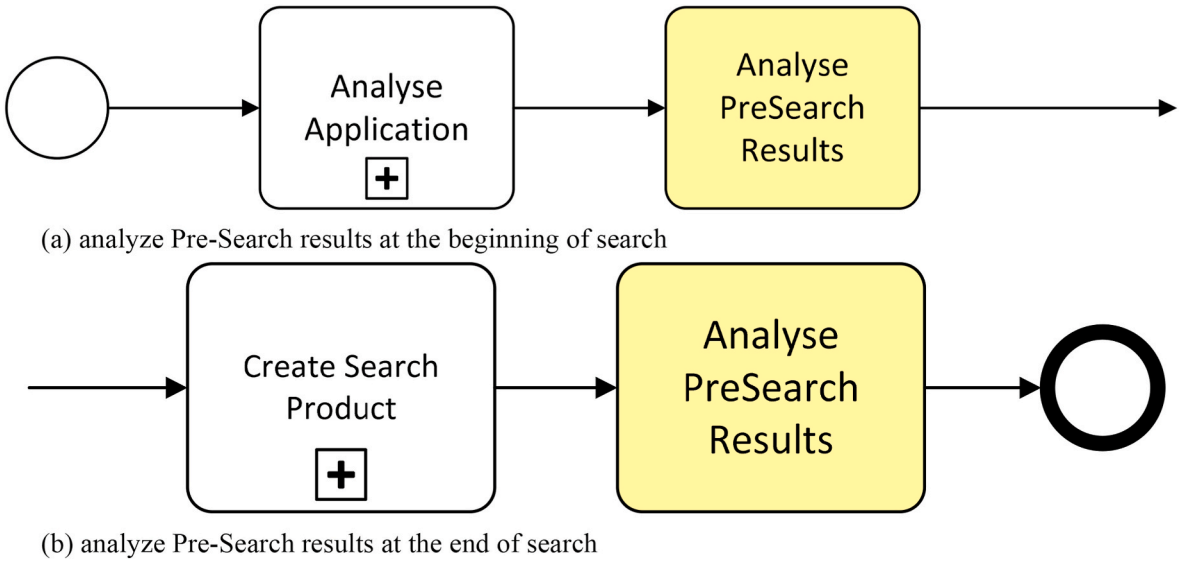


Fig. 11. Two ways of using Pre-Search results.

4.2. Strategy 2: chemical search

For the chemical search strategy, there are in total 7 search workflows, 5 patterns and 36 business requirements in these workflows.One pattern is shown in Fig. 10.

Compared to the examiners who apply the figure-based search strategy and who quickly scan through a large amount of documents, most examiners who apply the chemical search strategy usually analyze 150–500 documents for one search. Therefore, it is important for them to stay focused and know what exactly to search. This is an important reason of consulting internal tools before running external tools.When comparing detailed workflows of chemical search, we found two patterns as shown in Fig. 11.

Pattern (a) in Fig. 11 is also common to other search strategies. Pre-Search, as in its name Pre-Search, is a search that is launched automatically when a new search starts. It is an automatic search based on available information of a patent application, for example, the citation information from other national offices, citation information from the applicant and extracted keywords. It is recommended to analyze the Pre-Search result before any manual search in order to search more efficiently.

Pattern (b) in Fig. 11 is special. It does not appear in any other kinds of search strategies. Why examiners analyze the Pre-Search results after the manual search is finished is because of the following reasons.

A Pre-Search result contains 60–80 families. Compared to the situation where an examiner of figure-based search, who can scan through

Fig. 12. Citation-based search strategy supported by EPOQUE2.

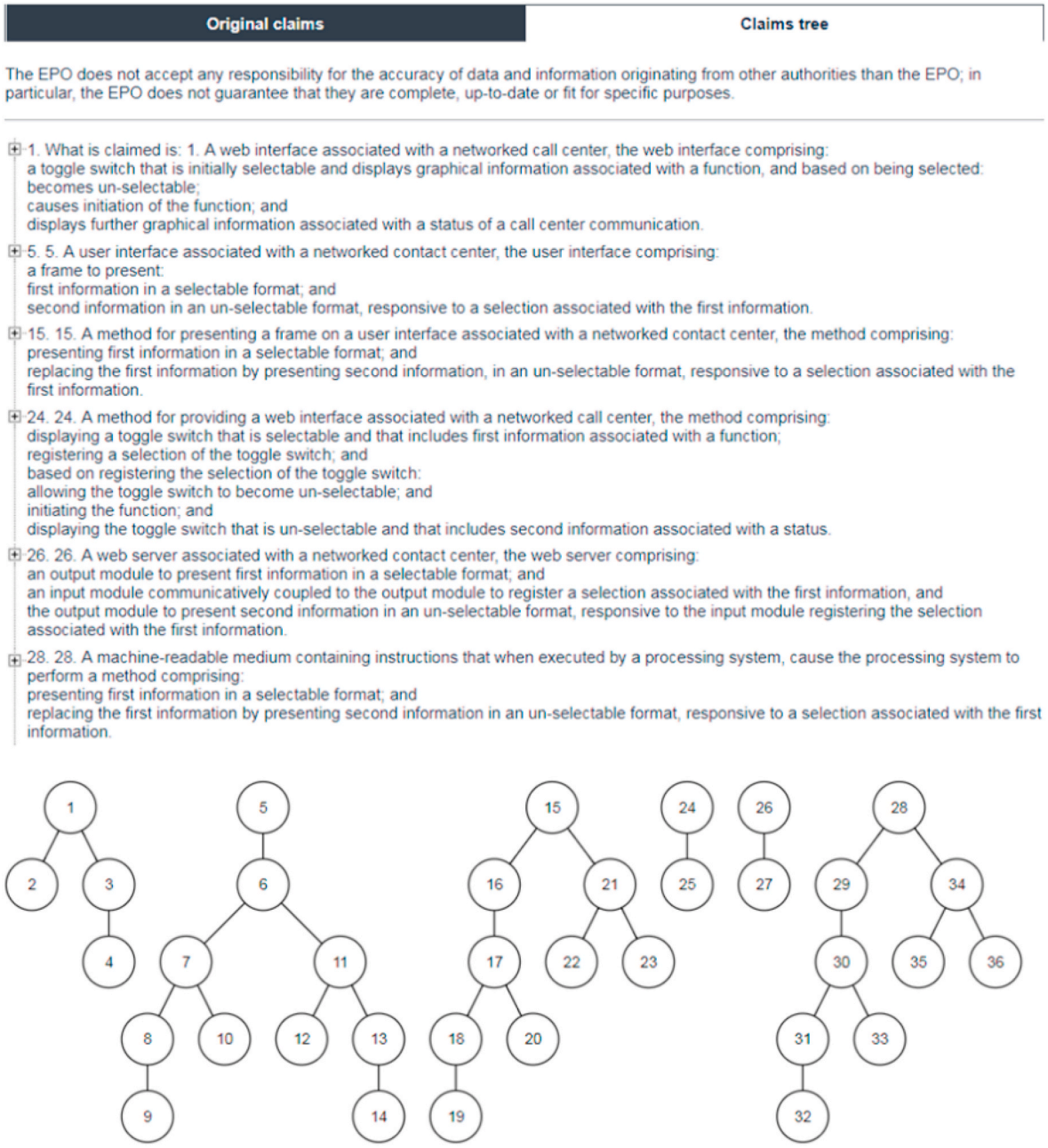


Fig. 13. A claim tree generated by Espacenet.

minimum 1000 documents, it is time-consuming for an examiner of chemical search, who usually analyze 150 documents, to study the Pre-Search result. Interesting enough, this idea is also reflected in a future search strategy (in Sec. 4.6) from other technical fields.

Accordingly, we have identified a few solutions: One solution is to improve ranking in Pre-Search results in these technical fields. Another solution is to allow end users to run Pre-Search at any moment during a search and allow them to refine the search results. Process information provided by Pre-Search shall be provided.

4.3. Strategy 3: non-patent literature search

"Prior-art" is not only made of patent literature but also consists of any other types of documents available in any forms to the public before the filing date of a patent application. We consider technical journals, conference proceedings, committee reports, meeting minutes and standards etc., also as "prior-art" in the fields related to Artificial Intelligence, Computer Implemented Inventions. It is also applicable in the fields, where many research/industrial topics are actively carried on, such as standards-related fields.

In particular, we had two interview meetings. One is about search in standards and the other is about search in books.

For search in standards, meeting minutes of standardization bodies are important because good prior-art can be often found in this set of documents. A library of search concept is also important, seeing that applicants often use terms deviated from the standards, to which they contribute. An internal tool called SeaStar [9] is used to support searching in standards.

For search in books, search is a means for an examiner to understand the real problem-solution of a patent application. In other words, search is done as soon as the problem-solution is identified. It is similar to the search behavior as discussed in Ref. [5]: search is not a linear process but

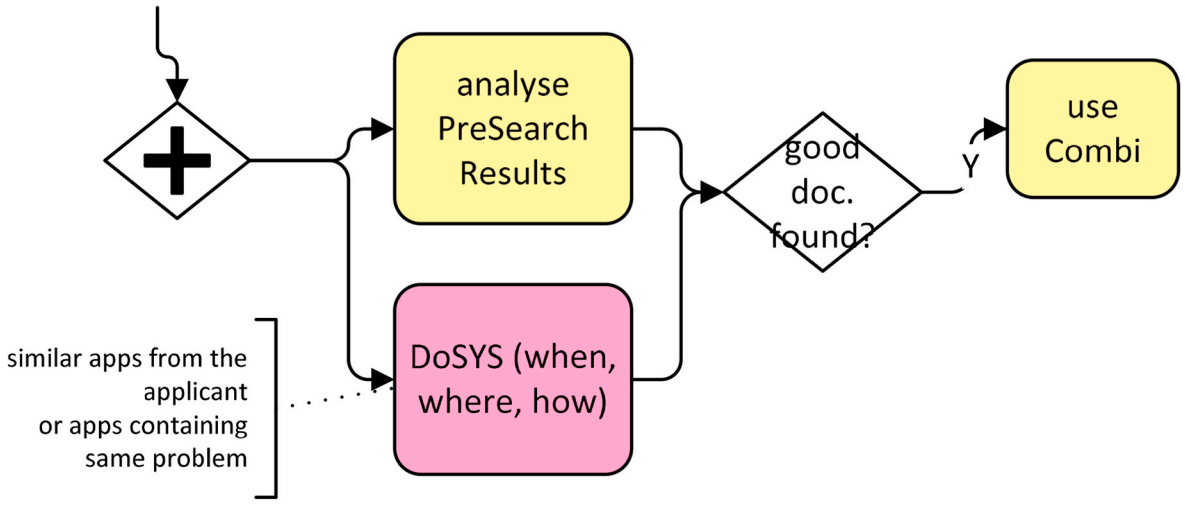


Fig. 14. Workflow of dealing with an application with known problem-solution.

a berry picking behavior in unknown terrain. Note that this search behavior could also be observed in the mixed search strategy in Sec. 4.5.

For this strategy, since main inventions in the technical field of the interviewee come from scientific studies. It is important to first understand the background of an invention. Books are the best resource to find background information and references in books are valuable for exploration.

Since references in books or any scientific papers are important to the examiners in these fields, it is requested to have a proper support of citation exploration.

A related search strategy is called citation-based search strategy. It was analyzed before the SWM project. Examiners, who use a citation-based search strategy, explore the prior-art literature by following references. They stop searching when all interesting references have been scanned through.

Fig. 12 shows how this search strategy can be supported by EPOQUE 2.0. It is a feature called 'breadcrumbs'. Citation information comes from various resources. One is an EPOQUE database called DOSYS (DOssier SYStem). It contains citations from search reports. An example of EPO search report can be found in Refs. [18].

Another important resource is called COMBI (COMBInation of citing and cited documents) – an EPOQUE preparation. Examiners use COMBI to explore the cited and citing PN and NPL documents. Part of the citation information of COMBI comes also from DoSYS.

Use of COMBI is a common task in all the existing search strategies. A typical pattern is to use COMBI as soon as a good prior-art document is found. It is useful for 1) discovering classes especially for the fields where CPC classes need to be improved, 2) final checking before writing the search report to ensure that important documents have not been missed, 3) exploring research background of applicants, and so forth.

4.4. Strategy 4: claim tree and search table driven search

With this search strategy, examiners can get an automatically generated search report after all search tasks are executed.

Typically, it starts with the claim tree of a patent application. A claim tree is either manually designed or automatically generated using existing tools. Fig. 13 shows a claim tree generated by EPO Espacenet.

During a search, examiners annotate found prior-art documents with claims. When the search is finished, a search report is generated based on the annotations.

Annotations include found CPC classes, examiners' description of real problem-solution, comments that will be used in the examination phase,[9] detailed search steps, manual drawings (such as chemical formula and process steps), key search queries and summary of search report.

This search strategy is used not only for automatically generating search report, but also for providing a good hint of what are the best prior-art documents to cite. One way of choosing the best prior-art documents is to calculate scores based on types of finding. In a content-wise dimension, some prior-art documents are annotated as 'novelty-destroy'. The other is annotated as 'inventive step'. In a time-legal dimension, the documents can be annotated with 'P', 'E' and 'In Time'.

4.5. Strategy 5: mixed search (figure, class and text)

Mixed search is the most common search strategy in the fields where CPC classes are not very well defined.

There is no obvious pattern in the search workflows. However, it is obvious that these interviewees use almost all available search tools. Examiners decide to apply a search strategy based on the content of application.

For example, depending on whether a problem-solution is known to the examiner or not, an examiner executes different search flows.

Fig. 14 shows part of the flow when the problem-solution of a patent application is known to the examiner. The initial prior-art set has two parts: one from Pre-Search and the other from DoSYS. When interesting documents are found in this initial set, he uses COMBI for a further exploration. After scanning through the prior-art documents, he can often find good documents and stop searching. This search strategy is a mixture of automatic search and citation-based search.

4.6. Future search strategies

One future search workflow has been suggested by an examiner is collaborative search. In the technical fields of the interviewee, colleagues are working not at all in the same way. What matters is that the invention is well understood at the end of search. Which prior-art documents will be cited in the search report is of less importance because there are many citation candidates once the problem-solution is identified.

[9] see the guidelines for Examination, G-VII, 5 (https://www.epo.org/law-practice/legal-texts/html/guidelines/e/g_vii_5.htm).

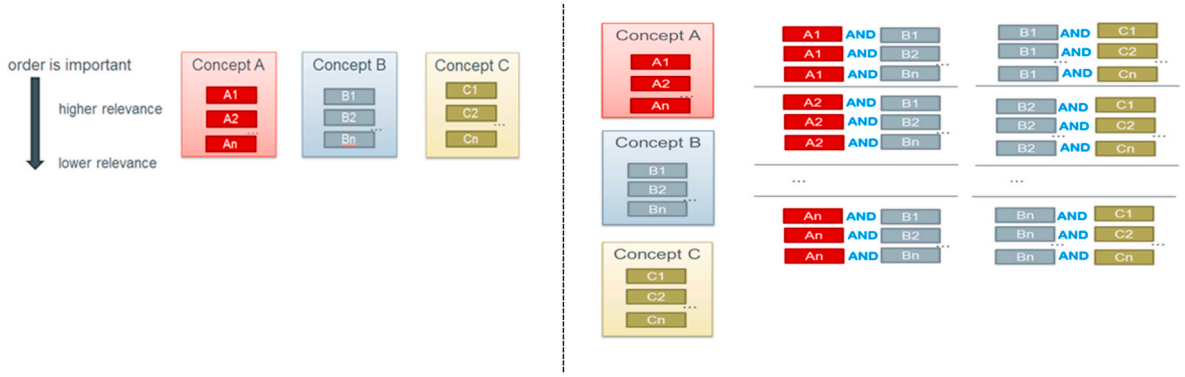


Fig. 15. Left: Search concepts in exhaustive search; right: Combination of search queries.

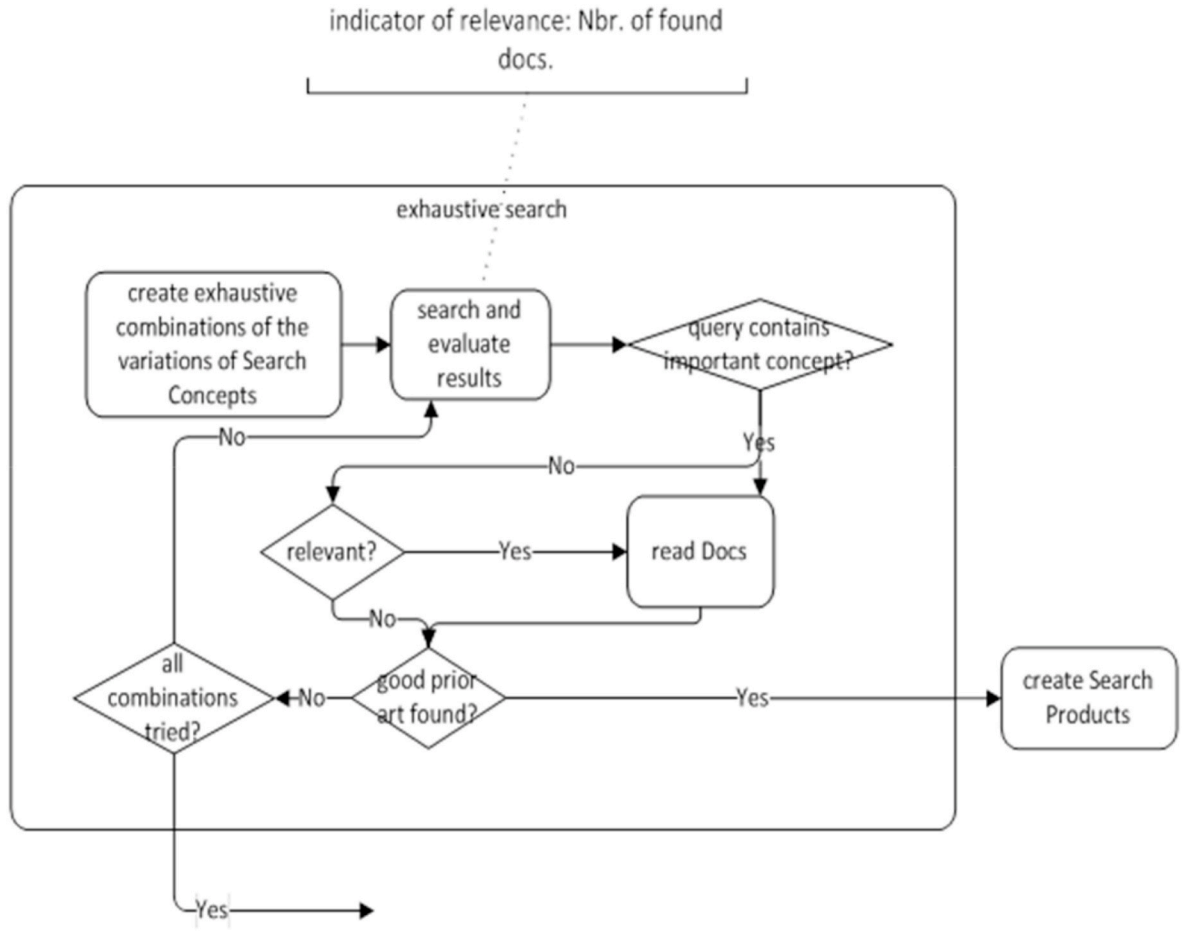


Fig. 16. Exhaustive search flow (partial view).

Therefore, allowing more than one examiners to do a prior-art search for a same patent application is meaningful. For supporting this search strategy, it requests a support of tools. An IP5[10] pilot project[11] has been launched on 1st July 2018 to support collaborative search and examination under the PCT.

Other future search strategies have been discussed based on the gaps between legacy EPOQUE and our new search tools, such as ANSERA/EPOQUE 2.0. One search strategy is exhaustive search in EPOQUE. It allows searching various independent claims on one go.

In this search strategy, each patent application contains several search concepts, each of which is expressed as a set of queries with different relevance levels as illustrated in Fig. 15 (left).

For example, concept A is expressed as queries$A_1, A_2 \ldots A_n$, where the most relevant query for concept A is A_1, A_2 is less relevant than A_1 and so forth.

Afterwards, an exhaustive combination of search concepts is created as illustrated in Fig. 15 (right). Each combination is one query made of one query from a concept. For example, the first combination is $A_1 \cap B_1 \cap C_1$. It is also the first search query (or combination of the 3

[10] Five Intellectual Property offices (https://www.fiveipoffices.org/about).
[11] https://www.epo.org/law-practice/legal-texts/official-journal/information-epo/archive/20190702.html.

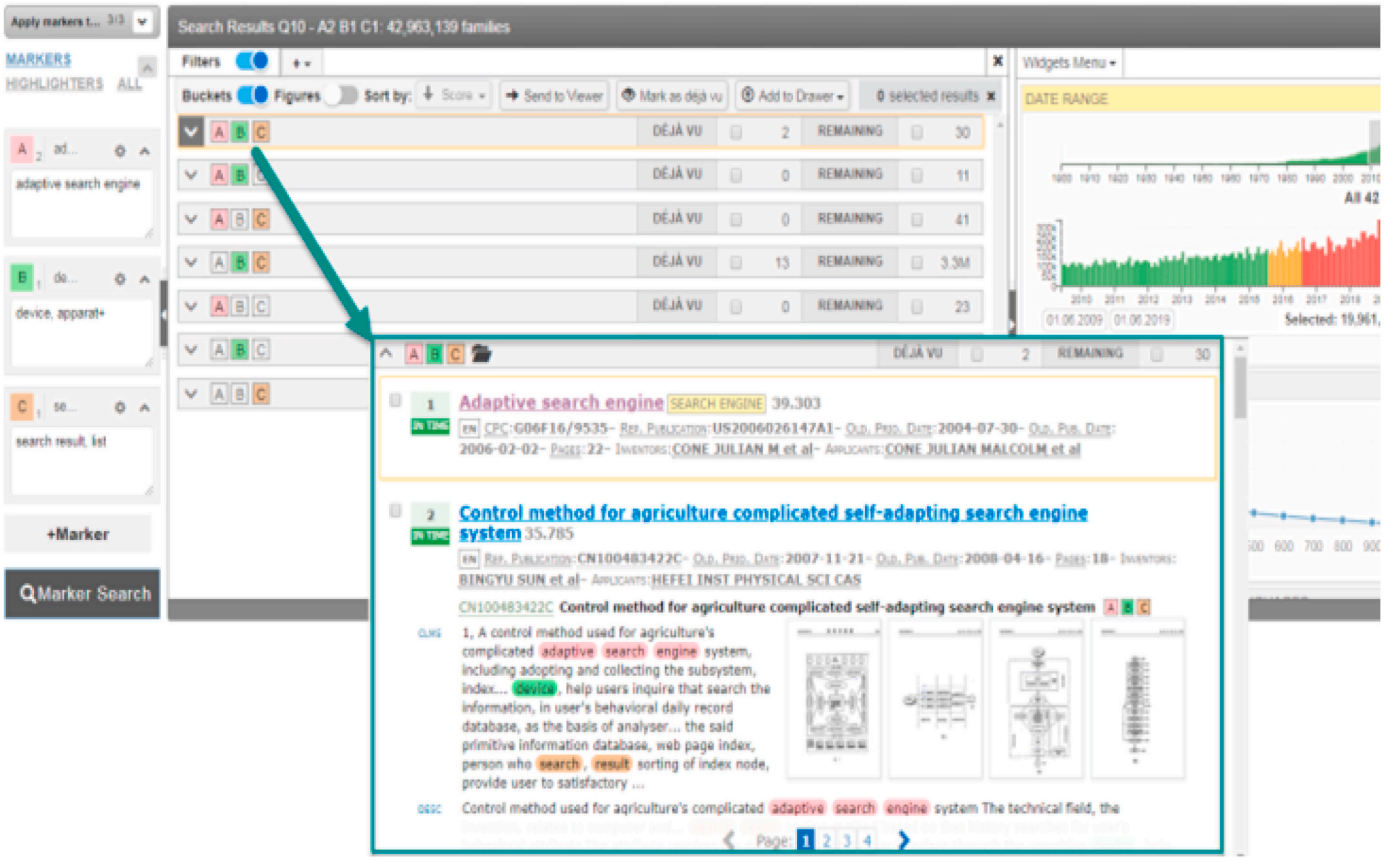


Fig. 17. Bucket view in EPOQUE 2.0.

queries).

The main workflow is illustrated in Fig. 16. Examiners stop searching when a good prior-art document is found after executing a few combinations. If all the combinations have been tried and nothing is found, then the search is complete.

Search products are the documents produced after a search is done. For example, a search product can be a Search Report or Written Opinion.

ANSERA/EPOQUE 2.0 uses markers and buckets. A marker is a search query with a dedicated color. A bucket is a way to group search results as illustrated in Fig. 17. For example, the first bucket indicates that all the three markers (i.e. **A**, **B**, and **C**) have search hits. The last bucket indicates that only marker **C** has search hits. There are a set of widgets on the right pane in Fig. 17. They are used to filter the search results or provide examiners with insights for modifying search queries.

Given that each concept combination is one search session, the exhaustive search workflow requires only one bucket - . All the rest are useless. If all combinations are presented in one search session, then it is difficult to have a clean overview of the search results.

To tackle this challenge, we have proposed the following solutions: 1) a better function of grouping/filtering or a sorting feature to obtain a cleaner view of search results; 2) allowing users to rank the buckets by allowing them to give weights to the markers; 3) allowing users to make hierarchical markers or marker groups. Fig. 18 shows the FBM model

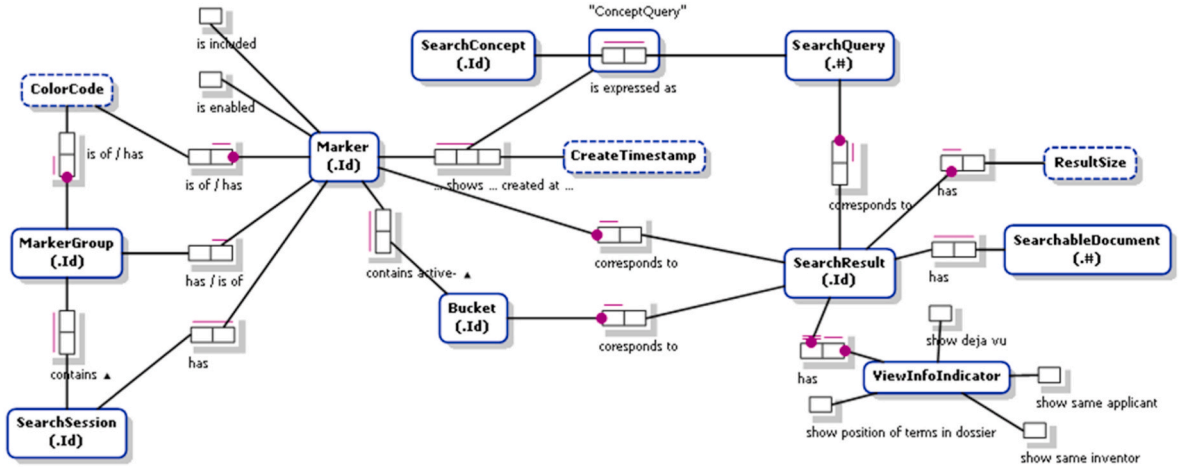


Fig. 18. Advanced bucket in FBM (partial view).

that supports solution 3).

Other gaps between the legacy systems and new search tools, and new search features were known before the SWM project. They are NPL search, Query Builder (including automatic translation of search terms), Concept Management (including ontology-based search), search relevant data from DOSYS, Annotation (including manual and automatic annotations), Search History and Search Session.

During the SWM project, these features have been validated and integrated in the search workflows of interviewees. The links between the existing workflows and future search features give rich contextual information for our ongoing and future projects.

5. Discussions

As presented in the paper, there is a large variety of workflows applied at the EPO. Since the coverage of search strategies is not guaranteed by the available set, other resources, such as models (and requirements in any formats) should be further gathered. After the SWM project, we have been continuously using the SWM methodology to gather requirements in projects.

A good search tool for one technical field may not meet the needs from the other. In a same technical field, every examiner has some unique behaviors in search tasks.

The SWM project showed that examiners will always continue a search even if automatic search has already given good results. As already discussed in the paper, Pre-Search algorithms work well for some technical fields; in chemical search, Pre-Search does not provide good results. For the fields that Pre-Search works well, examiners (as well) do not stop searching after the Pre-Search results are analyzed. The main reason is that examiners need to understand why the result is good in order to draft a meaningful search report. In addition, examiners tend to have the desire of finding an even better result to enhance the arguments in the search report. It is important for the reason of quality, which EPO is proud of. This issue potentially reveals a 'black box' effect of automatic search. In some ongoing projects, we have been already studying possibilities of rendering necessary and meaningful information of such a black box.

There are 45% of examiners who strongly believe that search is fundamentally a human effort and 52% who do not believe in a fully automated search in 2035 with the reasons given above. However, 70% believe that Artificial Intelligence and Machine Learning could provide a valuable support for search. The answer to the question on whether AI can replace the work of examiners or not is out of the scope of this paper.

One requirement, which has been mentioned several times by different interviewees, is about single search tool. Examiners would like to have a single search interface for all internal search tools, which should be further integrated with other EPO tools in the whole patent granting procedure.

6. Conclusion and future work

The paper is a review of the SWM project, which provided a structured highlight on the way how examiners work. We use the results to optimize our existing tools. In this paper, we have illustrated 5 search strategies, with which we have recorded the best search practices in various technical fields and worked on new requirements.

The results are beneficial not only for the future searching tools, but also for other aspects. For example, 80% of examiners who have answered the questionnaires claimed that classification is very important for efficient searches. 84.7% of them claimed that CPC is important for efficient searches. Cause and Effect models delivered in the SWM project have also revealed how good CPC classification can lead to efficient searches. An ongoing activity of automatic classification tries to meet this need.

There are in total 203 requirements elicited during the interview meetings. These requirements have been analyzed and taken to the board of our ongoing project – EPOQUE 2.0. For example, the feature called 'breadcrumbs' (as illustrated in Fig. 13), together with COMBI and DOSYS, is a current solution to support citation based search strategy. Other ongoing projects, for instance, a new graph database of citation based on existing legacy databases will be delivered to fulfil a number of non-functional requirements with regard to this search strategy, such as performance.

Some requirements concerning the search strategies in this paper have been already implemented while the authors are drafting this paper. For example, the feature of 'search in NPL' in the EPOQUE 2.0, which answers the needs of NPL search strategy, has been implemented in release versions 1.2 in mid-2019 and 1.3 in December 2019.

In the future, we will analyze other search strategies, e.g. search in Sequences. We need to continuously use the SWM methodology for understanding new search strategies. We will also use the methodology for gathering requirements.

Acknowledgements

We would like to thank our colleagues – Vincent Vuillamy and Rex Bosma – for their valuable support in the SWM project as well as our past and present colleagues who have contributed to the work described in the paper. We gratefully acknowledge the support of EPOQUE and ANSERA/EPOQUE 2.0 teams.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.wpi.2020.101989.

References

[1] J. Allan, J. Arguello, L. Azzopardi, P. Bailey, T. Baldwin, K. Balog, et al., in: J. Culpepper, F. Diaz, M.D. Smucker (Eds.), Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018), vol. 52, SIGIR Forum, 2018, pp. 34–90. June 2018.
[2] J. Cone, G.L. Franklin, G.J. Ryan, W.F. Stalker, Patent No. US2006026147A1, United States of America, 2006.
[3] D.M. Kosak, A.K. Lang, Patent No. US2002120609A1, United States of America, 2002.
[4] Z. Stekkelpak, Using Pre-search Triggers, United States of America Patent US8510285B1, 2013.
[5] M. Bates, The design of browsing and berrypicking techniques for the online search interface, Online Review (13) (1989) 407–424.
[6] R. White, R. Roth, Exploratory search: beyond the query-response paradigm, Morgan & Claypool, 2013.
[7] B. Jürgens, V. Herrero-Solana, Espacenet, patentscope and depatisnet: a comparison approach, World Patent Inf. 42 (2015) 4–12.
[8] E. Marttin, A.-C. Derrien, How to apply examiner search strategies in Espacenet. A case study, World Patent Inf. 54 (2018) 33–43.
[9] R. Oltra-Garcia, Efficient searching with situation specific and adaptive search strategiess: training material for patent searchers, World Patent Inf. 54 (Supplement) (2018) 29–32.
[10] D. Andlauer, in: S. Adams, T.L. Bereuter, N.S. Clarke (Eds.), Automatic Pre-search: an Overview, World Patent Information, vol. 54, 2018, pp. 59–65.
[11] C. Jonckheere, EPOQUE (EPO QUEry service) the inhouse host computer of the European patent office, World Patent Inf. 12 (3) (1990) 155–157.
[12] FBM Working Group, Fact-based Modelling Metamodel: Exchanging Fact-Based Conceptual Data Models, European Space Agency, Noordwijk, 2015. http://www.factbasedmodeling.org/Data/Sites/1/media/FBM1002WD08.pdf.
[13] T. Halpin, Object Role modeling (ORM/NIAM), ch. 4. Handbook on Architectures of Information Systems, Springer, Heidelberg, 1998.
[14] T. Halpin, Information Modeling and Relational Database, Academic Press, 2001.
[15] P. Spyns, Y. Tang, R. Meersman, An ontology engineering methodology for DOGMA, Appl. Ontol. 1–2 (3) (2008) 13–39.
[16] M. Uschold, M. Gruninger, Ontologies: principles, methods and applications, Knowl. Eng. Rev. 11 (2) (1996) 93–155.
[17] F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, P.F. Patel-Schneider, The Description Logic Handbook: Theory, Implementation and Applications, Cambridge University Press, Cambridge, 2010.
[18] K. Loveniers, How to interpret EPO search reports, World Patent Inf. 54 (2018) 23–28.