PREBEN HANSEN

Task-based Information Seeking and Retrieval in the Patent Domain

Processes and Relationships

ACADEMIC DISSERTATION
To be presented, with the permission of
the board of the School of Information Sciences of the University of Tampere,
for public discussion in the Auditorium Pinni B 1096,
Kanslerinrinne 1,
Tampere, on September 3rd, 2011, at 12 o'clock.

UNIVERSITY OF TAMPERE

UNIVERSITY
OF TAMPERE

ACADEMIC DISSERTATION
University of Tampere
School of Information Sciences
Finland
Swedish Institute of Computer Science (SICS)
Sweden

Distribution
Bookshop TAJU
P.O. Box 617
33014 University of Tampere
Finland

Tel. +358 40 190 9800
Fax +358 3 3551 7685
taju@uta.fi
www.uta.fi/taju
http://granum.uta.fi

Cover design by
Mikko Reinikka

Acta Universitatis Tamperensis 1631
ISBN 978-951-44-8496-4 (print)
ISSN-L 1455-1616
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 1093
ISBN 978-951-44-8497-1 (pdf)
ISSN 1456-954X
http://acta.uta.fi

ISRN SICS-D--55--SE
SICS Dissertation Series 55
ISSN 1101-1335

Tampereen Yliopistopaino Oy – Juvenes Print
Tampere 2011

ACKNOWLEDGEMENTS

In memory, and *for ever* present… Ulf Pettersson

"…In light and in openness, new views and aspects will be revealed…"
(Ulf Pettersson, at one of our many discussions)
How could I *ever* have done this without you, Ulf?

The situation of being divided between SICS and Tampere University during the writing of the thesis has provided me with valuable and stimulating experiences. In these two environments, many ideas and people have been in the crossroad of my research. Some just for a short time, while others hopefully for many years to come.

First and foremost, I would like to thank my supervisor, Professor Kalervo Järvelin. I am deeply grateful to him for his extraordinary patience with my specific research situation and for initially supporting my ideas. During the writing process of my thesis, things not anticipated have been unfolded. Our face-to-face discussions at the School of Information Studies in Tampere have always been inspiring, intellectual and very enjoyable. I would also like to thank Professor Pertti Vakkari, Associate Professor Sanna Talja, Professor Reijo Savolainen and Professor Jaana Kekäläinen, at the School of Information Studies at the University of Tampere, for encouraging feedback and suggestions for methodological readings. Thanks also to Leena Lahti and Mirja Björk who kindly helped me during my visits to Tampere University.

I also want to thank my dear college, Docent Jussi Karlgren at SICS, acting as my mentor during the whole work process with this thesis. Thanks for all the intellectual discussions and feedback on ideas and statistical issue and always open for new turns of a problem to solve.

I want to thank everybody that contributed with their knowledge and time for discussions related to this work, in particular Dr. Katriina Byström (SSLIS at the University of Borås), for showing great interest in my work and for important discussions. Professor Peter Ingwersen (Royal School of Information and Library Science, Denmark) for initial creative discussion on IR matters in the early stage of the thesis. I would also like to thank Professor Björn Gambäck, Gunnar Eriksson, Olof Görnerup at SICS for their support and knowledge. I also would like to thank, Christer Norström, SICS Managing Director, my former lab leader Magnus Boman and current lab leader Björn Levin for giving me the possibilities to pursue this work.

Finally, and most importantly, I would like to give my most sincere thanks to my wife Anette, and my children Linn and Andrea, for being close to me during this (long) period. They have, most bravely, and with a mix of curiosity and impatience, been supporting me in different ways.

Stockholm, August 2011
Preben Hansen

ABSTRACT

Information-intensive work tasks in professional settings usually involve dynamic and increasingly complex information handling tasks that include the gathering, assessment, assimilation, and creation of information. Understanding the factors affecting information handling processes, and their interaction, is important and forms the objective of this thesis. To reach this objective, the present thesis examines one information-intensive domain, the patent information domain.

The thesis addresses this objective through a longitudinal empirical study in a real-world patent information handling context, that of the Swedish Patent and Registration Office. Specifically, three main theoretical aspects of information access are investigated: information seeking and information retrieval tasks as performed within patent work tasks. These aspects of information access are observed via multiple data collection methods. Qualitative and quantitative data are collected for analysis. Although these three aspects of information access have been investigated in various ways, contemporary understanding of their inter-relationships in real-world situations is far from sufficient.

Based on the empirical observations, a framework for patent information seeking and retrieval is proposed. This includes identifying novel features of the search process, such as relevance judgement strategies, and of information needs within patent information retrieval. A set of important relationships between the task levels of information seeking and retrieval and work tasks are empirically described. During the study, extensive collaborative information retrieval activities were revealed and integrated into the general framework for patent retrieval. Features and conditions of collaborative information activities are outlined and discussed.

Finally, the thesis proposes a methodology for systematically studying empirical information seeking and retrieval processes as applied over a longer span of time in a real-world professional work setting. We developed a method for analysis, description, and systematic categorisation of patent IR sessions and modelling of session-based information retrieval. In addition, and schematic diagrams illustrate its application.

CONTENTS

BACKGROUND

SETTING

RESULTS

LIST OF FIGURES

LIST OF TABLES

List of Appendices

1
INTRODUCTION

In information-intensive work tasks, it is crucial for professional workers to stay informed and, at the same time, inform their colleagues in order to manage knowledge effectively and stay competitive, effective, and innovative.

Information-intensive work tasks in professional settings usually involve dynamic and increasingly complex means of information handling that include gathering, assessment, assimilation, and creation of information. Therefore, we need to enhance our understanding of factors affecting information handling processes, and how different components interact and relate to each other.

Information Access (IA) encompasses a wide range of processes, of which Information Seeking (IS) and Information Retrieval (IR) involve two different and sometimes opposite viewpoints and research areas, both important processes that will be of focus in the study described here.

Information seeking is commonly understood as the process performed by a human involved in searching for information through different information channels, such as paper-based, human, and those involving electronic IR systems. Information seeking involves the perception of, for example, the task problem, information needs, and relevance assessments. The research approach of information seeking is focused on empirical studies and on theoretical models and conceptual frameworks, to describe and explore the known elements and their presumable relationships. Aspects that have been given focus within this research field include information seeking strategies (e.g., Bates, 1989; Belkin et al., 1993, 1995) and user behaviour (e.g., Borgman, 1989; Kuhlthau, 1993a, 1993b; Wilson, 1997). Vakkari (2001a) proposes a model based on identified iterative information seeking and retrieval processes as well as various means of analysing these processes.

Approaches in IR research have been investigating research on IR techniques for storage, representation, searching, and presentation of information potentially perceived as useful and relevant for a human user or a group of users (Ingwersen & Järvelin, 2005). One such approach looks at lab-based IR. This line of IR research has

its foundations in the Cranfield project (Cleverdon, 1966) and has since then contributed with a vast body of research results, data, and knowledge with three emerging approaches: system-oriented IR, user-oriented IR, and cognitively oriented IR approaches. The aim of the system-oriented research is to develop and construct new algorithms for retrieval and presentation of (topically) relevant documents. One of the most important models used in system-oriented IR research is lab-based IR. The basic laboratory model does not involve the user, instead, it focuses on documents, requests and their representations, the database, queries, and the matching of the representations of the documents and the requests (Ingwersen & Järvelin, 2005, pp. 114–115). Since the system-oriented IR research approach neglects involvement of the user, research that is focused, for example, on the users and the information need can be found in user-oriented research into IR (e.g., Bates, 1989, 1990) and in cognitive IR research (e.g., Ingwersen, 1992) dealing with interactive communication processes that emerge in the transfer of information (e.g., Bates, 1989, 1990).

Usually, the operational IR systems described and evaluated were based on Boolean logic. The positive element in using Boolean systems is the possibility of creating structured and precise queries, while the downside is that people not skilled in Boolean logic have difficulties using such systems. The emerging Web technologies have now changed the scene for operational online IR in that it now may include a varied number of domains, a much larger set of online documents and document types, larger and varied user populations, and varied information access systems in which the IR component is just one part of a larger information management system (Ingwersen & Järvelin, 2005).

As stated by, among others, Belkin et al. (1995), Ingwersen (1992, 1996), and Saracevic (1996), the traditional lab-based IR approach alone cannot provide understanding and knowledge of the interaction between the user and the IR system as well as understanding of the human actor interacting with information sources. It has been claimed (e.g., Hansen & Järvelin, 2000; Ingwersen & Järvelin, 2005) that for understanding of information search and retrieval (IS&R) processes, the information seeking and information retrieval phenomena cannot solely be examined in isolation. For example, query formulation is often viewed as an *individual activity* but should be seen as related to the overall task at hand. Furthermore, the searcher performing the task is viewed as being rather *isolated*; however, it is obvious that the searcher is performing the task in a certain situation. Information retrieval and information seeking need to be viewed and understood as two processes that are integrated and closely related.

The patent domain provides us with a rich, complex, and information-intensive real-world platform on which a large number of information seeking and retrieval activities are performed and also information search in operational IR systems is performed and problem-solving are done daily and hourly. Such a platform is suitable for investigating in depth real-world search processes and studying work tasks, search tasks, and their relationships.

1.1 The work task and the IS&R tasks

In a professional work setting, a work duty can be described as a set of tasks that need to be performed. Most of these tasks can be considered to be work tasks. Work tasks can be further divided into subtasks that may be performed in order to accomplish the specific task(s) set by the organisation, group or team, and individual.

Work tasks may involve different tasks, such as search tasks. A search task can further include information seeking and also information retrieval tasks. The information retrieval task is explicitly considered a specific type of information seeking task (Wilson, 1999; Ingwersen & Järvelin, 2005).

Examination of the work task, and the levels of search tasks (information seeking and information retrieval task) forms the foundation of our empirical study of patent information handling activities and is applied to a real-life work task setting. In utilising these levels, often considered only separately, we may also develop an approach to their integration. By integrating these task levels and viewing them as intertwined, we believe, the present study will contribute to broader understanding. In general, IR systems have not been explicitly designed for specific work tasks, unless the system is designed in a highly specific domain with a well-defined knowledge structure and user environment.

Even though the number of analytical and empirical studies involving information seeking, human information-related behaviour, search strategies, and information channels and resources is slowly growing, the *relationships* between work tasks and information retrieval have not received enough attention (Hansen, 1999; Hansen & Järvelin, 2000; Vakkari, 2003; Ingwersen & Järvelin; 2005). One problem is that in studies of information retrieval, the user (or performer) is seldom present. At the same time, the information seeking research field has shifted focus from studying only the user and the user's behaviour in isolation, toward more contextual studies involving, for example, work tasks, interactive searching, and human–computer interaction technologies.

Motivation for studying the patent domain:
There are several reasons we wanted to investigate professional work tasks:

- Work tasks are seldom used in laboratory IR research as a context for the set of queries used in the IR experiments. Therefore, the outcome of laboratory-based experiments may say more about an algorithm than about the applicability of the results of the experiment in a real-world setting.
- Interactive IR experiments (e.g., Borlund, 2000) have tried to simulate a real search task in which test subjects will assume a certain situation, performing a set of queries with a description of the situation and some contextual components. The simulation may be conducted in a more or less complex setting, and most of the components are controlled and, through predetermination, also measured in a controlled way.
- Most of the a) laboratory-based and b) simulated experiments that use participants utilise students from academic settings do not have the in-depth competence and skills in performing real professional tasks to draw upon.

Both lab-based and simulated user experiments have their strengths, such as the controllability and designed measurements performed. However, exploring a specific domain also involves skilled professional workers performing real-world work tasks.

Accordingly, our motivation for studying work tasks related to information seeking and information retrieval tasks is as follows: first, we wanted to move the study of interactive information seeking and retrieval from laboratory-based settings into an environment where interactive IS&R activities actually are performed. We believe that, by doing this, we will reveal circumstances that may affect future conceptual and methodological frameworks for research. This would then allow us to study work tasks as well as interactive IS&R tasks in their natural environment and not separated from each other. The study of relationships between work tasks and the information seeking and retrieval tasks may reveal new knowledge and would then benefit 'an integrated view of information seeking and retrieval' (Ingwersen & Järvelin, 2005, p. VII).

1.2 The patent work domain

This section gives a brief presentation of the patent domain and workplace that is the target for our study. A more detailed description can be found in Chapter 5.

The study was conducted at the patent department of the Swedish Patent and Registration Office (SPRO)[1], a government agency. The overall goal of SPRO is to protect investments (ideas, inventions, designs, and trademarks) that individuals and companies have made into new technological innovations and to stimulate competitiveness in Sweden. The main work to accomplish that is done by handling incoming patent applications, which, in turn, involves tasks such as classification of patents, search, retrieval, and inspection and judging of relevant patent-related information.

The patent application:
The patent engineer (PE) basically handles patent applications written by professional patent bureaux, applications by companies' internal patent departments, and finally those patent applications written by private persons. There are both national patent applications, as well as international applications, which affects the handling process.

The patent application (PA) itself is a highly structured document consisting of several mandatory elements, such as abstract, background, description, claims, figures, and summary. The abstract is important because it contains a condensed and detailed summary of the invention, while the description gives a statement of the state of the art regarding the technology.

Finally, one of the most important parts of the document is the claim section, since it defines the various features of the invention for which the applicant wants to claim legal protection. The language in the patent application may have different levels of formalisation: e.g., the description has a more narrative form, while the claims section is more formal, for legal reasons. In almost all patent applications, one may find one

[1] SPRO, Stockholm, http://www.prv.se/.

or more images or figures to illustrate the technical details of the invention. Examples of images are chemical structure, circuit diagrams, and flowcharts.

Search types:
Within the patent domain, there are different types of searches. The goal with one may be to test whether it would be worth the effort to write an application, while other types of searches are concerned with the technological field and yet others are performed in different phases of the patent handling process (such as the 'prior art' search). A novelty search, on the other hand, is performed to identify the novelty or lack thereof as regards the proposed solution claimed in the patent application.

The aspect of relevance:
When judging a document for relevance, the patent engineer uses a very specific set of graded relevance criteria. These can be combined in different ways, expressing the level of relevance.

The patent handling process:
In general, the patent handling process is well structured and involves a certain sequence of stages. When a patent application arrives at SPRO, it is registered. The application is then reviewed and classified. After that procedure, the application is assigned to a patent engineer with the necessary expert knowledge. This is generally followed by description of the need identified and specific conditions for further processing. The search task involves various interactions with different sources, and the search outcome then undergoes relevance assessment and judgement. From the documents retrieved, information may be extracted and summaries may be written as reports that will be sent to the applicant. Finally, the PA may involve a series of exchanges between the patent office and the applicant before public announcement.

Motivation for studying the patent domain:
The patent handling process is a very information-intensive and focused work task. What makes it challenging for our purpose is that the patent work involves a) professional real-world work and search tasks, b) extensive and concrete IS&R processes, c) highly complicated problem-solving procedures, d) time-consuming search tasks (most of the time each work day involves search-related duties), e) a relatively unknown domain within the IS&R research field (at the time when the study was performed), and f) the patent application (the document itself) as a complex and challenging entity (with different source types; documents; and content types, such as text, drawings, and figures; and languages). Finally, the patent work also results in an outcome in the form of a report (in contrast to traditional searching that yields a list of hits). This means that the outcome (the report) is a consequence of the assessments of the search result. In the present thesis, we will be investigating some of these features.

Thus, the patent domain represents a platform from which several important and complex problems may be studied. This motivated us to pursue our goal of

performing on-site studies of real-world work and search tasks within this specific domain.

1.3 The goal of the study, and its research problem and methods

The goal of the study described here is empirical investigation of IS&R processes of real-world work tasks within the patent domain. We will analyse characteristic features of patent IR and, in addition, whether and how these features affect the information seeking and information retrieval stages. We therefore need to explore the characteristics of different task levels. If we consider IS&R processes important aspects of professional work tasks and, furthermore, deem necessary study of these processes in real-world (in our case, the patent domain) situations, then it is important to

a) *Describe* the overall patent handling process;
b) *Describe* the IS&R processes (sessions) within patent handling;
c) *Analyse* characteristic features of patent information retrieval (PIR) with regard to various aspects of IR (e.g., information need, source selection and usage, query formulation, relevance assessments, search task outcome, and search process structure);
d) *Analyse* both individually and co-operatively performed elements of PIR; and
e) *Develop* a methodology for analysing data of task-based PIR studies that is based on multiple data collection methods and then illustrate its application.

For the present study, the Swedish Patent and Registration Office was chosen as the setting. The patent domain provided us with a rich, complex, and information-intensive and challenging environment, in which both information seeking and information retrieval activities are performed. The units of analysis are at two levels: first a) at the overall patent information handling process level and secondly b) at the patent search session level, within the process.

The study was designed to cover two main problems, addressed below.

Problem 1 – the empirical issue:
The first problem is empirical and deals with describing the overall patent handling process and, more specifically, the IS&R session activities. We will investigate the relationships between work tasks and the IS&R task performance process. This involves analysing the processes of the various tasks as well as collaborative information handling in the patent domain.

The main research question is: What are the effects of work task features on the information seeking and retrieval process in the patent domain?

This main research question has seven separate sub-problems:

1. What are the effects of the work task features (work task (WT), information seeking task (IST), and information retrieval task (IRT)) on work tasks?
2. What are the effects of work task features (WT, IST, and IRT) on the deconstruction and formulation of the information need?

3. What are the effects of work task features (WT, IST, and IRT) on the types of sources and source content utilised?
4. What are the effects of work task features on query formulation?
5. What are the effects of work task features (WT, IST, and IRT) on relevance judgement (RJ) performance?
6. What are the effects of work task features (WT, IST, and IRT) on use of information for completion of the task?
7. How are collaborative information retrieval activities manifested within and in the course of the IS&R task performance process?

These sub-questions are further detailed in Chapter 3.

Problem 2 – methodology:
The second problem is related to development of a methodology for analysing the data of task-based PIR studies that is based on multiple data collection methods. Since we intend to investigate real work tasks and their characteristic features as well as features of real IS&R tasks, the data collection must be performed in a real-world setting. This, in turn, leads to the utilisation of data collection methods that can capture these features. Our intention is to capture as many, varied data as possible that reflect the patent handling process, which involves data generated by human activities during IS&R activities. This includes search logs, on-site observation of patent engineers performing their tasks, and their descriptions of their work – in electronic diaries. In order to do this, we need to utilise both qualitative and quantitative methods. In short, we will apply methods that a) combine qualitative and quantitative methods such as interviews (theme-based and with expert focus), participatory observations, electronic diaries, and database search logs and b) propose methods of analysing data in a systematic way.

For the present study, we will explore and describe real-world patent work tasks within the patent domain (at the Swedish Patent Registration Office) and the information retrieval and information seeking activities within the patent handling processes. Various features of patent IR will be investigated. We will also analyse individual and collaborative aspects of patent information retrieval.

1.4 Thesis structure and the research process

The present piece features both theoretical and empirical sections. Figure 1.1 gives an overview of the stages and the way in which the study was conducted.

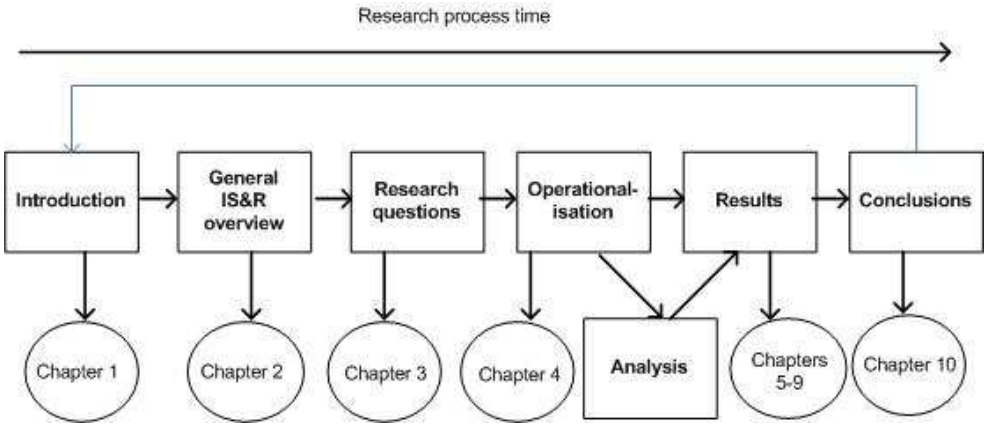


Figure 1.1: The research process

The structure of the dissertation is as follows. After the introduction provided by Chapter 1, Chapter 2 presents a general outline of the theoretical foundations, including established models within the information seeking and interactive information retrieval research area. The chapter also provides a discussion of information seeking and retrieval tasks embedded in a work task situation.

Furthermore, a specific section introduces the patent domain that will be the focus of our thesis. In Chapter 3, we discuss the research motivation and describe the main research questions for the reader. The main research question is partitioned into seven sub-questions. For each research question, we also describe the means for collecting data. Chapter 4 describes the design of the study and gives a detailed outline of the data collection and analysis process. We provide a detailed framework for the multiple qualitative and quantitative methods used for collecting data as well as how said data will be analysed. We also discuss the rationale for using these specific methods for our purposes. The chapter ends with on overview of the research steps and how the data will be handled.

The results of the study are presented in chapters 5 to 9.

In Chapter 5, the patent domain is introduced in general and SPRO in particular. The patent handling process and, specifically, the characteristics of the patent document are described in detail in order to embed the interactive information seeking and retrieval processes in a real-world context. In addition, a general conceptual model of the patent handling process is presented. An understanding of this context is important as background for the discussion of the analysis of the data in chapters 6–8.

In Chapter 6, research questions 1–7 are addressed from a descriptive viewpoint. Here we present the characteristics of each group of variables linked to the individual research questions. In this chapter, we also assign to each variable values identified in our data analysis.

Research questions 1–6 are addressed in Chapter 7, through cross-tabulation of the variables described in Chapter 6.

In Chapter 8, we separately deal with research question 7 and discuss the findings on collaborative information seeking and retrieval activities. Types of collaborative activities are described.

In Chapter 9, we present and describe a method of analysing and describing the captured features of interactive patent search sessions. Here we include both visualisations of query sequences and a schematic diagram of task processes, before, finally, Chapter 10 presents final discussion and concludes the thesis.

1.5 List of publications

In the course of preparation of this thesis, some of the results have already been published, in articles. Since the goal of this work was to write a monograph, we therefore present a list of articles that contain some of the results and material presented in this study:

Hansen, P. & Järvelin, K. (2005). Collaborative information retrieval in an information-intensive domain. Information Processing and Management (IPM) 41(5), pp. 1101–1119. Sep. 2005. Journal article.

Hansen, P. (2005). Work task-oriented studies of IS&R processes Developing theoretical and conceptual frameworks to be applied for evaluation and design of tools and systems. In: Theories of Information Behaviour. Fisher, K., Erdelez, S., & McKechnie, L. (eds). ASIST Monograph Series, pp. 392–396. Sep. 2005. Medford, NJ, USA: ASIST. Book Chapter.

Byström, K. & Hansen, P. (2005). Conceptual framework for tasks in information studies. JASIST - Journal of the American Society for Information Science and Technology 56(10), Number 10, pp. 1050–1061, 2005. Journal article.

Hansen, P. & Järvelin, K. (2004). Collaborative information searching in an information-intensive work domain: Preliminary results. Journal of Digital Information Management 2(1), 2004, pp. 26–30. Journal article.

Byström, K. & Hansen, P. (2002). Work tasks as unit for analysis in information seeking and retrieval studies. The Fourth International Conference on Conceptions of Library and Information Science: Emerging Frameworks and Methods. CoLIS4, Seattle, WA, USA, 21–25 July 2002, pp. 239–252. Conference paper.

Hansen, P. & Järvelin, K. (2000). The information seeking and retrieval process at the Swedish Patent and Registration Office. Moving from lab-based to real-life work-task environment. Proceedings of the ACM-SIGIR 2000 Workshop on Patent Retrieval, Athens, Greece, 28 July 2000, pp. 43–53. Conference/workshop paper.

Hansen, P. (1999). User interface design for IR interaction. A task-oriented approach. In: Aparac, T., Saracevic, T., Ingwersen, P., & Vakkari P. (eds). CoLIS 3: Proceedings of the Third International Conference on the Conceptions of the Library and Information Science, Dubrovnik, Croatia, 23–26 May 1999, pp. 191–205. Conference paper.

BACKGROUND

2

STUDY OF REAL-WORLD WORK-TASK-RELATED IS&R

This chapter provides the background for the study by discussing prior research and the concepts involved. Previous work in information seeking and retrieval (IS&R) research is presented.

The chapter is structured as follows. First, a general overview of literature in the IS&R research area is presented. In Section 2.1, we describe the basic concept of task and work task, followed by discussion of information access viewed in a work task setting (Section 2.2). Section 2.3 provides description of different approaches to information access. In sections 2.4 and 2.5, different models and frameworks related to information seeking research and to information retrieval research are presented. This is followed by a discussion of patent IR research (Section 2.6) and a presentation of collaborative information search, in Section 2.7. In Section 2.8, information use is discussed, before the chapter is closed by a summary (Section 2.9).

2.1 The concepts of task and work task

The concept of task is of increasing importance for a better understanding of IS&R processes. It is a fundamental concept to Information Science and Information Retrieval even though the models and methods that deal with tasks are heterogeneous (Hansen, 1999). The concept is utilised in the Information Seeking literature (e.g., Feinman et al., 1976; Mick et al., 1980; Kuhlthau, 1993; Kuhlthau & Tama, 2001; Rasmussen et al., 1994; Byström & Järvelin, 1995; Sonnenwald & Lievrouw; 1997; Solomon, 1997; Byström, 1999, 2002; Herzum & Pejtersen, 2000) as well as in Information Retrieval literature (e.g., Belkin et al., 1982a, 1982b; Marchionini, 1995; Ingwersen, 1996; Wang, 1997; Reid, 1999; Hansen & Järvelin, 2000; Borlund, 2000; Vakkari, 2001a).

Tasks and subtasks:
A *task* may be viewed as an abstract construction that may, in fact, contain smaller subtasks. It may also be understood from a functional point of view, from which a task is seen as a process wherein an actor performs a set of actions (physical and mental) in order to reach a goal. A task may be assigned to a human by another human, or the task may be constructed or designed by the task performer. A task has, both as a performed activity and as a formal description, a recognisable beginning and end. However, it may be difficult to assess when and where a task ends and begins, especially where the limits of a subtask of a main task are concerned (Vakkari, 2003).

On a high and abstract level, a *work task* is a sequence of activities that a person has to perform in order to reach a goal (Hansen, 1999). A work task can be a job-related task or a non-job everyday-life-related task[2] and may be either initiated by its performer or assigned (Hackman, 1969). The work task may be set, externally or internally, by a person, a group of persons, or an organisation, and within a professional workplace, there may exist a predefined set of work tasks that need to be performed. There may be established routines, formalised procedures, a predefined set of resources, etc. that are so obvious that the task performer or his or her employer does not reflect on their existence. In a work-related setting, a work task may lead to, or involve, a need for information, which, in turn, may initiate a search task.

A *task description* may be implicit or explicitly stated. The task description defines certain *requirements*, also providing a description of methods and strategies related to the requirements. Normally, a description also indicates that the task has a practical goal (a result) and it normally has a meaningful purpose (a *reason* for the task). A task that includes several specifiable smaller subtasks may involve individual requirements and goals for each of these. Each subtask may have different goals, requirements, and purposes; for example, a subtask may involve IS&R activities as well as other kinds of activities.

Subtasks may be accomplished separately and then brought together to generate a meaningful result. As an example, we may cite a situation in which the overall task is to give an answer (yes/no) to a request regarding water quality status from a microbiological standpoint. One of the subtasks may involve a search activity for seeing whether there are anomalies in the analysis process for the water. The seeking process is of great value for the microbiologist with regard to a final decision but not to the person who externally initiated the work task. Thus, IS&R activities may be subtasks but normally not the main goals of a work task. Furthermore, the IS&R activities are not independent from the work task. Finally, there may also be work tasks wherein a group of people work together to resolve a specific task or a group of tasks and each individual may perform his or her own subtask (Hansen & Järvelin, 2000, 2004, 2005).

Task characteristics:
The *characteristics* of tasks may have more or less impact on how the work task and its subtasks are approached, performed, and completed (Hansen, 1999). Work tasks

[2] Tasks in day-to-day life are usually non-job-related activities and may have cultural and social characteristics linked to, for example, entertainment.

may be constructed and perceived as *simple* to *complex* tasks (Byström & Järvelin, 1995), involving, for example, several subtasks, several sources, or a topic outside the searchers' domain knowledge. Tasks may also have a predefined *structure* (or lack of structure), and the structure may be a result of the planning stage of the task (O'Day & Jeffries, 1993a). Structured tasks have a designed course, whereas unstructured tasks may involve creative planning and flexibility. Also, tasks may be *subjective* or *objective*, where objective tasks may be understood as being external to the performer and imposed on him or her, independent of their performers (Hackman, 1969), while subjective tasks are viewed as internal to the performer and are often defined by him or her. In this way, one objective task may create and involve a set of subjective tasks that all may be distinguished from each other (Hackman, 1969; Byström & Hansen, 2005). Tasks can be *routine tasks or unique/specific tasks*. Repetitive or routine tasks may include specific subtasks, as well as specific tasks (Hill et al., 1993). More often than we think, we are *switching* between task activities rather than performing them in logical and serial ways (Preece et al., 1994; Hill et al., 1993; Belkin et al., 1993; Smith et al., 1997; Spink, 2004). Depending on shift in the information need, the task may take a new direction, involving different behaviour (O'Day & Jeffries, 1993a) and the continuity of the task may be stable or may shift in a new direction. *Task uncertainty* is another aspect to take into account. Kuhlthau's (Information Search Process (ISP) model of task uncertainty (1991) involves several stages of uncertainty, such as when a person becomes aware of lack of knowledge and understanding in order to formulate a personal point of view. It is also important to acknowledge *how* a task is *perceived* if one is to understand its relation to the need for information and IS&R.

Task performance:
Task performance takes place when a person is handling a particular item of (in our case) work, which means that the task is manifested through the person's goals, beliefs, strategies, and actual behaviour. From within the organisation, sets of more or less official and formal duties are involved in the work task and the organisation may outline different levels of tasks both implicitly and explicitly. Factors important in this process are the human searcher and his or her level of experience, task knowledge, and domain knowledge, as well as characteristics of the organisation, such as specific constraints and possibilities. Task performance can be divided into three main parts: task construction, task performance, and task completion.

Task performer's knowledge:
The task performed is a central part of the IS&R and often the performer of the actual IS&R task. Among the factors related to task performance are the task performer's prior knowledge, skills, and experience. Perception of the work task, along with prior knowledge and experience, may affect the information need, the search tasks, and relevance judgements (Ingwersen & Järvelin, 2005). While performing an IS&R task, the performer may have different degrees of knowledge about[3] a) the work task setting and its components, b) the specific type of task assigned, and c) the specific topic of the task. A task performer's behaviour within an organisation is generally

[3] It is necessary to mention these aspects of knowledge types related to the IS&R process, even though it is not the primary focus of this thesis.

guided by norms and value structures of the work organisation (Giddens, 1979). This knowledge may vary from person to person and in time. It may also be that a person possesses conceptual knowledge but lacks knowledge of how actually to complete the task.

On the *work task level*, knowledge of how to plan, structure, and perform the task stems from the task performer's knowledge of how the task is supposed to be performed (*procedural knowledge*) and *individual experience* (as from prior performance of similar tasks). With regard to the *IS&R* task levels, knowledge about *information sources* and information *systems* related to the task at hand are important – that is, understanding of the structure of the document representations and types, search strategies, electronic and human information sources, and *people and groups* (Hansen & Järvelin, 2004, 2005) as well as of how these are connected to the perceived information need.

2.2 Information access in the work task setting

Bennett (1972, p. 189) speaks of 'user task effectiveness in task performance' as an important element and thus points out that we need to look at how people actually are performing specific tasks. This implies that we must take into account the setting in which the user performs that task. This is also suggested by Rasmussen et al. (1994), Byström and Järvelin (1995), Kekäläinen and Järvelin (2002a), and Ingwersen and Järvelin (2005), who claim that users' work tasks and goals must be taken into account and understood when one investigates IS&R within a larger framework (see Figure 2.1, below). Recently, several attempts have been made at analytically bringing knowledge and empirical findings from IS&R fields closer to settings involving work tasks (Hansen & Järvelin, 2000; Vakkari, 2001a, 2001b; Järvelin & Ingwersen, 2004; Hansen & Järvelin, 2005; Byström & Hansen, 2005; Freund, 2008; Veinot, 2009).

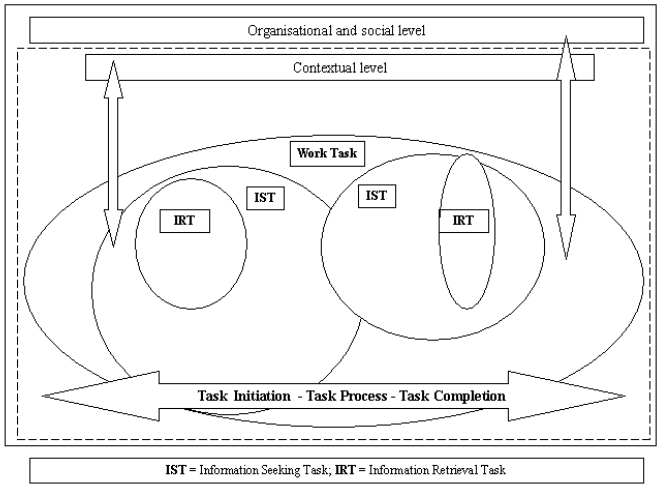


Figure 2.1: Task performance and relationships between the task levels in this study[4] [5]

[4] The figure is a revised version of a figure previously published by Hansen (2005).
[5] The dashed line incorporates the focus of our study.

The issue of context is often connected to tasks in general and work tasks in workplaces in particular. The concept of context has been discussed in depth in various research settings, meaning different things (e.g., from the human–computer interaction (HCI) perspective as a 'context-in-use' (Wixon et al.1990; Anderson & Alty, 1995)). Context of use is generally used to refer to the social, cultural, individual, and historical factors affecting how people manage their practices, whether these be job-related or daily-life-related. In the information seeking arena, Allen's (1997) model of 'person in situation' focuses on individual influences, situational influences, and individual and group needs as important factors.

Dervin (1997) concludes that it is very difficult to provide a description of how to approach the concept of context within the area of information seeking. It has proved difficult to establish a definition of the concept of context, which is reflected in the vast number of characteristics and attributes applied to context (ibid; Kari & Savolainen, 2007).

In our study, we apply a general definition of 'context' as the setting involving certain conditions (such as physical place and work duties), while a 'situation' is defined as a set of events or actions that may differ from one situation to another in consequence of the influences on a person's information behaviour, such as time constraints or lack of resources. For example, a classical IR situation features a common set of actions or events that may occur in different contexts, such as a medical vs. an academic context.

2.3 Approaches to information access

Information access is one aspect of the more general case of *information handling* and encompasses various types of information searching processes and practices (see Figure 2.2).

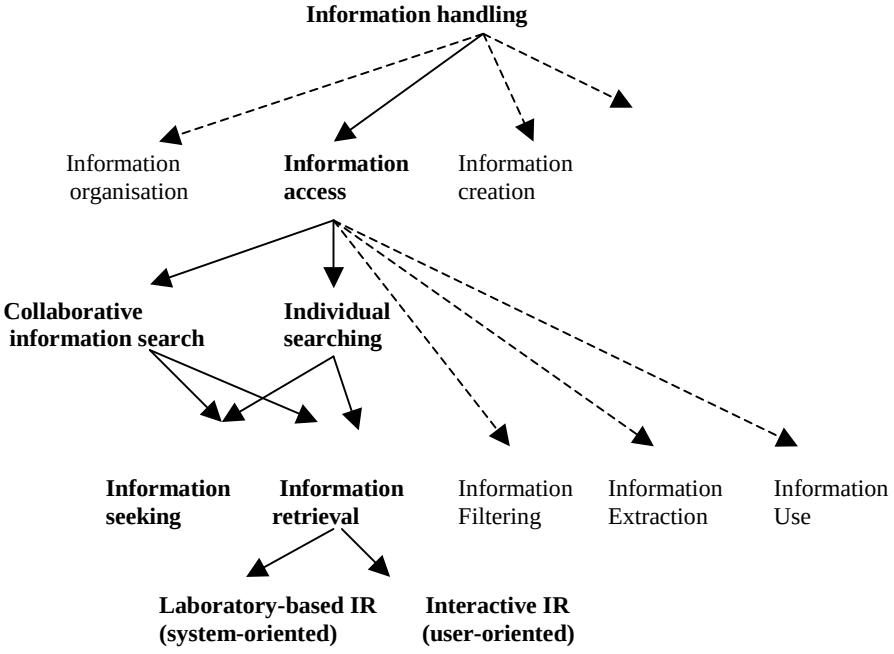


Figure 2.2: General levels of information access

These search activities may be performed *individually* or as a *collaborative* effort. Two examples of approaches for accessing information are *information seeking* and *information retrieval*. In addition, we may differentiate between the levels of *work task* as described above and the *search task*. The latter is further divided into the information seeking task and the information retrieval task. Wilson (1999) presented a similar division of activities with the corresponding levels of information seeking and information searching, surrounded by the main level of 'information behaviour'.

The two main research areas in study of information handling activities – information retrieval research and information seeking and behaviour research – represent many types of studies, ranging from lab-based (system-oriented) and tightly controlled experiments to studies of information search in natural settings, where studies of interactive and user-oriented search tasks can be found. Information seeking and retrieval (IS&R) is generally understood as encompassing complex and dynamic processes, given the great variations in the many components involved, such as retrieval systems, user groups, individual user behaviour, and user needs, as well as a variety of domains. Information retrieval research can be characterised by two major views: a system-oriented (or laboratory-based) and a user-centred view. The present thesis is concerned with a user-centred view of the interactive information retrieval area. There are now growing numbers of both theoretical models and conceptual frameworks that cover various levels of information seeking research and information retrieval research.

Next, in subsections 2.4 and 2.5, we present the background of information seeking research, followed by that of the information retrieval research area.

2.4 Research on information seeking research

In a work domain, a work task may lead to a particular information need, which may or may not activate a search task situation. A search task is carried out by an actor as a 'means of obtaining information associated with the fulfilment of a task' (Ingwersen & Järvelin, 2005, p. 20).

Since the beginning of 1960 (e.g., Taylor, 1968; Hackman, 1969; Bennett, 1972; Wilson, 1973; Feinman et al., 1976; Allen, 177; Bates, 1979a, 1979b), information seeking (behaviour) as a research area has received increasing attention. Many conceptual models and frameworks have been proposed and discussed. As Järvelin and Ingwersen (2004) point out, these models and frameworks cover a wide range of phenomena, such as information seeking stages, actors, seeking strategies, information needs, and sources. Studies in information seeking have mainly focused on the use of documents as well as on information channels that support different search-task-related activities. For this group of studies, the IR system is of limited importance, while the search processes, task levels, and information behaviour are of greater interest.

The connections between different activities and contexts that generate information seeking behaviour have been described in a number of general models and frameworks. Alongside more theoretical models, a few limited attempts have been

made to describe the relationships between information seeking and specific features of work activities more empirically. Below, we attempt to describe some of the most important models. In the associated studies, different features have been cited to explain the variation in, for example, use of various types of information and channels.

In 1981, Wilson proposed a model of information seeking behaviour that was based on the importance of an individual's physiological, cognitive, affective, and perhaps other needs. Here Wilson suggests that these needs may be seen as the context of, for example, a person or the environments involving work tasks. Thus, information seeking is seen as embedded in the activity or context that generates information seeking behaviour (Vakkari, 1999).

One influential approach is called the Sense-Making approach. Proposed by Dervin and Nilan (1986), this approach calls for a change of focus with respect to the human involved in the search activity. Dervin and Nilan suggest a shift in the consideration of the information seeker, from users to the 'actor'. The information systems should be viewed and assessed from the actor's point of view. At its base, the Sense-Making model employs three important labels: 'situation', 'gap', and 'use'. The sense-making approach regards the information need situation as the situation in which the actor needs to create new sense. This information need is the sense-making situation. The sense-maker (actor) is stopped by some kind of gap in a specific situation. To bridge the gap between the need situation and (information) use, the actor (sense-maker) examines the possibilities for overcoming this gap (that is, to answer the questions). Information use has been conceptualised as the different ways in which actors 'put answers to questions' (p. 22). The use of information is then considered situational.

The sense-making approach does not specify any relationship between components of the model or really model the work task aspect of information seeking. However, the model is important in that it focuses on the actor as well as on the sense-making situation. A sense-making model has been used in other domains also (Jensen, 2009).

Furthermore, some studies have focused on empirical examination of information seeking and search strategies, such as those of Kuhlthau (1993a), Ellis (1989), and Ellis and Haugen (1997). Kuhlthau discussed learning tasks and problem solving as a process from an information seeking perspective and considered empirical findings from longitudinal studies of students and library-users.

The IS&R process is described from a psychological perspective, including in its affective (feeling), cognitive (thought), and physical (action) elements, and is further described as featuring six stages of the search process: a) task *initiation*, b) topic *selection*, c) pre-focus *exploration*, d) focus *formulation*, e) information collection, and f) presentation. Kuhlthau's model is based on data from one type of task, so the model may be applicable to only one task type (student learning tasks), although the claim is more general. Kuhlthau developed her model further and applied it to the work domain of security analysts (1997). In this case study, she compared a person's perceptions at the start of his or her career with the perceptions held five years later. It was found that uncertainty in the information search process is an important element in the workplace.

In 1989, Ellis presented a set of features involved in information seeking:

- Starting, which involves the means by which the user begins the seeking process
- Chaining – following, for example, citations in known material
- Browsing
- Differentiating – using known differences in information sources as a way of filtering information
- Monitoring – current awareness searching
- Extracting – selecting relevant material from a source
- Verifying, which involves considering the accuracy of information
- Ending, which involves the means of closure

As Wilson (1981) did, Ellis emphasises that the interaction of these features within a seeking activity will depend on the specific circumstances of that activity at any given time.

These features are further elaborated upon in an empirical study of research scientists and engineers in an industrial setting, by Ellis and Haugen (1997). The features of information seeking behaviour described correspond to the seeking patterns of the real-life search situations of the engineers investigated. The features were starting, chaining, browsing, differentiating, monitoring, extracting, verifying, and ending. The empirical findings were based on data collected from several types of tasks, but no specific information about the tasks was given. The strength of Ellis and Haugen's study from 1997 is that the features that were defined in the paper from 1989 (Ellis) now were tested in a real-life situation. However, the interrelationships and any dependencies affecting or among these features are not discussed in depth. That is, the model may describe the process of information seeking and its activities, but the set of features does not explain how these features relate to real work tasks.

Leckie et al. (1996) present a model of the information seeking of professional engineers, in which they assume that information seeking is connected to different roles and tasks linked to these roles. The model describes particular roles and their related tasks as creating information needs, which have different characteristics and are, in turn, affected by factors such as source, individual characteristics, and environmental factors. Important in this model is the relationship between work roles and their connected tasks with an impact on the seeking process. Leckie et al. mention one specific factor, awareness of sources, pointing out that colleagues are a very important source. Leckie et al. continue by saying that, as engineers do, lawyers tend to rely on personal experience and knowledge when choosing information sources.

Task complexity is another feature that has been given attention, by Byström and Järvelin (1995), and by Byström (2000). They studied information seeking task performance in a real-life setting (among municipal workers in Finland). Their study focused on levels of task complexity and how it affects the task outcome. They viewed the task-based information seeking as a problem-solving process. The study examined interrelationships of components such as information channels at the information seeking level. However, it did not investigate the IR level in greater depth. The framework introduced does not discuss the integration of different levels of both IS and IR.

Vakkari (1999, 2000b, 2001a, 2001b, 2003) has investigated the work and search tasks as important components in the understanding of IS&R. Vakkari (1999)

presents a model with interlinked components of task complexity and information actions in work environments. On the basis of variations in information-related actions, Vakkari proposes relationships between categories of information activities, complexity of tasks, and problem structure. Vakkari and Hakala (2000) presented a study of students writing up their master's theses over a four-month period. Among other things, the students' understanding of the task during its performance and the use of search terms and tactics were investigated. Vakkari (2000a) found that a person's problem stage during task performance is related to the use of relevance criteria.

Vakkari (2001b) presents a longitudinal study showing that stages in task performance were systematically connected to the information sought, the search tactics used, and the usefulness of the information found when one was writing a research proposal. On the basis of a set of hypotheses, Vakkari also suggests a theory of the task-based IR process, which is an extended version of Kuhlthau's ISP model.

Vakkari (2003) also reviews studies that deal with the relationship between task performance and information searching by end users. Descriptively, Vakkari highlights important aspects for pursuit of task-based studies, pointing out that, before 2003, the object of the studies had almost always been the research process in academic settings. Others had been scarce. Vakkari concludes that there is a set of limitations that need to be considered in task-based studies, of which the following are relevant for our work: few studies taking tasks as a starting point, almost always an academic setting, lack of longitudinal studies, and studies seldom focusing on the whole searching process.

Byström and Hansen (2002, 2005) discuss both theoretical and conceptual foundations for task-based research, and tasks are defined at three levels that are relevant for information studies: work tasks, information seeking tasks, and information retrieval tasks. Byström and Hansen (2002) argue that work task performance provides a common ground for IS and IR studies and that this approach is useful for bridging the gap between IS and IR research. Byström and Hansen (2005) discuss the concept of task in the context of information studies in order to provide definitional clarity for task-based IS&R studies. Central task levels are defined and the analysis is aimed at providing a conceptual starting point for empirical studies in the relevant research area.

Pharo (2002) developed a method of analysing Web information search processes, for understanding how work tasks may affect information seeking and searching in the Web context. The study had a task-based focus and, through generalisation, the outcome of the study addresses task-based IS and IR. In its methodology, this study is relevant since it used multiple data collection methods (log statistics, a questionnaire, interviews, observations, and video recordings). Pharo used triangulation in order to describe the users' search sessions.

Järvelin and Wilson (2003) report and discuss important features in relation to how conceptual models may contribute to scientific research. They discuss task complexity as well as task categorisation in terms of five task categories (see the work of Byström and Järvelin (1995), as referred to above):

- Genuine decision tasks

- A known, genuine decision task
- A normal decision task
- A normal information processing task
- An automatic information processing task

In order to specify a task as belonging to one of these five task categories, one often needs to relate the task to levels of complexity. The level of difficulty for each of the task categories depends on three components: the task input, the task performance process, and the task outcome. For example, for automatic information processing tasks, the type of task result, the work task process, and the types of information used can be described in detail. In another case, a known and genuine decision task may involve an *a priori* known type of result while the procedures for performing the tasks are not known. Therefore, the process for the task cannot be determined (Järvelin & Ingwersen, 2004). These categorisations may be very useful for studies examining and determining task types. Furthermore, these task types are also related to the level of actor knowledge or experience. A specific task type may be experienced as difficult by one person but as a normal task by another person.

Järvelin and Wilson (2003, as originally reported by Järvelin & Repo, 1983) suggest another interesting categorisation. Three, orthogonal categories of information types are described:

- Problem information (PI)
- Domain information (DI)
- Problem-solving information (PSI)

With problem information, the structure, properties, and requirements are described (similar to background description), while in problem-solving information, it is the method, how to treat the problem, that is covered. This involves how a problem should be formulated and treated, what information should be used for solving the problem. Finally, the category of domain information involves the facts and concepts in the specific domain of the problem.

In summary, two important classifications have been introduced: categorisation of tasks and categorisation of information types. Both have been empirically tested by Byström and Järvelin (1995).

In a study reported upon by Savolainen (2007), the findings were based on interviews with 20 environmental activists in Finland. The problem investigated was information overload and how people coped with it. The findings showed that people used two strategies to handle information overload: the filtering strategy and the withdrawal strategy. As the author points out, the study's main contribution is that it provided empirical knowledge of different strategies when monitoring everyday events through media. Furthermore, the two strategies are often used together in an everyday context.

Information seeking performance:

Information seeking is initiated with a recognised need for information and then a decision to react to it (e.g., Wilson, 1999). A person or group of people probably start with some notion of what is wanted or needed within a specific situation. Preece (1994) calls this the *intentional level*, and Ingwersen (1996) mentions it as the underlying intentions in the desire for information. At the *functional level*, there is a

sequence of actions that are performed by a person, a group of people, or a machine as a response to the recognised need.

The *task construction phase* involves the recognition of a need for information and a decision to try to satisfy this need on the part of an individual or group of individuals and consists of a set of preconditions and goals for the performance and the planning of time and resources. This phase corresponds to the initiation stage described in Kuhlthau's ISP model (1993a). The task may be either assigned or self-generated (Hackman, 1969) and may refer to operational demands and desired goals with the task at hand. Task construction plays a major role for the other parts of the task process but is very difficult to observe directly (Byström & Hansen, 2002).

A central concept in the construction phase is the formation of the *information need*, which occurs when a user recognises a desire for information created by some kind of everyday life or work task. Taylor (1968, as reported by Ingwersen, 1992) and Belkin et al. (1982a, 1982b) suggested that the recognised 'incompleteness' or 'Anomalous State of Knowledge' is the mental state for the development of the information need. The information need is then transformed and expressed as a statement representing that information need. When identified, the need for information may lead to the adoption of an information seeking strategy in which different information *sources* are considered and interacted with. These sources may be human-, paper-, or computer-based, such as an IR system (Hansen & Järvelin, 2004). Those channels and sources perceived to be available and also experiences (negative or positive) of past use are likely to affect the decision to refer to them again (e.g., Gerstberger & Allen, 1968). It is the practical actions of task performance that we are able to observe directly (Byström & Hansen, 2002).

As part of the construction phase, seeking *strategies* are plans and actions for acquiring and retrieving information. Bates (1990, p. 578) defines a strategy as 'a plan, which may contain moves, tactics and/or stratagems, for the entire information search'. Bates also relates the different levels of information seeking to different task situations. The information seeking *strategy* is linked to the selection of source; these strategies may be adjusted in all contexts (e.g., Kuhlthau, 1993a; Vakkari & Hakala, 2000), or they may be stable and formalised procedures closely linked to stages in a professional task. This also corresponds to the stages of selection and exploration in Kuhlthau's ISP model (1993a). Kuhlthau's fourth stage, 'formulation', is, in our context, found partly in the construction phase and partly in the performance phase.

The actual *task performance* consists of practical actions and conceptual actions taken in order to satisfy the need and reach the desired goals and are related to the task performer's perception of the task requirements. The task performer now activates prior knowledge concerning various information channels and sources such as personal networks and other formal and informal information channels (e.g., Kuhlthau, 1993a). This phase is called 'execution' by Marchionini (1995) and is in line with what Hackman (1969) terms 'hypotheses' and 'process'. In Kuhlthau's model, this is the fifth stage, called 'collection', and involves the most intense activities between the user and the information system (Kuhlthau, 1993a).

The *task completion* phase for information seeking tasks may be described as a situation in which the information seeking process is successfully completed (or

unsuccessfully terminated) and sufficient relevant information for meeting the perceived information need and task requirements is collected. This may lead to the closing of the task or a reconstruction of the task or in leaving the task uncompleted (Feinman et al., 1976). If the task performer faces difficulties in formulating his or her information need (Ingwersen, 1992), problems may arise also in defining relevance criteria. Kuhlthau calls her closing phase 'Presentation' (Kuhlthau, 1993a), and Marchionini (1995) refers to 'evaluation and use'. Either the beginning or the end of a task, and perhaps both, may not be anticipated and may be recognisable only in retrospect. However, in some specific cases, an individual subtask may constitute a well-defined item of work that has a recognisable start and end and may also create a single result. Such well-defined subtasks will be investigated in the present study.

In summary, research into information seeking focuses on the user's (or actor's) information behaviour, including features such as understanding of the task and problem, information need formulation, and relevance assessments. Existing empirical studies encompass academic students and professional practitioners as well as everyday information seeking behaviour. These studies often provide us with an understanding only internal to the domain studied. Furthermore, they generally do not inform us how findings on behaviour and features could be applied to IR processes.

However, as pointed out by Järvelin and Ingwersen (2004), the benefits of information seeking research are threefold: they a) provide a theoretically grounded understanding of information seeking through models and frameworks, b) give empirical descriptions of information seeking behaviours and processes in different domains, and c) provide support to design of information systems. However, the latter have been most problematic within the information seeking research area. Järvelin and Ingwersen (ibid.) conclude that research in information seeking must be extended both toward the task context and toward the information systems context.

2.5 Information retrieval research

In this section, we describe the background of information retrieval research and, in particular, the user-oriented and interactive information retrieval (IIR) approach. Information seeking studies (described in the previous section) and IR studies serve as the background from which the present study sets out. Accordingly, the study is not only an information seeking study, nor a system-oriented traditional IR study. It has its focus at the intersection of these research areas: interactive information retrieval (IIR).

The *system-oriented and lab-based IR research*, such as the Cranfield[6] and TREC[7] experiments, is concerned with exploring ways of improving the matching algorithms of text representations and queries for retrieval of relevant documents. The Cranfield experiments, begun in the 1960s, are often mentioned as the starting point for classical computer-based IR evaluation activities (Cleverdon, Mills, & Keen, 1966). The goal with the Cranfield studies was to test different ways to improve the retrieval

[6] The Cranfield tests (I and II) developed and used a model for evaluation of the effectiveness of IR systems. They were organised by ARPA (the Advanced Research Projects Agency).
[7] Text REtrieval Conference (see http://trec.nist.gov/).

effectiveness of IR systems' enhanced indexing methods. This was done in highly controlled test settings and usually included a test collection (documents) and a set of topics and set of documents judged to be relevant for the topics used. In this approach, measures such as precision and recall were (and still are) used. This approach does not address task types, information need types, and different types of users and their behaviour in interaction with the system. Situational and contextual aspects and factors are not considered, overall. Recent IR-related experimental platforms that engage considerable research effort are the CLEF[8], NTCIR[9], and INEX[10] IR evaluation campaigns.

According to the traditional IR model (see Figure 2.3, below), IR systems[11] are generally built on the idea that all information (usually text-based) can be stored; organised; and, through various text analysis processes, also represented[12] in such a way that it can be matched with a query representing an information need and then retrieved.

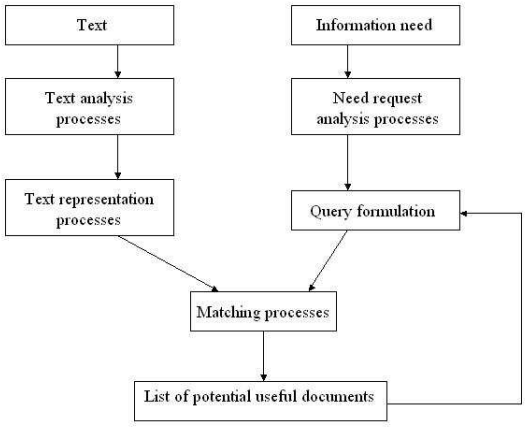


Figure 2.3: Traditional IR model

As can be seen above in Figure 2.3, while the traditional lab-based IR model does not include the information-seeker, user-oriented IR research (such as that following the cognitively oriented approach) involves users, to varying extent.

Interactive IR, including the cognitive IR approach, focuses on the user and his or her behaviour. The focus is usually on the interaction between the system and the user (or the cognitively oriented IR). This intersection between system- and user-related (cognitive) features of IIR studies is not without complications. Even though studies in interactive information retrieval are performed in a relatively controlled setting, it involves studies of human behaviour that may entail uncontrolled conditions – for

[8] The Cross-Language Evaluation Forum **(**see http://clef-campaign.org/).
[9] NII-NACSIS Test Collection for IR Systems (see http://research.nii.ac.jp/ntcir/).
[10] The INitiative for the Evaluation of XML Retrieval (see http://clef-campaign.org/).
[11] In this study, an IR system is an *electronic* information system with the task of matching a query to representations of a set of documents and, from this matching process, presenting a retrieved subset of documents.
[12] This representation is usually in textual form and could consist of one or more sentences from the full text, parts of the text, or any of various descriptions of the text.

example, when the researcher sets out to perform studies in real-life work-based environments with users performing real job-related tasks. This is a problematic issue that has also been highlighted by Ingwersen and Järvelin (2005) and Kelly (2009). Still, many studies within the area of IIR are conducted in a controlled system-based setting (a laboratory), as with various interactive tracks in the TREC and CLEF work. In general IR research, the system-based studies are concerned with results and experiments that can be validated and replicated, and so provide reliable outcomes. At the same time, there is increasing demand for contextual features and components to be involved in IIR research studies – for example, people, tasks, interaction, and socio-organisational elements. Furthermore, from a computer science point of view, IIR work may also be concerned with the specific features of the *means* (e.g., the user interface) through which the interaction takes place. However, this issue will not be investigated in the present thesis.

As mentioned above, closely related to interactive IR – and sometimes seen as included therein – is the *cognitive IR research* approach. This research approach is seen to consist of a set of cognitive structures and of processes that involve interaction between system characteristics, the user's characteristics, and the functionality of the user interface (Ingwersen, 1992). Interactive IR studies are also conducted with individual platforms such as TREC, CLEF (e.g., as iCLEF), and INEX. In these evaluation campaigns, components such as the effect of a user interface are examined with the purpose of suggesting support for the user during the search process. According to Järvelin (2007, pp. 972–973), there may exist several cognitive approaches. These approaches involve different conceptual models and levels of focus, coverage, and analysis. One such approach is the simulated work task.

The Cognitive Viewpoint has its roots in a workshop held in Ghent in 1977, and its most influential components are discussed in depth by Ingwersen and Järvelin (2005, pp. 23–30). De Mey (1977) proposed several, interrelated dimensions – for example, that information processing takes place at different levels; that the actor is influenced by his or her experiences as well as the social, organisational, and cultural environment; and, finally, that information is situational and contextual. These aspects have clearly been a vital part of the cognitive IR approach. Fidel (1984, 1985) investigated cognitive styles employed by experienced online searchers when performing their daily work tasks. The searchers were then categorised as either operationalist or conceptualist searchers, where the former used a large set of IR system capabilities, such as focus on high precision and modified queries non-conceptually, while the conceptualists focused on modifying concepts and terminology for high recall. Our study connects some IR components, such as relevance and request formulation, from a daily work task standpoint.

Ingwersen (1992, 1996) has elaborated upon the cognitive viewpoint and extended it into the research area of Information Science, also introducing the cognitive information concept in information seeking and retrieval. In 1992 and, in more depth, in 1996, Ingwersen described a cognitive model of information transfer. This model has five important parts: a) the information object, b) the IR system, c) the interface (or intermediary), d) the cognitive space including the individual users, and e) the social and organisational environment including work tasks. Between these exist cognitive transformations and influences and interactive communication of cognitive structures (Ingwersen, 1992, p. 148). The interaction takes place between the actor

and the interface. The entities and the relationships between them are important and vital for the present study. From the cognitive theory for interactive IR (Ingwersen, 1992, 1996), the principle of polyrepresentation (or multi-evidence) was developed, to describe the cognitive overlap of information objects represented through cognitively different information structures. The cognitive approach has been further elaborated upon and a cognitive framework of (longitudinal) IS&R presented (Ingwersen & Järvelin, 2005, p. 274). In this model, interaction and perception are the central interactive IR processes. For example, the perception of the work task is the central factor affecting IS&R. In addition, interactive IR takes place 'via requests, information acquisition, relevance assessments and feedback' (ibid., p. 275).

Wilson (1999) describes a set of user-oriented research models related to information behaviour studies, which includes information search. In his model, Wilson points to the importance of the contextual aspects and states that IR is embedded in information seeking processes, which, in turn, is one of the information behaviour activities. As Ingwersen and Järvelin (2005) point out, Wilson considered the cognitive model of information transfer (Ingwersen, 1992, 1996) to be part of the IR-related branch of user-oriented research and not really connected to the information seeking processes.

As mentioned previously, Vakkari (2001b) shows that stages of task performance were systematically connected to the information sought, the search tactics used, and the usefulness of the information found when one was writing a research proposal. He proposed a model suggesting that information seeking should be viewed in relation to stages of work tasks and task performance. In this model, Vakkari incorporated not only work tasks and domain knowledge but also IIR-related elements, such as relevance assessments and degrees of relevance, as well as aspects of information use. This model represents a task-based approach and contributes to the IIR research.

One solution to the problem of balance between, on one hand, the demands of repeatability and validity (and use of proper measurements in evaluation) of an IIR system and, on the other hand, the urge to investigate how people (individuals or groups) interact with information systems was the development of IR experiments using simulated work task situations.

Borlund and Ingwersen (1997) and Borlund (2000) used the application of a simulated work task when investigating both Boolean and best-match interactive IR performance evaluation. This approach mixed features of system-based and cognitively oriented research for the evaluation of IIR systems. The motivation for using a simulated work task was that this would be closer to a real-life situation in the study of users performing search tasks with an IR system. The simulated work task situation consisted of two descriptive parts: the simulated work task condition and an indicative request (ibid., pp. 115–116). The simulated work task may describe the reason for the information need, the overall problem that needs to be solved, and the objective or goal of the search.

Hansen and Karlgren (2005) introduced an additional level of general description of a domain and a work task description that may be used to investigate the interaction between human and system (p. 638). As part of a cross-lingual IR study, they performed an experiment-based study and proposed an extension to the idea of using simulated work tasks by introducing and applying an expanded work task description.

This was done in order to move toward a more realistic study setting. The first level included a description of the domain and general work tasks usually performed within the specific domain of interest. The second level included a situational description that featured the topic of the query and a search task description (see Figure 2.4).

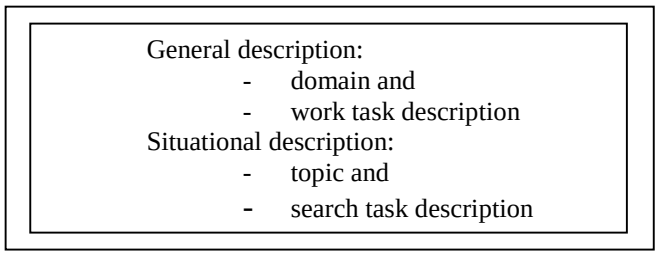


Figure 2.4: The two-level scenario description framework

The study showed that work task scenario descriptions had an observable effect on the retrieval process. Giving subjects more information about the work task meant that they spent more time working on the search task, and by self-report the subjects used differing strategies to pick out documents for perusal, dependent on the scenario. Furthermore, effects on results by traditional relevance ranking were detectable. This may be an argument for extending the traditional IR experimental topical relevance measures to cater for context effects (ibid.).

Elsweiler and Ruthven (2007) proposed a task-based evaluation methodology wherein one would utilise a combination of a naturalistic approach and controlled experiments in a laboratory setting. They concluded that in studies of personal information management (PIM) situations, both system and user should be examined when the users actually perform their own tasks. These studies show the importance of moving toward more realistic study settings for understanding IIR.

The interactive information retrieval process:

The *traditional IR* approach is mainly concerned with improving the effectiveness of automatic searching techniques and has therefore been criticised for neglecting issues such as cognitive and interactive aspects (Saracevic, 1995, 1996; Ingwersen, 1992, 1996).

IR systems apply different *IR techniques*, supporting different search strategies. *Information* or information objects vary in representation and structure. There are different types of information objects and different information *resources*. *Users* have different preferences, experience, and knowledge of the subject domain and IR, and they also apply different strategies when performing a search task (Belkin et al., 1993). In addition, users have different tasks and goals, with various characteristics, that influence the way the user approaches the IS&R activity (Hansen, 1999).

As in the case of the IR level, *task construction* resembles to a great extent the previous task level of information seeking. However, since IR tasks are seen as part of the information seeking phase, the construction of the IR task is done in a specific context and situation and may then be affected by the characteristics of the work task.

As regards *task performance*, whereas information seeking tasks may focus on satisfaction of the full information need, composed of, for example, different information types and topics, as well as using several consultations of channels and sources, the information retrieval task may focus on the satisfaction of a portion of an information need through a single consultation of one or more sources. Thus a task-performer may seek information from one or more sources in a search episode (Belkin et al., 1995). For example, utilisation of either single or multiple electronic databases through a search interface would constitute one information retrieval session.

An *information need* is identified as a gap in a person's knowledge base (e.g., Belkin et al., 1982a, 1982b) and recognised as an 'Anomalous State of Knowledge'. The mental state of the information need is then transformed and expressed as a statement representing that need. The information need may then be formulated as a request for information.

This *request* is then transformed and formulated in a *query* by the searcher in order to explore one or several IR (database) systems. A query may consist of one or more search terms or similar query attributes, such as classification codes or dates, and can also have logical operators as separators. A term is any string of characters (e.g., letters and numbers) with no space between them. For example, words, classification codes, and abbreviations may be considered as terms. The process of formulating a query is a result of the perceived information need and requires various knowledge levels, such as domain and subject knowledge, as well as knowledge and experience of interacting with certain specific sources (cf. Hsieh-Yee, 1993); domain knowledge is the knowledge a person has in a specific domain.

Here, *source selection* involves the decision to approach an electronic information source. This implies that the task-performer has some prior knowledge of how to operate the source. The selected source is most likely to contain documents that are relevant to the query (Baeza-Yates & Ribeiro-Neto, 1999, p. 451). Selection of sources is seldom done randomly.

Q*uery reformulation or subsequent query*. In the interactive process of retrieving information, the task-performer may need to reformulate the query as a result of, for example, inappropriate term selection or overly narrow/broad term selection.

A *session* is a coherent unit or set of queries submitted by the user during an interactive episode between the user and the IR system. A session may consist of more than one search. It might be represented, for example, by the actions bracketed by opening and closing of an IR system. In our case, a search session involves the interaction with one IR system. However, this definition is problematic, since a specific search strategy may be repeated two times, with two different search systems, with different contents or functions, either serially or in parallel.

R*elevance judgement* is a central concept in information sciences. A large amount of literature exists on the topic. Relevance judgement may be understood as the process of assigning a (binary) value of relevance to an object or information – expressing the level of relationship between the user's information need and a given document at a specific time (Wilson, 1973; Mizzaro, 1998) in a set context. Thus the relevance

assessment is merely situational and implies that the search task is connected back to the work task at the same time, connected to contextual conditions and how well the search will support the accomplishment of the task (Wilson, 1973; Schamber et al., 1990; Saracevic, 1996; Vakkari, 2003). As stated by Cosijn and Ingwersen (2000), relevance always implies a relation (p. 537). Mizzaro (1998) proposes a model wherein relevance may have the following components: *topic, task, and context*. For example, the component 'task' is referred to as the activity the user will perform with the retrieved documents (p. 8). Kekäläinen and Järvelin (2002b) describe the application of different degrees of relevance; the user assesses not only the document as relevant or not but also how relevant it is (not relevant, partially relevant, or fully relevant). Topical (or subject) relevance involves the relationship between the query topic (the search task) and the topic covered (the information object). As Cosijn and Ingwersen (2000, p. 539) point out, the relation is basically system-oriented, although the request (later transformed into a query) is formulated by a user. System (or algorithmic) relevance is about the relations between the features (e.g., words) in a query and an information object. The similarity is usually measured statistically, according to Cosijn and Ingwersen. Other manifestations of relevance include cognitive relevance, which deals with the relation between the cognitive information need of the user and the information object, and motivational relevance, involving the relationship between goals and motivations of the user and the information object.

There are different approaches to *relevance feedback* (RF). Relevance feedback may a) involve the user marking relevant or non-relevant documents. The information retrieval system then uses features (e.g., terms) from these documents to enhance the query. In the next approach, b) the system may suggest a list of new terms to the user, from which the user then makes a selection. The list of terms is then used to augment the query (Hearst, 2009). Finally, there is c) a third approach, called pseudo-relevance feedback. Here, the system assumes that the highest-ranked documents also are relevant and from these documents the system identifies terms and automatically augments the query (Croft, 2009). The outcome of the reformulation of the query will lead to the new query being 'moved towards the relevant documents and away from the non-relevant ones' (Baeza-Yates, 1999, p. 118).

Task completion. The information retrieval task is completed through examination of the results and reflection upon them leading to an end to the process (possibly after an additional iteration of the phases). In some cases, retrieval is aimed only at fetching of known documents, which then may not even be read, just identified. As described in the discussion of the information seeking level, sometimes the task requirements and the result of the retrieved information do not match, and this may lead to closing of the task or a reconstruction.

2.6 Patent IR research

Information search of patent documents and related Intellectual Property (IP) material is an increasingly important issue for the business, scientific, and legal community. The patent domain is also of interest for a growing lay and research community.

As have other professional work environments, the patent domain has quickly changed from being paper-based to being almost entirely addressed via electronic

databases for searching and other information handling tools. Over the last 20–30 years, much of the research and development of systems and tools for handling patent documents has been performed within the database management research area (Joho, Azzopardi, & Vanderbauwhede, 2010). In the last 10 years, the IR community have shown increasing interest in and awareness of the specific characteristics of the patent domain (Leong & Kando, 2000). Today, there are several information evaluation campaigns for research on patent documents.

Patent IR:

In the IR community, research on patent information retrieval has been performed both as individually reported research activities and around evaluation campaigns. Currently, there are three major evaluation campaigns that deal with patent IR: the NII Test Collection for IR system (NTCIR); the Cross-Lingual Evaluation Campaign (CLEF, which involves the CLEF-IP track[13]; and, finally, a series of symposia, conferences, and workshops initiated by the Information Retrieval Facility (IRF[14]).

The NTCIR hosts a series of evaluation workshops. It was originally focused only on retrieval from Japanese documents and cross-lingual information retrieval, with the first NTCIR workshop held in Tokyo in 1999. The workshops usually draw attention to research in the IR and Natural Language Processing (NLP) research fields. The NTCIR evaluation deals with both traditional lab-based IR testing and more realistic evaluation involving users. The NTCIR-3 Workshop, in 2001, was one of the first concerted attempts to improve patent document retrieval, and since then there has been a separate patent evaluation track. The main task has involved cross-lingual and monolingual retrieval tasks as well as patent translation and patent-mining tasks.

The CLEF evaluation campaign promotes research into multilingual information access. To realise this, CLEF develops infrastructure and test collections. One aspect of this is work on cross-lingual patent retrieval (CLEF-IP). In multiple-language search, problems such as general concepts and terms, acronyms, and new words and concepts used in patent applications complicate the assessment and performance.

More recently, another platform for intellectual-property-related research has emerged. The Information Retrieval Facility (IRF) is an independent, not-for-profit research institute. The institute began operation in 2007 with the goal of promoting and facilitating IR research for industrial take-up. In order to do this, the IRF facilitates large-scale IR experiments and works via various events, such as the IRF symposium, to bring people from the IP industry and academia together. The IRF also organises the CLEF-IP track and an IRF report is issued each year (e.g., Piroi & Tait, 2010).

The TREC evaluation campaign started in 1992. For two years now (from 2009), there has been a track called ‘TREC-CH’, dealing with chemical IR in order to develop and evaluate technology for searching chemical documents. The interested participants are mainly patent-searchers and chemists.

[13] http://www.ir-facility.org/clef-ip.
[14] http://www.ir-facility.org/.

Research in patent IR:
To understand the context of the patent domain better, it is necessary to undertake studies of the various users involved in patent handling, the tasks they perform, and the different information handling processes that occur. However, in moving to a specific domain, such as the patent domain, highly specific characteristics will come into play. Patent documents have several specificities and characteristics, which suggests that we need to go beyond general-purpose IR research.

Retrieval of information:
Retrieving documents, extracting patent information, and analysing patent text documents are some tasks that require text processing. For example, the patent document is generally written in a highly structured way, but sometimes it employs very generic terms and concepts, in order to cover as much as possible of the invention described. This means that the person who wrote the document is handing over this description to the patent engineer for interpretation. This specific problem involves reduction in precision.

Since a patent document may contain different types of text, research has been performed on techniques for enhancing patent analysis. Tseng et al. (2007) describe various text-mining techniques suggested for the analytical process of patent handling. Among these are text segmentation, text summarisation, and feature extraction. The authors describe a methodology intended to improve the analysis process. They also describe a typical patent analysis scenario involving the following stages: a) task identification, b) searching, c) segmentation, d) abstracting, e) clustering, f) visualisation, and g) interpretation (ibid., p. 1217). Because of the problematic aspect of long and complex sentences within the claims, which often incorporate multiple descriptive elements, Parapatics and Dittenbach (2009) discuss from an NLP perspective how to process the claims automatically such that several separate parts may be extracted and analysed. Another domain-specific patent-related issue is chemical IR. As do patent documents, chemical documents have specific characteristics (Zhu & Tait, 2008); they also involve an information need that differs even more distinctively from general-purpose IR These specific characteristics are found also in patent document and involve chemical names referred to in various ways, describing transformation of chemicals and visualising chemical relationships, among other things. According to Zhu and Tait (ibid.), this also leads to increasing interest in extending document-centric retrieval more toward entity-centric retrieval.

Search tasks:
Bonini et al. (2010) present and discuss a list of patent search tasks. These are patent search, patent analysis, and patent monitoring (p. 32). In a study by Tseng and Wu (2008), a list of different search tactics or tasks is provided and discussed: a) to find source information; b) to develop, select, and combine search vocabulary items; c) to link related information or patent search Web sites; d) to screen search results; and e) to store and manage retrieved patent information. Tseng and Wu's a–b correspond to Bonini et al.'s patent search task, and Tseng and Wu's tasks c–d correspond to the patent analysis task. However, the monitoring task does not correspond to the 'store and manage retrieved patent information' search task/tactic described by Tseng and Wu (ibid., p. 34).

Users and patent search behaviour:
More recently, studies of patent users have been conducted. Newton (2000) presents a study covering 277 patent information specialists at the British Library Patent Information Centre. This study shows that users, through emerging Internet technology, increasingly utilise patent databases on the Internet.

Bonini, Ciaramella, and Corno (2010) emphasise the variety of users involved, since more and more people have access to patent information sources. Professional patent-searchers prefer advanced functionality with a higher degree of control for managing the search parameters, while the occasional user requires ease of use.

Joho, Azzopardi, and Vanderbauwhede (2010) investigated differences in search requirements among different types of patent-users, such as patent analysts, inventors, researchers, managers, and patent-searchers within companies with respect to IR systems. Their study provides a more detailed picture of what Bonini, Ciaramella, and Corno describe (2010, pp. 20–21) – the different patent-users performing the same search task – but emphasises the importance of the user context, such as the patent role, work task, and educational aspects. Their conclusions were that patent searching is inherently interactive and that the characteristics of the search tasks are important and need to be supported.

Multilingual aspects:
Classical IR research has focused more on general search systems and therefore has neglected the characteristics and uniqueness of, for example, patent searching (queries) and patent documents. This has recently changed through the availability of test collections for dedicated evaluation purposes.

Patent IR is considered a difficult task. One of the difficulties is the vocabulary used in patent documents, with its highly specialised technical and juridical words and terms outside everyday language. Patent documents have several, different sections, such as the abstract, the description, and the claims. These sections may be written and added to a document over time, thus forming a complexity along a timeline. Furthermore, parts of a document may contain several languages.

Jochim et al. (2010) argue that query translations could be seen as query expansion since the queries are expanded with their translations (ibid., p. 58). Since the NTCIR-7 workshop (Fujii et al., 2008), one of the patent retrieval tasks has been the task of patent translation via machine translation (MT) techniques and methods.

Functionality:
Bonini et al. (2010) suggest that there is an increasing need for tools and functions that automatically facilitate patent information tasks such as patent analysis and patent monitoring. These are especially needed for a growing group of occasional users. The authors suggest improved database quality and focusing on issues such as more semantics-based solutions that may improve recall of the search process while keeping precision constant. Such semantic solutions rest on the knowledge bases of a domain model (e.g., an ontology, thesauri, or a taxonomy) and on domain-specific data (p. 36).

Another interesting issue within patent IR is what Bashir and Rauber (2009) call the 'retrievability' of every single document. They investigate whether it is possible to identify a document, through its characteristics, as being of high or low retrievability. The retrieval of documents is performed through identification of content-based features that could be used to classify a document as having higher or lower 'retrievability'. This is important in recall-oriented domains, such as the patent domain, since it is vital to have access to all available and relevant documents.

2.7 Collaborative information search

There remains no widely accepted definition of collaboration, sometimes also referred to as co-operation. As Foster (2006) correctly points out, research related to collaborative information seeking and retrieval is an interdisciplinary phenomenon including studies especially from areas such as HCI, computer-supported co-operative work (CSCW), and information science. Thus definitions of collaborative information seeking are developed from the disciplines and circumstances in which they have been used. In the present study, collaboration is specifically related to information retrieval, so we proceed from the following broad and preliminary definition of Collaborative Information Retrieval (CIR):

> CIR is an information access activity related to a specific problem-solving activity that, implicitly or explicitly, involves human beings interacting with (an)other human(s) directly and/or through texts (documents, notes, figures, etc.) as information sources in a work-task-related information seeking and retrieval process either in a specific workplace setting or in a more open community.

This is indeed a rather broad definition, as our study and its empirical observations show. Collaborative information retrieval means active and explicit retrieval of information for dealing with a specific task. Sharing information, on the other hand, is usually about sharing information already acquired. Sometime these do coincide.

While collaboration is understood as an increasingly important feature of IS&R, there actually is very little *empirical* knowledge concerning the collaborative IS&R processes within organisations or teams. An early example can be found in the work of Allen (1977), who studied the differences between the information seeking behaviour of engineers and scientists. Allen points out important aspects of the information seeking behaviour that are relevant for our study: the importance of personal contacts and discussion between engineers and that there are gatekeepers in organisations. Allen also studied patterns of communication within a small research laboratory and found a typical communication network. Such networks featured central points (persons) around which communication was centred.

Pinelli et al. (1993) discuss engineers' information seeking behaviour from within a conceptual framework. That framework assumes that, in response to, for instance, a task, specific types of data, information, and knowledge are needed. The engineer then chooses from two alternatives: create information or search existing information. When an engineer decides to seek information, there are two types of information channels available: informal or collegial networks (oral interpersonal communications

with colleagues, work with gatekeepers, and personal collections of information) and formal information systems (libraries and librarians, information specialists, and information retrieval systems).

Recent research into IS&R extends our knowledge of how people access, retrieve, and judge information. A selection from the relevant research is reviewed below.

Karamuftuoglu (1998) discusses what he calls *social informatics*, which seeks to include the relationships between humans within an IR process. This is a very valuable and important aspect. Also, Fidel and colleagues (2000) describe a project focusing on collaborative activities of members of a work team within an organisation performing IS&R tasks.

Sonnenwald and Pierce (2000) studied information behaviour in a dynamic work context involving command and control (C2) at the battalion level in the military. The phenomenon that the authors highlight they call interwoven situational awareness, which is defined as individual, intra-group, and inter-group situational awareness. Though the authors do not talk in detail explicitly about information search, their findings provide valuable insight into intra- and inter-group communication aimed at acquisition of the information needed.

Hansen and Järvelin (2000) investigated the IS&R processes carried out by patent engineers. One of the main preliminary results in this study was that the patent engineers were involved in multiple types of collaborative activities.

In the work of Herzum and Pejtersen (2000), the importance of providing support for people when they search for information is discussed. Two case studies were conducted involving engineers. The authors found that people searched for documents in order to find people and searched for people to obtain documents. Furthermore, they interacted socially to acquire information without engaging in any explicit search activity. These findings provide further knowledge that people do engage in collaborative IS&R activities.

Foster (2006) and also Reddy and Spence (2008) argue that collaborative information searching is inherently embedded in everyday work practices. Foster (2006) presents a literature review describing current research into collaboration related to seeking and retrieval tasks. The task of information seeking involves both social and collaborative approaches, while the information retrieval task involves mainly collaborative elements such as collaborative filtering and collaborative querying. Foster concludes that research in the field of CIR needs to address the conditions that influence development of systems handling collaborative information activities, such as ‘direct and indirect collaboration during information tasks’ (p. 352), which, in turn, requires a multidisciplinary approach.

Reddy and Spence (2008) conducted an ethnographic study of a multidisciplinary patient care team in an emergency department. The goal was to identify information needs within a medical team and to identify situations that trigger collaborative information seeking activities. Seven categories of information needs were identified. Furthermore, three triggers for CIS activities were identified: lack of expertise, lack of immediately accessible information, and more complex information needs.

O'Day (1993a) described four levels of sharing information in collaborative group situations: a) sharing of results with other team members, b) self-initiated broadcasting of interesting information, c) handling of search requests made by others, and d) archival of potentially useful information in group repositories for others to use. Romano et al. (1999) describe how user experiences with IR have informed the development of a system prototype for a Collaborative Information Retrieval Environment. The system prototype is dedicated to supporting collaborative information searching. Hertzum (2000) reports on how information seeking is interwoven with co-operative work and investigates the role of people as information sources during the work of a system design task. The author found that software engineers were looking for practical experience rather than hard facts and were also looking for commitments rather than information. Empirical studies of collaboration in IR among end users have been scarce, but increasing interest today is resulting in contributions such as the SearchTogether system by Morris and Horvitz (2007), a prototype that allows distance collaboration among a group of users during searches for information on the Internet. The system supports the following activities or functions: awareness, division of labour, and persistence – such as storing a search. The study was performed in an everyday life situation (travel search).

In HCI and CSCW, we find a large body of literature wherein attempts are made to facilitate finding of information through social networks (e.g., the Answer Garden approach of Ackerman & Malone, 1990). Furthermore, McDonald and Ackerman (1998) describe the 'Information Lens'. They conducted a five-month field study of how people in a medium-sized organisation find the expertise to construct, maintain, and support their software systems. The study deals mainly with how people share information through expertise identification and expertise selection.

Research in Computer Supported Collaborative Work (CSCW) addresses collaboration in organisations and work groups, and also systems supporting collaboration, such as organisational memory, organisational information handling, and information sharing. Harper and Sellen (1995) studied the professional work setting of the International Monetary Fund (IMF) and report that the reuse of documents will most often involve paper documents or 'paper-like' equivalents. Even though the study mainly deals with sharing of information, the findings are important. One interesting finding reported is that social interaction is not as important for the sharing of *objective* information as it is to the sharing of *interpreted* information. Harper and Sellen do not explicitly address the retrieval of information. Traditional *human communication* may be *asynchronous* or *synchronous* – asynchronous by post and through book/journal reading, synchronous through human face-to-face real-time communication and *ad hoc* social interactions. *Computer-mediated communication* may also be asynchronous through e-mail, searching of the Internet, and log viewing, and synchronous through videoconferencing (e.g., Erlich & Cash, 1994; Haake et al., 1999). Also, they may be *loosely* or *tightly coupled* activities (Tang et al., 2006). In *loosely coupled activities*, the system takes advantage of recommendations from other people through observations of their information seeking behaviour, such as search paths and annotations; recommendations based on usage rates; and explicitly stated recommendations. *Tightly coupled* activities may in the context of IS&R include sharing queries and strategies for their refinement, and feedback and judgement phases involving others (Haake et al., 1999).

2.8 Information use

A factor that is seldom discussed and not often part of the information seeking is *information use*. As a result of an information seeking activity, certain information is retrieved and may be used and utilised in different ways (e.g., for reaching a task resolution). The information collected may be used as a whole, in part, or in combinations in order to contribute to the accomplishment of the task (Wilson, 1973). As pointed out by Savolainen (2009), the concept of information use is still ambiguous and can refer to a) the use of information during the search process or involving judgement about relevance of information for decision-making or problem-solving or b) the use of information in an end product. Kari (2007) provides a more general description when he talks about the outcome of information search. This involved both use as a process and the effects of the process. The former implies that when a person is using the information, he or she also does something to or with the information, while an effect means that the information does something to the person.

2.9 Summary

The present study has its background in the interactive information retrieval field, with a user-centred view of information retrieval. In this chapter, we established the relevant background of concepts, theories, and models behind work on information access in work task environments that is related to the study. The platform for our study is based on the intersection of two research areas focusing on information access: information seeking and information retrieval research. More precisely, the present study is concerned with the user-oriented and interactive part of IR research. In many of the models from the IS and IR research areas, there are very few actual attempts to relate these models to each other on a theoretical basis.

Within information seeking, a set of concepts, such as information need, search strategies, and relevance assessments have been empirically tested, and the outcomes have then been built into established theories, models, and frameworks forming extended models. This development in the theory (Vakkari, 1998) shows a tendency toward incorporating work tasks and work task situations. However, one concept has been examined to a lesser extent: that of information use, which could mean either using (consuming) information or using the search outcome to create something new, such as a report in which different parts of the retrieved information may be used. Furthermore, the concept of work tasks and its relation to search tasks are sometimes not clearly described. Information seeking research places the actor in the focus, resulting in empirical description of people in different domains and situations.

In traditional IR research, the user- and cognition-oriented have made advances, especially since 1990. Ingwersen and Järvelin (2005, p. 255) list a set of achievements within user-oriented research, among them research models and theories such as the polyrepresentation and assumptions about the importance of the work task and work task situations.

Different task levels and their characteristics have been discussed, as have previous approaches and attempts to gain knowledge when performing studies in real-life work tasks and IS&R settings. Studies in real-life situations are considered to be more representative of the users' real behaviour (Kelly, 2009, p. 28). However, Kelly cites as a drawback the lack of control the researcher has over the setting, which may be problematic. Another of the achievements that one could cite is the Evaluation Package for IIR proposed by Borlund (2000) that attempted to bridge classical IR studies and user-oriented approaches on the basis of the simulated work task. Empirical studies in relevance are moving toward multi-grade relevance judgement scales (e.g., Kekäläinen & Järvelin, 2002b). However, there is still no general (or standard) way of grading relevance assessment. One reason may be that the grades have different meanings within different domains. Another important issue in IIR studies is the stage of information use. The literature shows that studies are needed on how the information retrieved is used – how the reason for retrieval of information is related to the search task.

Collaborative information access has recently gained interest, with studies recognising that collaborative information search may actually be part of, and included in, traditional IR theories and models. The characteristics, requirements, and processes of collaborative searching have not yet been fully examined and understood, and empirical studies may introduce additional aspects to both traditional IR research and interactive IR research. The notion that groups of people seek information together will have an impact on models and theories of information search.

Patent searching has, until recently, been an issue only for the patent domain itself. However, because of Web-based technology and applications, practical and research-related issues and problems within the patent domain have surfaced and gained more attention from traditional IR, interactive IR, and information seeking research. The patent domain may be a real-life work task environment in which problematic issues within IR- and IIR-related areas can be empirically investigated.

Experiments under the traditional IR approach may involve a limited set of variables, while empirical studies performed in a real-life setting involve a larger number of variables. The primary concern of the present study is with real-life IS&R processes observed in real professional work tasks.

SETTING

3

THE RESEARCH SETTING AND RESEARCH QUESTIONS

This chapter describes the overall purpose and goal of the study. We present a general overview of the study in Section 3.2, then, in Section 3.3, the research questions and the variables used.

3.1 The purpose of the study

The purpose of this study is to empirically investigate text-based information seeking and retrieval related to work tasks in the patent domain. It was anticipated that, through selection and examination of a real-world information-intensive domain and its work tasks, enhanced understanding of the information seeking and retrieval activities could be obtained. The main goal of the study is to describe the overall patent handling processes and the IS&R activities.

As explained in Section 1.3, if we consider that IS&R processes are important aspects of professional work tasks and, furthermore, that these processes ultimately need to be studied in a real-world situation, we need to pursue necessary and appropriate studies that address these aspects. We must a) describe the overall patent handling process and b) describe the IS&R processes (sessions) within patent handling. We also need to c) analyse the features characteristic of patent information retrieval, such as information need, source selection and usage, query formulation, relevance assessments, search task outcome, and search process structure, as well as d) analyse aspects of patent IR, both individually and collaboratively, and finally e) develop a methodology for analysing data of task-based patent IR studies on the basis of multiple data collection methods and exemplify its application.

The set of points mentioned above can be divided into two major research problems, one empirical (points a–d) and one that is methodological (point e). The research problems and the detailed research questions are presented in Section 3.3.

In order to deal with this problem, we construct a conceptual framework describing, in general terms, the task performance process and the factors/variables involved. The conceptual framework is both derived from a real-world setting and based on IS&R literature. We also have utilised real-life tasks and work procedures performed in real-life time frames in the development of the conceptual framework.

For the present study, we apply an exploratory and descriptive methodology, in that the study aims at describing and classifying the phenomena. A combination of qualitative and quantitative methods for data collection and analysis will be used. Chapter 4 gives a detailed description of the methodology used.

3.2 Overview of the study

As mentioned above, the study is based on real-life observations and literature reviews addressing relevant aspects of IS&R. The categories and variables used in this study are drawn from the observations as well as from the literature and form a framework in which the relationships between variables may be studied. In the present study, we use the concepts of levels, categories, variables, and attributes.

A patent work task has a task definition, a process, and its performer as well as associated activities such as source selection or relevance judgement. These patent activities are established in the literature, and we refer to them here as 'categories' associated with patent tasks. In each category is a set of variables that have been predefined and/or observed as describing the actual performance within each level of categories. The work task and the search tasks and their associated categories and variables as used in this study are described and presented in Appendix A.

The structure of Figure 3.1, below, is based on our research design, which has three related, nested levels: work task, information seeking task, and information retrieval task. The framework in the figure is a refinement of a general structure for the patent domain and derived from several sources. This refinement was based on analysis of the patent literature as well as on the interviews with the two senior patent experts and the pilot study. Figure 3.1 depicts a framework of major *levels* and *categories* with *variables* considered as relevant to our study of patent work task processes and IS&R manifestations. A final, detailed list of all variables in the categories (too many for this figure) can be found in Appendix A. Each level encompasses a set of categories of variables identified as belonging to that specific level. In turn, each category includes a set of variables that describe identified activities. Finally, each variable has two or more attributes that describe its variation. In some cases, we do not have categories, only a variable describing an activity.

At the top right in Figure 3.1 are three descriptive categories: the patent task performer, the patent application, and the patent work task. These categories correspond to the categories of variables 1a–c on the left side and consider a) the patent task (such as task type, task planning, and task constraint) and b) the task performer, with variables such as different knowledge types and completion time, related to the patent work task and patent work task process. Each of these has one or more attributes.

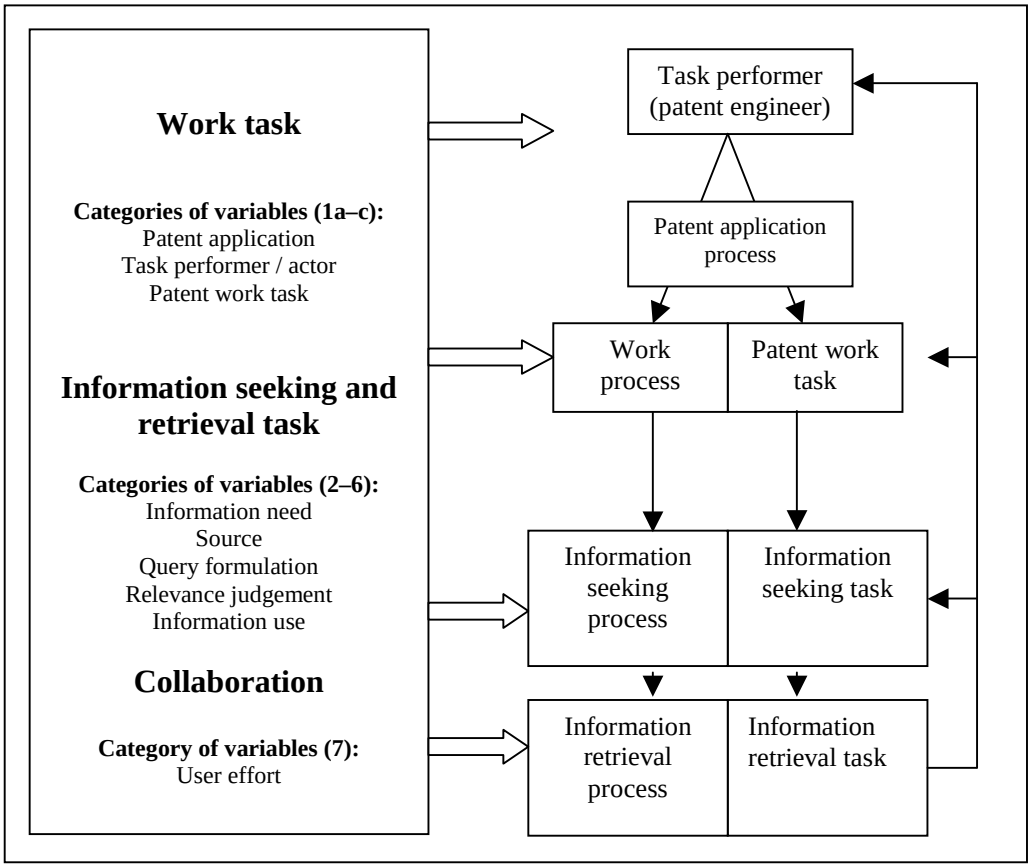


Figure 3.1: Study set-up and application of variables

Further down on the right side we have search tasks such as the information seeking task and the information retrieval task. These involve a set of interrelated categories central to the information seeking and retrieval processes, such as information need, source, and relevance judgement. These processes will be described through the categories of variables 2–6, including the category of collaboration (7), on the left.

Finally, each of these categories has a group of variables, attached to which is a set of research questions. The details for each research question can be found in Section 3.3.

3.3 Research questions

The study sets out to cover two main problems: the first research problem is empirical and the second methodological.

Problem 1 – empirical issues:
The first problem involves describing the overall patent handling process and, more specifically, IS&R session activities. We will investigate the relationships between work tasks and the IS&R task performance process. The main research question is

> *What are the effects of work task features on the information seeking and retrieval process in the patent domain?*

This main research question is divided into a set of seven sub-questions (see below). Each of these seven questions corresponds to a specific level of the work-task-based IS&R process and to a specific category of variables (Appendix A contains a full list of the variables used in the study) referring to important issues in the study. Even though the *task performer* is not the focus of our study, there are elements here that affect the study: the task knowledge of the patent engineer, subject area experience and IR knowledge, and collaboration.

The work task level
Category: Work task
Research question 1:
What are the effects of the work task features (WT, IST, and IRT) on the work task?

This problem will be studied via the variables a) type of patent applicant, b) type of patent task, c) task structuring, d) problem formulation clarity, e) user knowledge of task topic, f) perceived task difficulty, g) task constraints, and h) completion time. Data for these variables were collected through interviews, electronic diaries, observations, and post-task interviews.

The IS&R task level
Category 2: The information need
Research question 2:
What are the effects of work task features (WT, IST, and IRT) on the decomposition and formulation of the information need?

This problem will be studied via the following variables: perceived information need (2a), information need structuring (2b), information need change (2c) and deconstruction (2d), expressed information need and representation thereof (2e–2f), and information need (for information need formulation and task resolution) (2g–2h). Data for these variables were collected by means of search logs, electronic diaries, and observations.

Category: Source
Research question 3:
What are the effects of work task features (WT, IST, and IRT) on the types of sources and source content utilised?

This problem will be studied through the lens of the variables of number of sources (3a), source type (3b), and type of content (3c). Data for these variables were collected through search logs, electronic diaries, and observations.

Category: Query formulation
Research question 4:
What are the effects of work task features on query formulation?

For this research question, we will investigate the following issues: number of unique query terms (4a), types of query elements (4b), synonyms (4c), and number of terms in the query used in a query string (4d). Also considered are the numbers of combinations of query elements used in a query (4e) and the number of unique

classification codes used per task (4f). Data for these variables were collected via search logs, electronic diaries, and observations.

Category: Relevance judgement
Research question 5:
What are the effects of work task features (WT, IST, and IRT) on relevance judgement performance?

This problem will be studied through the following variables: relevance judgement application in the task performance process (TPP) stages (5a); application of RJ (as sequenced and/or aggregated) (5b); type of document elements judged for relevance (5c); and, finally, relevance judgement degree in types of patent tasks (5d). Data for these variables were collected from search logs and electronic diaries.

Category: Information use
Research question 6:
What are the effects of work task features (WT, IST, and IRT) on information use for completion of the task?

This problem will be studied via the following variables: the type of information used (6a), the types of information elements/components used, and for what the information retrieved is used (6b). Data for these variables were collected through electronic diaries and observations.

Category: Collaborative activities
Research question 7:
How are collaborative information retrieval activities manifested within and in the course of the IS&R task performance process?

The phenomena of collaborative information seeking and retrieval are studied in detail through observations of each task performance stage. Data for these variables were collected through search logs, electronic diaries, and observations.

Problem 2 – methodology:
The second research problem, as described in Chapter 1, involves the development of a methodology for analysing the data of task-based patent information retrieval studies on the basis of multiple data collection methods and exemplifying its application. Since we intend to investigate real work tasks and their characteristics, as well as features of real IS&R tasks, the data collection needs to be performed in a real-world setting. This, in turn, leads to utilisation of data collection methods that can capture these features.

This problem is addressed through capture of as many and varied data as possible that reflect the patent handling process. Qualitative and quantitative methods are utilised. In short, we a) apply methods that combine qualitative and quantitative methods, such as interviews (theme-based and expert-focused), participatory observations, electronic diaries, and database search logs, and b) propose ways of analysing data systematically.

How the data were collected for each of the variables related to the research questions described above will be reported and described in Chapter 4, especially in sections 4.4

and 4.5. Chapter 4 also includes a description of how the data were analysed (Section 4.6). Sections 4.1–4.3 contain discussion of using both qualitative and quantitative data collection methods (Section 4.1) as well as a process-related approach (Section 4.2) and provide a general outline of the data collection process (Section 4.3).

4

DATA COLLECTION AND ANALYSIS METHODS

The decision on how to collect data on-site in a real workplace has some implications for our selection of both data collection techniques and analysis methods. Our study is guided by two methodological approaches. One approach focuses on the collection and analysis of a combination of qualitative and quantitative data (see Section 4.1), and the other approach applied is a process-based one (see Section 4.2).

A field study is performed in a natural setting among professional workers or people in everyday situations. Field studies often involve the researcher being part of an organisation's social and cultural settings such as work procedures in order to understand the specific domain under investigation. In a field study, one may investigate individuals' behaviour, the behaviour of groups, and that of larger populations in communities. One way to collect data is to observe what people do. One can make direct observations of the organisation through, for example, participatory observations in order to collect data related to procedures or people's behaviours. Secondly, it is possible to analyse the material the specific organisation is working with and that may involve having at one's disposal databases, notes, reports, etc. Thirdly, one may use interviews, focus groups, or other means to collect data from the members of the organisation.

We use several data collection methods in order to gain a fuller view of the problem at hand. Section 4.3 presents and describes the data collection process and the specific methods used in the study. In that section, we describe in more detail our integrative approach to collection of data from a user-oriented point of view via qualitative methods (interviews, electronic diaries, and a questionnaire) and quantitative methods (system transaction logs). In Section 4.4, the research process and levels are outlined, including a general framework. Then, we describe the data collection process and the specific methods used in the study, also described by Hansen and Järvelin (2000), as well as the types of data collected (see Section 4.5). In Section 4.6, the analysis process and procedures are described.

4.1 Qualitative and quantitative methods

The main approaches to data collection are collection of qualitative and quantitative data. The two are usually *not* mixed together; rather, they have their own traditional foundations and procedures. Our study required combined collection and analysis of qualitative and quantitative data (e.g., Kuhlthau, 1991; Brannen et al., 1992; Strauss & Corbin, 1998; Seale, 1999), to allow description of the real-life patent information handling process. It is said that a researcher using a quantitative approach is looking through a narrow lens at a very specific set of variables while a researcher examining qualitative aspects has a much broader perspective. In qualitative data collection settings, the researchers themselves usually collect the data (e.g., when observations are used).

The distinction between qualitative and quantitative data collection methods seems to lie in the production of knowledge and the design of the research process (Brannen, 1992, p. 3). However, there is the possibility of combining research methods. Brannen (ibid.) describes several ways to use a combination of methods: a) multiple methods (triangulation between methods or within methods), b) multiple investigators (in a team or other group), multiple datasets (application of the same method at different times or of multiple methods at the same time), and d) multiple theories.

Methods of *data analysis* may also be integrated. The reason for combining several methods of analysis in our work was to capture the process of patent information task phenomena from as many viewpoints as possible: we wanted to collect a diverse, and as rich as possible, body of data so as to allow extensive analysis of the data, for us to be able to discover and unfold unforeseen behaviour or events. This diverse body of data will provide us with a more thorough view of the task performance process as well as the IS&R process to be analysed. Using the methods as complements to each other can lead to plausible conclusions about information seeking and retrieval in work tasks and processes.

Data analysis using quantitative data may support, for example, revealing and locating specific instances in the qualitative data that differ. During data analysis, quantitative data can help highlight, for example, patterns in observations. Qualitative data can support a quantitative approach with conceptual development and make data collection easier. In the analysis phase, qualitative data can support validation and interpretation.

Creswell (1998) points out that among the characteristics of qualitative research is that the research builds a complex picture and that the study is conducted in a *natural setting*. Denzin and Lincoln (1994) define qualitative research as involving several methods that take an interpretative approach to the subject. That means to study things in their natural settings. This involves a variety of empirical methods, such as interview, observations, description of encounters between people and information, and description of routine and problematic moments in people's work tasks. One possible path to follow when studying the connection between work tasks and search processes in a natural setting would then be to apply a longitudinal research design. By this, we mean a study performed over a longer span of time, observing a task from a starting point to the end. Longitudinal[15] studies are becoming more popular and

[15] Longitudinal studies are studies wherein, for example an actor/subject is followed for a longer time. How long the unit in this 'longer time' may be depends on the process. It may be one week or a year.

have been applied by Vakkari (1999, 2000a), Kuhlthau (1991), and Hansen and Järvelin (2000, 2004, 2005). These studies may unveil phenomena that are otherwise not observable. However, work tasks in different domains may be designed in different ways and the time frame for a task to be finalised may differ, making it therefore problematic to follow *in situ*.

In a qualitative study, the coding of the observations is a central activity and a tool for systematically uncovering patterns and categories. It is important to stress that coding of data is a flexible way to analyse said data. The coding of data may be either open or selective. One specific problem is related to the problem of counting: How do we deal with evidence without counting? In qualitative analysis, categorisation or coding cannot be a mechanical process (Dey, 1999, p. 146). Dey suggests that we talk about categorisation instead of coding since it is the conceptual aspect of the analysis that is important. According to Dey, when categorising, we need to reflect on how we use the categories. We may create or assign categories, then continue with exploration of the connections between the categories, and conclude by focusing on a core set of categories either as a selection of categories or through application of an integrative view of the categories (ibid., pp. 146–147).

For this thesis, we apply a procedure called *open coding*, involving categories being initially assigned (by grouping of data) in view of the phenomenon being studied. (Strauss & Corbin, 1998, p. 223; Creswell, 1998, p. 57).

4.1.1 Concerns

Since we are collecting a set of different but complementary data for analysis, *triangulation* (Seale, 1999) may be used for data collection. According to Seale, 'triangulation involves using diverse sources of data, so that one seeks out instances of a phenomenon in several different settings, at different points in time or space' (p. 54). The problem with just one observer is that there might be a bias problem, but this can be reduced through collection of different types of data for the same phenomena – e.g., via mixing of data collection methods such as interviews and observations.

One advantage of using triangulation and combining different types of data can be that one type of data can validate the other types of data collected – for example, observation and log data. Criticism of the use of triangulation in general tends to focus on the validity and reliability of these combinations, even though they are used on the same setting (ibid., pp. 56–61). However, the main focus, crucial for this study, is obtaining as rich data as possible, for *understanding* of the process and phenomena.

Another important issue in analysis of the data is the question of *generalisation*. In research based on a quantitative approach, one often chooses representative samples and gains a sense of the probabilities in order to estimate the chances of an event occurring in the population. In these cases, the sample contains a large number of cases.

In contrast, the qualitative approach usually studies a small sample of cases individually and in depth. The underlying phenomena in our study (with a random sample) could not be regarded as representative, because of self-selection by the patent engineers. It is also problematic to argue that the tasks monitored in our study would be distributed evenly across the underlying overall set of work tasks.

However, the 10 patent engineers (out of about 200 PEs) involved in our study at that time individually decided (voluntarily) to participate. They were not selected or filtered in any way by SPRO. In addition, the tasks they performed were not in any way different from their daily routines. When the participation was settled, the actual timeframe for the data collection was established by SPRO. Of course, the time of year and the specific topic area the PEs focused on might result in a certain focus in the patent applications handled. That is, the participating PEs randomly accepted the invitation but did not represent all possible topic areas within SPRO. One thing that was checked was whether the study's timeframe would interfere with critical duties. From the beginning of the study timeframe, data were collected from the first 54 work tasks. One issue that arose was that in some cases, a PE did not handle patent applications within his or her main area of focus.

The issue of *replicability* or reliability is another important factor that is frequently discussed. One can distinguish between internal and external reliability (Seale, 1999, pp. 140–157). Internal reliability concerns to what degree another researcher when applying the same approach would match the constructs of the original researcher. In order to make data analysis more reliable, a researcher might, for example, make the data recordings as concrete as possible – that is, gather data as they occur in an observation, for example, and not reconstruct the scene later on. Another way would be to make the participants themselves provide the data in written form. Furthermore, recording the data in a mechanical way, such as via an electronic diary (see subsections 4.5.6 and 4.5.7), would provide another way of enhancing reliability.

External reliability involves the issue of replication, of the whole study. Would a person studying the same setting arrive at the same findings and draw the same conclusions? Some problems of external reliability can be overcome through extensive and detailed description of the situation and the specific cases. Further, the theories, methodologies, and coding procedures used should be described as carefully as possible. However, external reliability (again, replicability of the whole study) is still very difficult to achieve with complex and unique situation and setting components. One solution to this problem is to offer a description of the study, the setting, and the methods used. As Seale (ibid.) argues, the use of low-inference descriptors in field notes and transcription, as well as systematic coding, would enhance reliability.

Finally, there is another aspect of reliability, *reliability of coding* of the data. Coding can be viewed as a sort of indexing device via which the researcher slowly moves toward something more stable and concrete (ibid.). We apply a procedure (see Section 4.6) that uses a set of stages involving data coding and re-coding for checking reliability.

4.2 The process-based approach

A *process-based*[16] approach is applied in order to capture the data needed for our research. This kind of approach is used when one is investigating phenomena over time. Usually, there is a pre-conceptualisation of the different stages or process taken under observation. In our case, we needed to use tools for data collection and analysis that

[16] A process may be defined as 'the linking of actions/interactional sequences, as they evolve over time' (Strauss & Corbin, 1990, p. 157).

monitor the work process as a sequence of stages and actions. In our study, we have defined the beginning and the end of such a process as the task of handling and resolving a specific patent application. This process includes information seeking and retrieval processes. The beginning is then defined as when the patent engineer has been assigned to a specific patent application and starts to address the content. This might, for example, be by reading the application or ordering some references in order to start the patent handling process. The end of such a process is defined as the moment when the patent engineer decides that all relevant information has been collected for making a decision on the problem at hand. This could, for example, involve writing a report to be sent to the applicant or registration of the application in the patent database.

Process-based approaches have been utilised by, for example, Kuhlthau (1991), Byström (1999), and Vakkari (2001a, 2001b). Kuhlthau (1991) focused on different stages of the information search process (ISP) in which the information seeking is viewed as a process of sense-making. The information search is seen as a process that involves thoughts, feelings, and actions. Kuhlthau (ibid.) investigated how this process developed. For monitoring of this process, data were collected longitudinally.

Vakkari (2001a) also used a process-based approach when examining the relation between students' problem stages in the course of writing their research proposals for a master's thesis. The issues investigated were the information sought by the student and the relevance assessments of the information found for this task. Also, in this case, a longitudinal study was performed. Vakkari states that the task performance process generates information searching that is the starting point of information seeking. To analyse successive searches made by a student during the process of preparing for a master's thesis, the period of observation was four months. Also Byström (1999) applied a process-based approach, when studying information seeking in public administration offices in Finland (ibid., p. 62).

In general, a process-based approach has the power to unpack complex work practices and activities that may be hidden within formal procedures and that might shed some light on why and how things are being done.

The following sections present and describe the data collection process and the specific methods used in the study (4.3), then the high-level research process stages (4.4).

4.3 Data collection: An outline

Our study set-up is based on the following reasons:

- *The work task situation*. We wanted to implement the study in a real work situation with workers performing actual work tasks that include search tasks, in order to be able to tell what actually took place in an operational workplace setting. A small (but growing) number of studies focus on people performing real-world work tasks involving real information needs. In this study, the emphasis is on the patent domain.
- *Information seeking and retrieval processes*. We wished to identify and describe the task performance process of patent engineers by monitoring and observing information handling activities on-site. Only a few studies in a *real-world* online

IR setting have sought to investigate and analyse IR systems from a task-based and information searching perspective, such as that of real information needs.
- *Information systems*. The study set out to involve operational and multiple IR systems and other information access systems. The patent domain makes use of a large set of databases and other types of sources.
- *Combining methods for data collection and analysis*. We collected both qualitative (interviews, electronic diaries, and observations) and quantitative (log statistics) data from users performing information seeking and retrieval tasks.

The project involved co-operation among the Swedish Patent and Registration Office, the University of Tampere, and the Swedish Institute of Computer Science (SICS). We wanted to perform the study in a setting that is IR-intensive – i.e., one that featured a relatively large number of information retrieval activities. We found that the patent engineers at SPRO performed information seeking and retrieval activities almost daily. Another factor that attracted us was that they were using several types of information sources. Furthermore, since we wanted to study the task process of handling patent applications, the frequency of IR activities was important. The study was designed to collect data in a real-life work situation that involves IR-intensive activities. The study is intended to be both open and holistic in its approach.

In all, 10 professional patent engineers at SPRO participated in the study. The informants were selected by SPRO, with the selection guided by two general criteria: coverage of a range of fields of technology and several years of work experience at SPRO. Since we were interested in the task performance process and, especially, the relationship of work tasks and the information seeking and retrieval processes, we needed to monitor the work task processes.

The data were collected over six weeks (May to late June 2000). One major problem with the handling of the patent applications was related to time. The processing of a normal patent application, from its arrival at the patent office to when the application is approved as a patent (or partially approved or not approved) spans 1–2 years, yet it was not feasible within the confines of our study to monitor the handling of patent applications for such a long time. In these one to two years, depending on patent task type, there are certain periods of intense activity. These times when the patent engineer or applicant handles the patent application (for example, in response to a request from the patent engineer) are sometimes followed by long pauses while one waits for replies. Especially when the intense activity involved processing of the application by the patent engineer, these periods were considered well-defined and natural subtasks to be observed. These subtasks had a ‘lifetime’ of approximately 1–6 days; therefore, we decided that they would be suitable work units for observation.

The preliminary goal was to collect description of approximately 50 patent application tasks by using electronic diaries and to observe around 10–15 subtasks physically. These tasks were monitored and observed as fully as possible. Our initial idea was that if we could focus on the 1–6-day excerpts of IS&R processes for observing and collecting data, including interviews covering patent activities preceding the observed excerpts, the excerpts could form the basis for a patent process description. Prior to the main phase of the study, a pilot study was conducted in order to validate our methods. One of the outcomes of the pilot test was the making of some changes

regarding the diary approach and construction of a keyword scheme for use later in the observation.

4.4 The research process and levels

Figure 4.1, below, gives an overview of the series of methods utilised in this study. They were used to capture as thorough data as possible. In parallel with the data collection stages, the process of analysing said data is shown along a timeline, which is divided into six weeks, representing the main period in which the study took place.

At a more detailed level, Figure 4.1 depicts three layers: the first highlights the two types of users participating in the study, and the second addresses, in detail, the components of the data collection methods used (the methods are described in depth in Section 4.5). Finally, we show the phases (1 to 5) wherein the data were analysed. Between the data collection and data analysis processes, we have added a timeline to connect these two processes and offer a sense of when the various activities occurred.

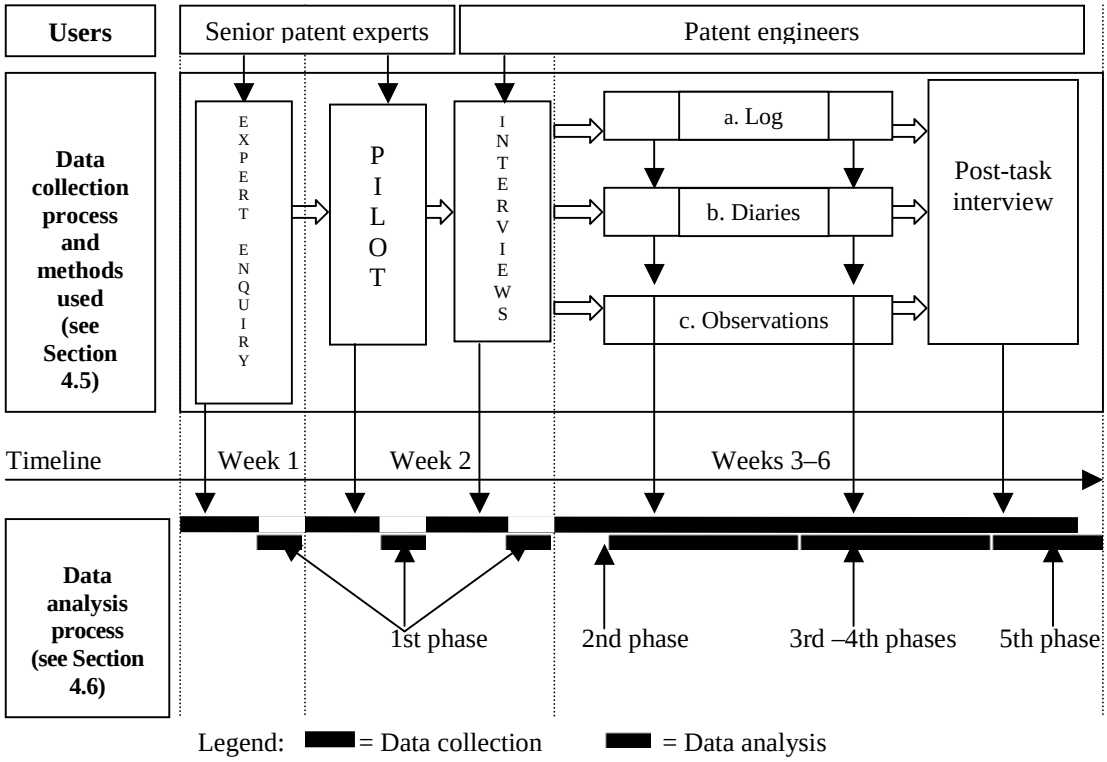


Figure 4.1: The data collection process and overview of the analysis methodology

In the first layer, senior patent experts and patent engineers were used. In the data collection layer, we performed an enquiry with senior patent experts during the first week, and, through data analysis, we acquired concepts and knowledge of patent handling procedures. In the second week, we performed a pilot study with the expert users, followed by a session of interviews with the patent engineers who would be involved in the main study. The goal with the pilot study was to test the protocol and with the interviews was to gather further aspects for consideration in the main study. At the start of the second data analysis phase, we designed the final protocol that would be used in

the main data collection process. The main data collection process was completed in weeks 3 to 6, closing with a post-task interview with all patent engineers. In parallel but with a slight delay, the third and fourth phases of data analysis were initiated and continued while, at the same time, data were collected continuously. This was because the diaries were sent back to the researcher as soon as a task was completed. Finally, the analysis of data in phases 3 and 4 actually continued after the interviews were performed.

Each specific component in Figure 4.1 will be further described in Section 4.5, below. The methods used created a large and varied set of data. The data analysis process is further described in Section 4.6.

4.5 Data collection and datasets

The following section is concerned with the methods[17] used for collecting data and the types of datasets handled in our study to allow monitoring of work task processes. Our original, fundamental research idea was to investigate a real-life information-intensive domain. This basic point of departure ruled out some methods of investigation from the start.

With respect to the issue of gathering data from a realistic situation, we needed to include real tasks performed by real users. First, we could create an experimental environment (fixed to a set point in time) to be monitored in which real work tasks could be performed by a large set of either professional or non-professional patent workers. Secondly, we could simulate (Borlund, 2000) the work tasks and show them to professional patent engineers. The first scenario provides the possibility of gathering a large group of subjects and thus allows collecting as many critical data as possible in order for us to draw significant conclusions. However, it would be difficult to create and capture real-life work tasks for this kind of set-up, as it would be almost impossible to get an intellectual property domain such as the patent domain to hand over these real-world work tasks. As for the second alternative, it would be possible to create a set of simulated patent work tasks that a group of people could perform. One problematic issue here is the time aspect. A patent task usually takes about 18 months to complete, and search sessions may be scattered over several days or even weeks. It would be difficult to anticipate exactly what the next sessions would look like, since a search session is dependent on the preceding search sessions.

In the end, we decided that neither simulated patent search tasks nor gathering a large group of (professional) participants would satisfy our initial research goal. For this reason, we decided to perform the study in a real professional setting at the actual place where the work tasks were performed. This would give us a realistic context of real work tasks. However, applying a study in a real-world environment and using the methods described does imply some problems. One such issue involves the ‘information need’ (Kelly, 2009, pp. 81–82). It is usually difficult to capture the original and ‘true’ information need. One such difficulty lies in ensuring that the

[17] On account of privacy issues, the researcher was asked to give preference to not using video recordings during the interviews and observations. There was concrete discussion of this, and the researcher decided to follow the advice of SPRO and so respect the integrity of the patent engineers and the applicants.

needs are appropriate to what is being studied and not over-engineered. Among the other difficult aspects are the level and sub-levels at which the information need should be broken down. When using real-life information needs, we may ask specifically how these needs emerge and how they are broken down, as well as how they are transformed into queries.

Below, we describe in detail the different methods used in our study, including discussion of the reason for application of each method.

4.5.1 Enquiry – senior patent experts

Since the researcher was new to the domain, we decided to undertake a pilot expert enquiry in order to gain better understanding of the domain, the setting, and people working in that domain. It was also important to synchronise with the responsible persons with respect to SPRO issues such as the time limits, the number of available people, and the researcher's intentions and demands for carrying out the study. We used a method wherein we asked two senior patent experts to describe the work of patent engineers. Both experts had extensive domain and subject experience in working within SPRO (20–25 years). They also had a very high level of knowledge about the information seeking and retrieval issues. The data were collected by us asking the experts to describe the patent domain and the characteristics of the patent document, the formal patent handling process, different types of resources, and the people involved in the handling process. The enquiry was performed very informally. During the discussion, a schematic description of the process was proposed and the experts were given an opportunity to react to it. The reason for this close co-operation regarding the study was to adapt the study to the real-life context at SPRO. At this stage, we collected written notes for our analysis.

From our point of view, it was necessary to apply an enquiry with domain experts. This was to ensure that the concepts, terminology, and general work processes were correctly understood before performance of the actual study. It was also necessary to get a preview of the work situations of the potential participants in the study, in order to know more fully the specific circumstances under which they were working. This was done to plan better for monitoring their work in a later stage of the study.

4.5.2 Pilot – verifying and testing the study's design

After the meeting with the senior patent experts, we designed a framework for how the study could be performed. This overall framework was presented to SPRO and then tested with two expert patent engineers who would not participate in the main study. The pilot was concerned with verifying the

- Overall procedure applied,
- Methods of data collection embedded in the study,
- Time required for the individual parts of the data collection, and
- Outcomes of the data collection methods applied.

Conducting a pilot study made it possible to identify problems with our data collection instruments and whether the tutorial was informative enough (Kelly, 2009), so it was necessary to include it in our platform of methods. The pilot also made it possible for the two test persons to identify problems and to assess the test framework as a whole and as constituent parts, and thus validate the approach. The two persons

taking part in the pilot were asked to give comments on the viability of the data collection. With a small number of changes, the original set-up was accepted for the main study.

Without the pilot, we might have encountered several problematic situations during the data collection, which, in turn, would have resulted in delays and inaccuracy in the data collected, among other things. Via the pilot, several things were accomplished: the procedure were validated, the pilot showed the responsible person at SPRO that the study featured the proper elements relevant for the domain, and we minimised the likelihood of having to repeat certain stages of the data collection procedures.

4.5.3 Group introduction and tutorial – patent engineers

Before we started the main study, it was important and necessary to introduce the participants (patent engineers) to the project and the data collection methods and to inform them more fully about these. The tutorial was given as a group introduction serving two purposes: providing an opportunity for a) the researcher to present the study and its goals and b) the participants to identify the others involved in the study. At the end of the session, one of the patent engineers decided to withdraw from the study because of integrity issues.

While this stage of the study did not involve data collection, it was important for another reason. We could have decided not to include it, but, as noted, during the tutorial, one participant decided not to join the study. This shows that the tutorial served a purpose: it saved us the inconvenience of having to reschedule or find a new participant in the middle of the study.

4.5.4 Interviews

We decided to perform two sets of interviews, the first one prior to the main data collection phase and involving all subjects participating in the study and the second as a post-interview, done after the main data collection phase when deemed necessary.

The pre-task interview was performed as a semi-structured interview (c.f. Appendix B) based on the categories described in Chapter 3. A set of questions was designed to collect data about demographics, experience and knowledge levels, and contextual factors and also descriptions of how the participants usually search for information, what sources they use, how they use information, etc. The interview was performed in an informal non-hierarchical way. The predefined questions served as ‘bookmarks’ upon which the interview was based. The data collected during the interview took the form of written notes and tape recordings.

A follow-up post-task interview session was held, with more open-ended questions, since it was seen as necessary to clarify and expand on specific issues. The data collection comprised adding notes to the previous data collected.

The interviews were performed in order to get individual statistics (pre-interview) and allow gaining feedback on anomalies seen in the main data collection process. The pre-interview could have been done in written form (on paper or electronically) in connection with the tutorial, with the post-task interview left out. However, since the electronic diaries (see below) and the search log files did generate some

inconsistencies and incompleteness, we needed to get feedback through the later interviews.

4.5.5 Participatory observation

In parallel, continuous on-site and real-time observation were made of patent engineers in their work with patent applications. This method involved a researcher sitting beside the participant, observing as he or she performed the work and search task. The goal was to observe the whole task performance as well as focus on particular actions and behaviours. The researcher used a protocol including a list of key questions to be asked when appropriate (see Appendix C). This protocol also structured the observations. Awareness of unexpected situations and activities was emphasised. Whenever a subject made a move relevant for the study, the researcher stepped into the scene and asked about that 'shift' in task performance. The subjects were encouraged to 'think aloud'. This means that, in parallel to the participants keeping 'diaries', the researcher moved about, in an 'ethnographic' way, at SPRO every day, observing people performing their tasks. Written notes were collected during the observations.

There has been some criticism of 'think aloud' approaches (Kelly, 2009). However, one can ask participants to 'think aloud' in ways different from actually talking about everything that is being done. This might involve just making comments now and then. Furthermore, we used a method that involved asking questions when we thought there was a need for them. The questions therefore responded to a specific action or event and we thus avoided one of the drawbacks Kelly pointed out (ibid., pp. 87–88), that of distracting and exhausting the participant. Observation is a very intensive data collection method. On the contrary, the participatory observation often resulted in a bit of discussion and clarification, though there was the possible downside of being time-consuming (when asked, the participants indicated that they did not find the discussions time-critical).

4.5.6 Electronic diaries

Diaries can capture factual data without the interference of an observer. In this study, an *electronic diary* was constructed (c.f. Appendix D), containing a set of suggested stages/steps in a proposed task process (e.g., there were questions about the starting phase and the ending phase). These stages/steps were presented as suggestions, and the goal was to collect data about the construction, performance, and ending of each subtask. Also included were a set of reminding keywords for the participant to check when taking the notes. The diary contained an empty field for logging information on online sessions too. The participant was provided with a Web-based template with a form for completion. This could be submitted through e-mail or by upload via a Web page placed at participants' disposal. The data were collected over two months. Participants were asked to send the diaries back to the investigator at the end of each working day. This ensured that the investigator could make a quick check of the content, and if there were any problems or peculiarities, the investigator could immediately go back to the PE and ask for clarification or complementary information. The diary was handled electronically. Each participant received a blank diary sheet. The participants were asked to copy the original sheet as many times as necessary during the data collection period. Each participant was given an ID key for

use in combination with the date to mark up each diary. These diaries were then sent back to the investigator. The investigator could monitor the archive for incoming diaries.

The data collected consisted of written statements and descriptions of processes as well as search logs that the individual participants created on the interaction with the various information systems during the task performance process. On the downside of using diaries was uncertainty about the level of dedication to filling in everything requested. This may, however, have been limited through careful design of the diary forms. Another possible constraint was that the predetermined sections might not cover all possible activities the patent engineer went through.

4.5.7 Construction of an electronic diary

The construction of the electronic diary allowed the subject to insert both set data items and more descriptive content. The diary's design basically followed two steps:

1. Our expectation of how the information handling process would proceed in the patent domain. The framework describing the expected process is outlined in Chapter 2 (see Figure 2.1).
2. The outcome of the
    a. expert interview, with some domain-specific changes, and
    b. piloting, with further minor additions and enhancements.

The diary was designed so as to reflect the execution of a patent handling task from the moment the patent engineer starts working with the actual patent application. The diary was also designed with its various sections and components corresponding to a real case of a patent work task. Furthermore, throughout the form, in all section and subsections, were fields asking for answers corresponding to our research questions.

Each diary was divided into two parts: section A and section B (see Appendix E). Section A required the date, the participant's name, a timestamp, the ID of the relevant patent application, the official task category, and the current stage in the patent handling process. Section B contained five subsections, and the patent engineer was asked to fill in information describing the work performed for each day and task.

Such a diary should be as open as possible so that other types of information can be inserted too, which might yield unexpected, new insights into the patent handling process. Furthermore, to support the participants in filling out the forms in an ordered and focused way, we designed the diary so as to capture detailed information for the five above-mentioned subsections:

a) One on *initiation*, which asked for information on the history of the specific patent application and how the PA was prepared
b) One on *construction* of the task, including the formulation of the problem and the information needed
c) One on the planning of the tasks, with information on the overall task and task strategy planned, as well as planning of the information seeking strategy
d) A task performance subsection, divided into two parts – Part 1 was concerned with the information seeking stage and asked for information on types of information sources and on types of information, relevance assessments, and the information need, while Part 2 dealt with the information retrieval stage

and included requests for information on sources, types of information, queries, and how relevance judgements were made

e) Finally, a subsection on completion, concerned with how the task was ended and how relevance judgements were made in relation to the final report

Also, the kind of information used for completion of the task was of interest. For each of the subsections, the participants were asked to write in a generous and fluent manner.

4.5.8 Logging of data

In connection with the diaries, we asked the subjects to share the log files from their searches if possible. In most cases, they agreed to do so and copied all of the search information into the daily diaries. The log data collected were from the participant's local machine and covered all interactions with the system, in 'client-side' logging. This method of collecting interactive search activities provides a comprehensive log allowing the possibility of annotating each action. The diary included fields for the logging information[18]. For some of the databases, it was not possible to extract this information. In most cases, we could acquire a full search history for the specific task at hand. The log statistics came from several commercial and in-house databases. Some of them came through patent databases and database hosts (Dialog[19], STN[20] – Chemical Abstract, EPODOC, PAJ[21], and INSPEC[22]) and others from patent classification and index systems (such as ECLA[23], PCT[24], IPC[25], PAJ, and US patent class definitions[26]). A source of ambiguity with logging data submitted through the diaries was that it was problematic to collect the whole set of interactions. Sometimes this was due to difficulty of extracting it from some sources, and the participants sometimes forgot to copy parts of the search set. We decided to use log files in order to be able to complement the data collected through the electronic diaries. The log files made it possible to describe a participant's actions and interactions from both a qualitative (electronic diary) and a quantitative perspective.

4.5.9 Summary of the types of data collected

For the main data collection process, we decided to use three distinct methods. To collect data for all work tasks, we used electronic diaries and also collected log files. In addition, we decided to follow 10 of these tasks through participatory observation. The log files gave us quantitative data but also qualitative data, in the form of annotations and explanation of certain searches made. The electronic diaries resulted in both quantitative and qualitative data, and the observations collected mostly qualitative data. Most importantly, these three methods resulted in a combination of qualitative and quantitative data for interactive information retrieval processes.

---

[18] Log information was handled in such a way that it did not reveal any unauthorised information.

[19] http://www.dialog.com/.
[20] http://www.stn-international.de/.
[21] http://www.jpo.go.jp/.
[22] http://www.theiet.org/publishing/inspec/.
[23] http://test.espacenet.com/ep/en/helpv3/ecla.html.
[24] http://www.wipo.int/pct/en/texts/articles/atoc.htm.
[25] http://www.wipo.int/classifications/ipc/en/.
[26] http://www.uspto.gov/go/classification/.

There is often a trade-off between having a large or a small number of participants, if you want to reach a certain level of depth and insight of single elements or the behaviour of participants. We wanted to reach this depth and therefore we decided to monitor a set of work tasks for each single participant.

One weakness with our strategy may be that we did not observe enough tasks for each individual participant for us to be able to discuss individual differences. However, we did have indicative examples of the problematic elements these patent engineers face when searching patent documents.

One may also argue that it would have been enough to use only one of these methods for data collection. However, this would have had its constraints. We could have collected only log files and then performed much statistical analysis. However, this would not have answered our research questions concerning the effects of work tasks on different stages of the IS&R process. Using only electronic diaries would have limited us both statistically (absence of log files) and qualitatively (lack of observations), and the context of the interactive IR would have been too limited to enable us to reach our original goal.

Finally, using only participatory observations for data collection might have been possible. From each observation, we could have collected both qualitative and quantitative data. However, this required that both the participants and the researcher actually have the resources for performing such complete observations – in terms of time, number of persons, and contribution of personnel and organisational resources from the patent office. It was decided that 10 work tasks including search processes in full would be observed, and SPRO contributed with the appropriate resources for this task. We wanted to have a meaningful population of participants that produced a meaningful set of work tasks that could be analysed, not so much for the sake of being statistically appropriate as to ensure that the work tasks observed would have enough depth and context to be informative in producing new insights and knowledge.

Finally, below, we show a condensed summary of the datasets collected through the various data collection methods described above. Table 4.1 shows that each of the data collection methods used resulted in extensive datasets, representing different aspects of the patent work and patent IS&R. The effort required for the analysis of the datasets was demanding. However, they also provided us with an interesting and useful foundation for investigating the real-life patent work involving IS&R processes.

Table 4.1: Summary of the quantity of data collected for analysis

Method	Type of data	Quantity
Diaries	Written notes	328 pages
Search logs	Electronic log files	2,007 pages
On-site observations	Written notes	240 hours
Notes during observations	Written notes	60 pages
Interview tape recordings	Tape recordings	18 hours
Transcribed tape recordings	Written notes	233 pages
Additional documentation	Written notes	171 pages
Total number of hours		258 hours
Total number of pages		2,799 pages

4.6 Data analysis

Data analysis was performed in an iterative procedure, on account of its nature (qualitative and quantitative) as well as the timeframe of collection. Furthermore, the data were transformed in various ways to conform to one coherent type. This was necessary to allow use of the data in an integrated matrix for explanatory analysis.

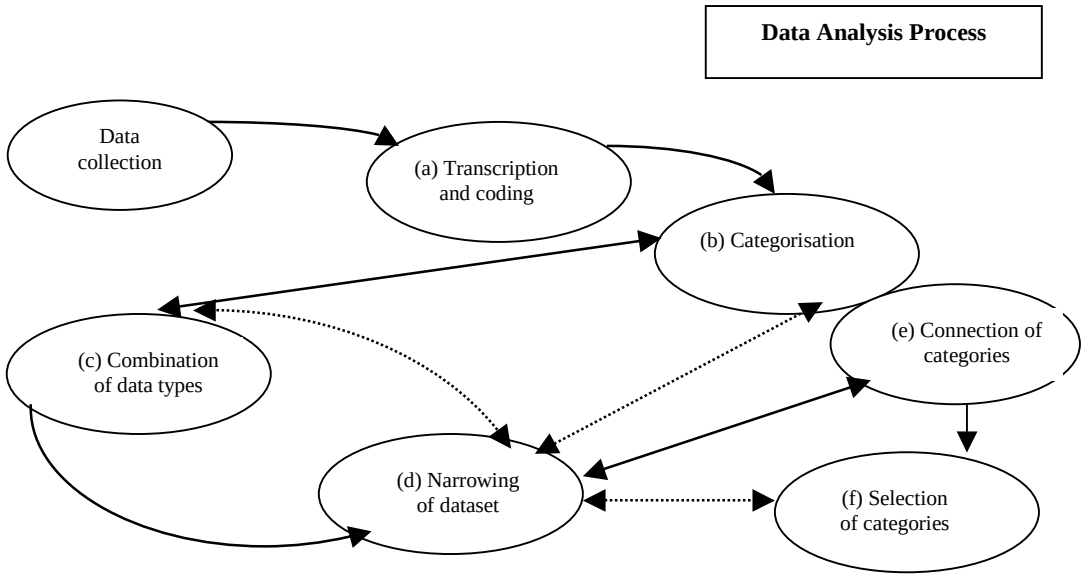


Figure 4.2: The data analysis process

a. In the transcription and coding, the data that required transcription were transcribed (from tape recordings) and then all data were coded.
b. We used an open coding method (Dey, 1999) to categorise the data. After further analysis, the original data could be re-categorised.
c. The various data types were then ordered into a protocol.
d. We applied a data and category reduction process, which was also applied after the re-categorisation phases.
e. Different categories were connected and tabulations were constructed.
f. Finally, a main set of categories was selected for further analyses.

Phase 1 – work with domain experts:
In the first analysis stage, the data from interviews with two patent experts were collected and analysed. The data were collected in written form, then a) *transcribed* and b) *coded* into a set of descriptive procedures, and categories were constructed. Based on these procedures and categories derived from the first data collection and knowledge of general information seeking and retrieval processes, a first version of a *diary* for data collection was designed, along with a protocol to be used for analysis of the data collected (phases c and d in Figure 4.2, above) was designed. The protocol was then tested in a pilot study. Additionally, each of the participating subjects was *interviewed* and written notes were collected and analysed via content analysis. The sessions were also tape-recorded and transcribed, then coded for values to be used.

Phase 2 – the design of the task protocol:
In order to obtain a workable tool for the analysis, we designed a *protocol* (Appendix E) for the description of individual tasks and for the categorisation. Data were collected from the interviews, diaries, logging, and observations. For each individual patent work task, a single protocol was assigned. This means that one protocol could include data from several diaries (one for each day of work on a specific task). The design of the protocol was guided by

a) A set of predefined variables;
b) The results of the pilot test; and
c) The categorisation of data from diaries, interviews, observations, and logging.

The protocol consists of two parts. The first part provides formal information about each individual work task observed within the study and could be regarded as a registration form. Each protocol includes an internal number that corresponds to a specific task performed by a certain patent engineer. Furthermore, each task was assigned a formal task category defined by SPRO (A, A+ITS, etc.). At the end of the protocol, the method used to collect the data was registered.

The second part of the protocol featured a set of relevant and predefined *variables* representing various categories. One example is the 'Source' category, which is represented by a set of identified variables. The variables in this category include number and types of sources used. Each variable then has its own range of attributes. In the example in Figure 4.3, below, the range of attributes is paper-based, human-based, and digitally based sources. To each attribute, a value may be assigned.

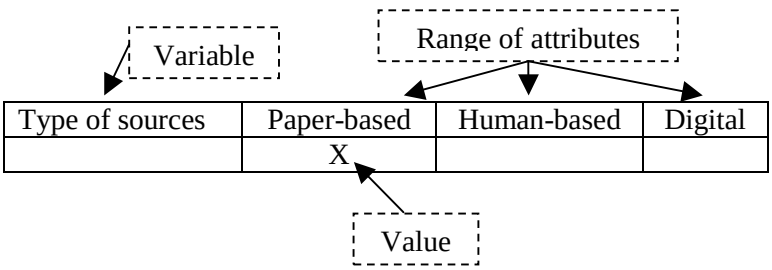


Figure 4.3: Example 6 – types of sources

Finally, we added new and important attributes to the protocol when the data allowed us to do so. The reason for this was that the data collection was done in multiple ways at different points in time, so unforeseen events could lead to incorporation of new, important aspects. Since the data collected were both quantitative and qualitative, we needed to transcribe some of the datasets into a coherent set of values. By means of content analysis, the descriptions made in different forms were transformed to single values or numerical sets so that it would be possible to compare them with other data.

In the next phase of the analysis, we deleted redundant variables and attributes. The reason for this was that no values could be gathered for these variables and attributes. We also separated the variables related to the descriptive work from the variables related to the main research questions. We identified and constructed the final set of variables (see Appendix A) for analysis and then designed the protocol accordingly.

In the next stage, we ordered the observed tasks according to the formal types of tasks defined by SPRO. We then performed content analysis of the diaries and interviews and of the extracted data and inserted them into the protocol scheme as values. We also analysed the search logs that the patent engineers submitted with their diaries. A final version of the protocol was created, including the data inserted from the observations, diaries, and log statistics. A protocol was created for each work task observed and followed, for 54 separate protocols. Also, a final list of all transcribed interviews was made. The following is an excerpt from Tape 4; Participant 4 (May 2000)[27]:

PH: When you perform an assessment, what are you focusing on? What is the important element for a positive assessment? Is it the claim, the description?

N: It is something in between, because you try to understand what the invention is about. The claim can sometimes be too fuzzy and written so it covers more than it should do.

PH: Is it the problem solving, rather than description of problem or claims, that is important? Can I find the problem solution in the documents I found when searching? Is there a relationship between the terms you used in a search and them describing the problem so that ???[28]

N: Yes

PH: Would you say that you made a relevant assessment based on how you have perceived the problem rather than how the problem was solved?

N: Yes

PH: That you match the right document to the exact relevant image, as it ???

N: If one should aim for an exact formulation – it is very seldom that two persons write exactly the same thing, so the interpretation is the problematic issue. The interpretation is usually based on other terms than what is written in the claims that I try to find in the relevant documents. In addition, you need to go back to the original patent application. That is more about judging whether the retrieved document is relevant. Then you need to make an assessment, a final

[27] The transcript is translated from Swedish. Its validity has been checked by one outside person.
[28] The use of three question marks ('???') indicates that the transcriber could not hear what was said.

assessment for real relevance. In summary, the matching process begins with the perceived understanding I have.

PH: The next step will be …?

N: Then you have created a document folder. Then I think one should go directly to what the claims say in the application.

PH: What type of information is important? We have already mentioned synonyms and classification codes. You have also mentioned that text in combination with images and the description. By 'information', I mean what sections of information are important. Is it paragraphs, the whole document, or the terms that are important in the process of matching?

N: I think it is the terms, but not the terms only – rather, the terms in their context, how the text is ???. Not the whole document. The parts of the document where these terms are located.

PH: The images for example: do they have any significant importance at all? Would you manage without them?

N: In most cases, when it is about technology, I never look at images. Sometimes I look at images and then only to understand the text better, but never to make decisions based on images.

Phase 3 – construction of a matrix:
The data collected resulted in a rich dataset that needed to be structured and handled properly. In the *third analysis stage*, we designed a large matrix (see Appendix F for an excerpt from the matrix) to carry all the data in order to get a workable overview covering all data from the 54 protocols. This enabled us to see trends and patterns. On the horizontal axis of the matrix, all the variables with attributes were listed, and on the vertical axis were all 54 observed work tasks. All data from the 54 protocols were then inserted into the cells of the matrix. Since some tasks had to be excluded from the study because of incompleteness, only tasks that qualified are presented in the matrix.

First, all relevant variables and attributes were checked so that sufficient data were collected. Owing to the nature of this study, the data collected were of different types and sometimes it was impossible to collect data for some tasks. The matrix contained both numerical and categorical values, and in some cases we needed to transform these in a coherent way for comparability. Some were of purely *numerical* nature. In other cases, the data were *categorical* (e.g., 'paper'/'human'/'digital', topic knowledge inside/outside one's knowledge domain, or 'Image'/'Paragraph'/'Abstract' /'Section'/'Reference'/'Term'/'Code'). To be presented in a matrix and compared with other attributes, some of the values needed processing into numerical form.

The result at this stage of analysis involved the final assessment of the variables connected to the research questions, also including the descriptive variables.

Phase 4 – cross-tabulation and correlation comparisons:
In the analysis described directly above, we performed analysis of individual variables in order to a) describe the phenomena and b) observe patterns of processes on the basis of the characteristics of the individual variables and their attributes. In

this stage of the analysis, we wanted to investigate further to see whether there were any relationships between variables. This would give us the possibility of examining our research questions further. For our purpose, we decided to use two types of correlation: *Spearman's rho* (Siegel & Castellan, 1988) correlation and *Yates* χ^2.

Spearman's correlation

In statistics, Spearman's rank correlation coefficient, or Spearman's rho, often denoted by r_s, is a non-parametric measure. It assesses how well an arbitrary function could describe the relationship between two variables, without saying anything about the nature of the relationship between the variables.

The procedure applied was as follows: first, we calculated all variables and mapped them to a large matrix (see Appendix F for an excerpt). Then, all correlations were grouped and categorised along the main task stages (IR, IS, and WT). The set of correlations is from a large number of variables with two to eight subclasses.

In general, for each variable and task type, there is a different set of values. Given the nature of the variable and the procedures for collecting data, different variables included different types of values, both *continuous* (e.g., minutes) and *discrete (*e.g., the abstract section or images) values (Gravetter & Wallnau, 2000, p. 25). Furthermore, the values for each variable were categorised along patent task types. In order to make comparisons between different task types, we needed to create a procedure to normalise the different types of values for each variable.

The normalisation was done in the following way (see Table 4.2): if one variable (e.g., 3b, for task length in hours) has values between 0 and 50 (hours), we reclassified these hours within a one to three intervals [H, M, L][29]. If we have a maximum value of 50, the intervals needn't be of the same length, as in H = [25, 50], M = [12, 24] and L = [0, 11]. Each interval [H, M, L] was assigned a numerical value [3, 2, 1]. A value of, say, 12.53, will be recorded in the 'M' interval with a value of 2. The limits between classes allow us to classify any value into one of three intervals. All tasks for any given type were classified into three different intervals. The number of tasks (with a task length value) was then multiplied by the values for the interval it belonged to (M = 2). In the example below, there were 13 tasks with a value, and the final product was 27.

The average task length by task type was calculated by dividing the number of tasks [13] for a certain task type [A] by the sum of task length intervals [27]. In the example below, we get the value 2.07. This procedure was repeated for each variable and task type within that variable class (see the example in Table 4.2, below).

[29] H = High; M = Medium; L = Low.

Table 4.2: Example of a table with normalised values

Hours per task

3b	A			PCT1			PCT2			AITS			AS			C		
H/M/L	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1
#	4	6	3	6	5	1	0	1	7	4	3	2	0	0	5	0	0	7
Total	12	12	3	9	7	2	0	2	7	12	6	2	0	0	5	0	0	7
tot. # / *n* = task		27/13			29/12			9/8			20/9			5/5			7/7	
avg. / task type		2.07			2.42			1.12			2.22			1.00			1.00	

Legend: H/M/L = High/medium/low; # = Numbers of values within each class limit; tot. # / n *= task = Sum of task length intervals divided by task count; avg. / task type = Average task length by interval.*

Secondly, in calculation of a correlation, six pairs were used, in general. In some cases, there were only four or five task categories present. This means that sometimes the correlations were based on four to six pairs of values. Since we encountered a large number of correlations from .700, we decided to settle with only those from .900, to avoid an overly cumbersome procedure for measuring correlations. Therefore, only correlations with a score above .900 are considered and discussed. For example, the tabulation was then done as follows:

Task type	V1	V2
A	2.07	8.64
PCT1	2.42	8.00
AITS	2.22	8.40
AS	1.00	3.66

Then we computed this correlation. Finally, because correlations between two variables are symmetric, only one correlation coefficient needs to be computed for each pair. Some 880 cross-variable tables were calculated.

χ^2 correlations

We performed a second dependency test, using Yates χ^2 for significance testing. A general definition of χ^2 is that χ^2 is based on frequencies and used in determining how well the data obtained from a test match the expected data. χ^2 is applicable both to qualitative and to quantitative variables. The goal is to test the results' statistical significance, in order to rule out results that may have been caused by chance.

Phase 5 – post-task interviews:

Finally, after scanning all incoming data from the diaries, logs from information systems, and observations, we performed a post-task interview when we thought this was necessary or when we otherwise needed to clarify an issue. The purpose was to collect additional and clarifying data to supplement the previously collected data.

Summary:

We have described the procedure for data collection and analysis as well as for connecting the procedure to a timeline. Multiple collection methods have been used, to collect both qualitative and quantitative data. The process for handling the data collected and

analysing the patent work tasks at SPRO has been described and justified in the previous chapter. The steps and handling of the data are described in Figure 4.4, below.

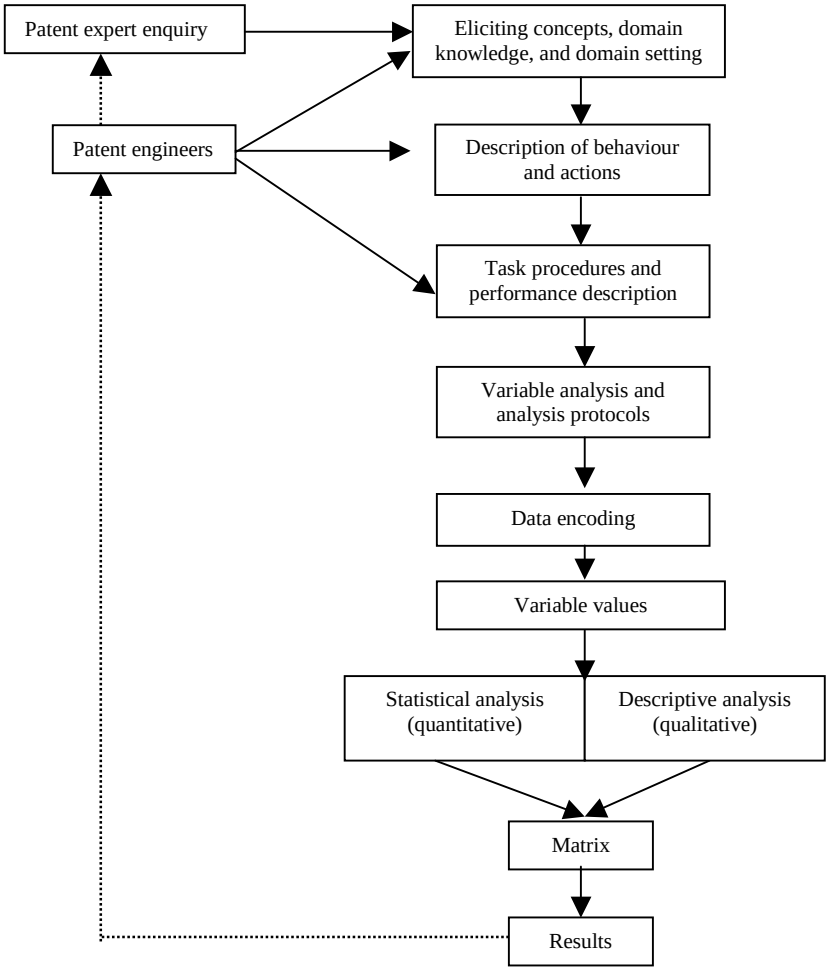


Figure 4.4: Research steps and handling of data

The figure shows that we first elicited domain-specific and IS&R-related concepts, domain knowledge, and work task procedures relevant for our study by interviewing senior patent experts at SPRO. Similar information was gathered when the patent engineers were involved, through interviews, observations, and electronic diaries.

Once the basic concepts and setting were established, we analysed the search processes and the actions performed by the patent engineers. Then, when we had a basic understanding of the work task performance, we described that task performance in a structured way. This structure provided an outline of the foundation for extracting and identifying relevant variables to be used in the analysis of our collected data. A protocol for analysing the data was constructed. The data were then encoded, and variables were assigned values. On the basis of the encoded data, we performed both statistical and descriptive analysis. The variables were then mapped to an extensive matrix containing all variables identified. That matrix was used for making of descriptive comparisons and determination of statistical correlations. From

this matrix, results were calculated and interesting findings reported. In some cases, we needed to go back to the patent engineers for clarification of some data we had collected (especially from the electronic diaries). Overall, this was needed only in conjunction with the electronic diaries.

RESULTS

The results of this thesis are presented in chapters 5 to 9 as follows:

Chapter 5: The Patent Domain
Chapter 6: Descriptive Analysis of the Work and IS&R Task Processes
Chapter 7: Cross-tabulation and Relationships
Chapter 8: Collaborative Search Activities
Chapter 9: A Method for Analysing and Describing Search Sessions in Interactive IR

5

THE PATENT DOMAIN

In this chapter, we present a general description of the patent workplace domain and a more specific description of work tasks performed at the Swedish Patent and Registration Office and the work performed. Section 5.1 describes the patent office and the general work processes performed at SPRO. In Section 5.2, the types of patent applications are presented, and Section 5.3 describes the structure of the patent document itself in detail. In Section 5.4, a set of search types used within the patent domain is presented. This is followed by Section 5.5, where the relevance criteria for judging a patent application are described. In Section 5.6, the patent classification system is detailed, and, finally, Section 5.7 presents a general conceptual model of the patent handling process.

5.1 The Swedish Patent and Registration Office

The study was conducted at the Swedish Patent and Registration Office (SPRO)[30]. This is a government agency with offices in three cities in Sweden – Stockholm, Sundsvall, and Söderhamn – and consisting of three departments: the patent department, the design and trademark department, and the marketing department.

The Patent Department, in which this study took place, is located in Stockholm. The main goal of the department is to protect investments (ideas, inventions, designs, and trademarks) that individuals and companies have made in new technological innovations and developments and to stimulate competitiveness in Sweden in a fair way. The handling of the patent applications, which is done mainly through classification, searching, retrieval, inspection, and judgement of relevant information usually (but not always) within the patent domain, ensures that each possible invention is processed properly. The Patent Department processes national and international patent applications, where a patent protects technical solutions and inventions and gives its holder exclusive rights to manufacture, sell, import, or use the invention, for example.

[30] In Swedish: PRV, Stockholm, http://www.prv.se/.

The Patent Department at SPRO consists of 12 technical units (in which, in total, about 200 patent engineers, or examiners, process patent applications); a documentation unit, which verifies the accuracy of patent applications; a patent information unit (i.e., a library) and other facilities for management of patent documents; several attorneys; and administrative staff.

The overall objectives of SPRO are to provide an efficient and appropriate system for registering intellectual property rights and to ensure that the skills, knowledge, and resources that SPRO maintains also benefit the society in an efficient and appropriate manner (SPRO Annual Overview, 2004, pp. 1, 5). In general, the patent engineers handle applications written by a professional patent bureau, those from an internal patent department at a company, and applications by private persons.

The patent engineers are placed either alone or in pairs in each office. The engineers worked in small teams and sub-units representing different topic areas, such as 'optics' or 'mobile devices'. Each team had a 'senior' engineer or mentor, usually a patent engineer with many years of experience in the patent domain, whom all patent engineers could consult whenever needed. All patent engineers have 18 months of basic SPRO course training as well as special training in searching databases. They also continuously participate in internal education. Some of the patent engineers also had completed language courses. In all, 56% had between 19 and 36 months' experience of professional patent work and 33% had 3–28 years of patent work experience.

5.2 Types of patent applications

In the patent domain, there is distinction among national and international patent application types. We encountered the following types of patent applications in our study:

National applications:
The *A application*[31] is a national patent application type wherein the applicant has not asked for any priority on account of an application in another country – that is, a new, unexamined Swedish patent application. The key outputs of an A application are a search report and argumentation. The report deals with patents that are related to the current patent application and what the area of technology for the specific invention looks like in general. Furthermore, an evaluation of the documents found is done, and if any of the *documents* found are '*against*'[32] the current patent application, this is reported. A *B application* is the same as the above-described type of patent application but is used when one has applied in another country first; when applying for a patent in Sweden, the applicant asks for priority from the first patent application date. When an A or B application is returned to SPRO by the applicant with revisions that include new claims or arguments, this patent application is treated as a *C*

[31] An 'A' application is a *national* patent application.
[32] The expression 'document against' the patent application is used when the patent officer has found a document in the patent databases that partly or fully matches the incoming application and, therefore, renders the incoming PA unable to be approved in its proposed state. Some parts need to be adjusted or the incoming application is rejected.

application[33]. When a patent application has been approved, the applicant has the opportunity to make final adjustments and smaller changes to the claims, in which case the later patent application class is *K application*.

International applications:
When an A application (national) is turned into an international application, it becomes an *A+ITS*[34] *application*.

In a *PCT application*, the applicant wants to apply for an international patent. The patent application is sent to one of the patent offices that can process PCT applications and the applicant can then apply in all relevant countries at one time. A PCT application has two phases: a PCT phase 1 report (PCT1)[35] concerns a 'novelty search' and has a international application as its basis. PCT phase 2 (PCT2)[36] involves 'evaluation of the level of patentability'. A PCT phase 2 here is a continuation of PCT phase 1, with additional argumentation. A report is issued for each phase. After this procedure, the applicant can apply for a full patent in each country, and for each application the report(s) from a PCT application will act as guiding documents for the examination in each country. In the second PCT phase, there are two procedures: A Written Opinion (WO) must be written if the examiner makes the decision that there are shortcomings in the PCT application. If a WO is not needed, the final report is the phase 2 report. A PCT phase 2 application also involves examination and assessment.

Finally, there are patent searches that are done on demand. These are called *assignments*[37]. Assignments usually are allowed 6–10 hours of investigation. One reason for making such a search from the applicant's perspective is to check the market for a certain topic or in a given area and see whether the right conditions exist for proceeding to the next stage in the development process. Each year, SPRO handles approximately 3,500 national patent applications and 6,000 international ones[38].

5.3 Patent document structure

Patent applications are highly structured documents and are composed, in general, of the following mandatory elements:

- A title page
  a. The name of the applicant
  b. A title
  c. Dates
  d. Classification code

[33] A C application is a former A application and treated separately.
[34] ITS = International Type Search Report.
[35] PCT I = *Patent Cooperation Treaty, phase 1*. The first phase is for an international patent application involving 'level of invention'. This phase occurs only once for each application.
[36] PCT II = *Patent Cooperation Treaty, phase 2*. Phase 2 is for an international patent application involving possibilities for application as a patent, searching for novelty, industrial applicability, etc.
[37] For an assignment, a search made on request, the abbreviation is 'AS'.
[38] The figures are for 2003 and are based on personal contact by telephone with Stig Edhborg of SPRO.

- Abstract
- Background
- Description
- Patent claims (see Figure 5.1)
- Drawings, figures, and chemical structures (see Figure 5.2)
- A summary (see Figure 5.3)

The *title* page contains an overview of the subject of the invention and provides information that is important when one is searching information such as the invention titles and bibliographical data. The title page allows for an initial brief assessment of the relevance of the patent document and contains bibliographic data such as the *publication number*; several types of dates, such as the *date of priority* (if the applicant has applied to another patent office); and the *classification code*.

The *abstract* contains a condensed, detailed summary of the invention as described in the description, claims, and drawings, while the background deals with what issues the invention is an attempt to resolve and how the invention improves its specific problem in comparison to other inventions in the same area. The abstract is written in English even though the original application may be written in some other language.

The *description* often contains a statement of the state of the art for the technology, a presentation of the problems the applicant claims to have solved, a short description of the invention, and a description of a set of 'embodiments' of the invention. The text in the descriptive section should be written such that a person skilled in the art could be guided to construct the invented object. The law requires that all possible aspects of the invention be described; therefore, the resulting section usually contains a description of

- The technical field to which the invention relates;
- The prior art;
- The figures and drawings;
- The problem to be solved; and
- A solution, with examples of how the invention could be implemented.

One of the most important parts of the patent application is the *claims* (see Figure 5.1). They specify different features and parts of the invention for which the applicant wants to *claim the legal protection*. The language in the claims is very formalised, for legal reasons, but also for a coherent understanding by all parties involved both at national and at international level. How the claims are formulated is of great importance and may be the reason if the application is successfully approved. Furthermore, the patent claims usually have a hierarchical structure, which generally affects the search process. The claims can be related to different forms of performance that are described in different parts of the description portion of the patent application. A claim has two parts:

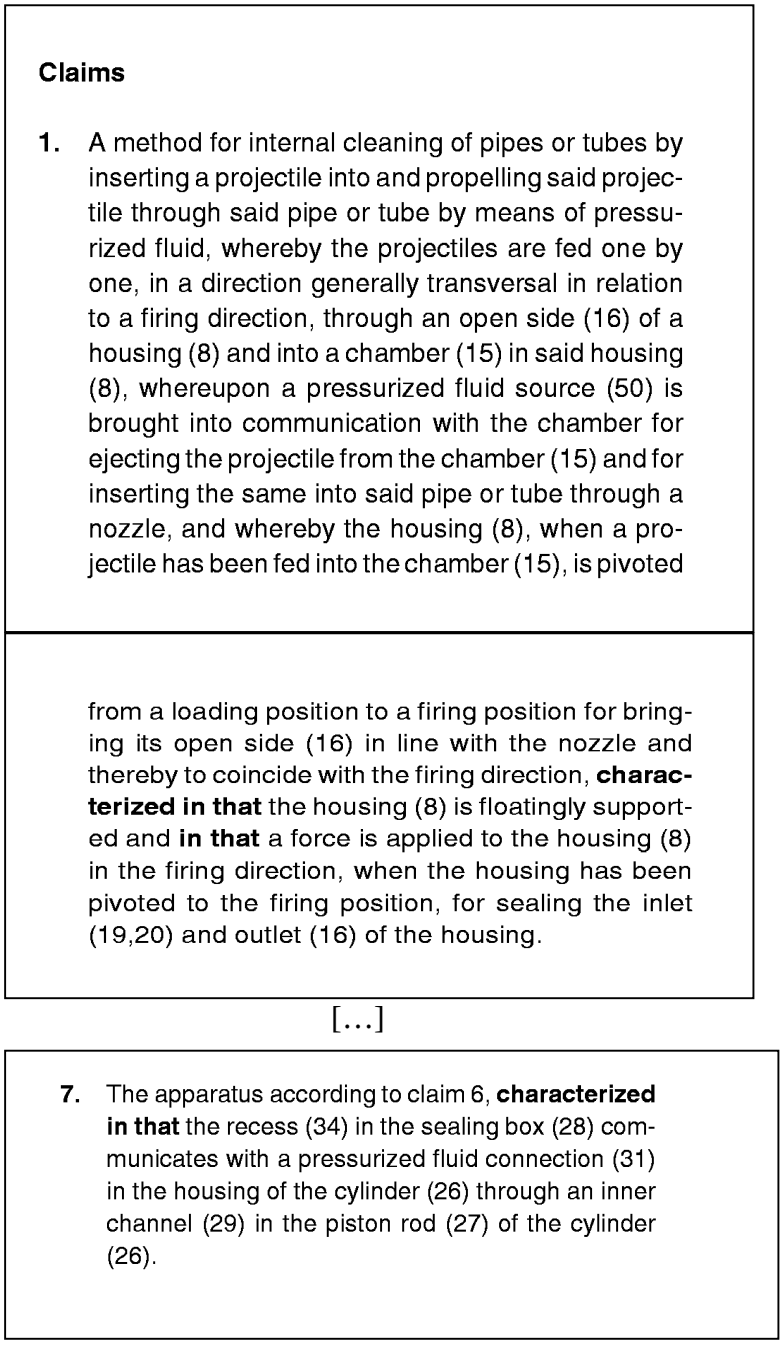
- The 'designation', which describes the 'prior art'
- A 'characterising portion', which describes the technical solution for which the applicant seeks protection

It is very common that several claims are made within one application.

There are two types of claims, the *independent* claim and the *dependent* claim:

- The statements in the independent claim stand on their own (see claim 1 in Figure 5.1).
- The statement in the dependent claim (see claim 7 in Figure 5.1) may
    - depend on a single claim or
    - depend on several claims. Every single claim then expresses different embodiments of the invention.

Each dependent claim is narrower than the independent claim upon which it depends.



Claims

1. A method for internal cleaning of pipes or tubes by inserting a projectile into and propelling said projectile through said pipe or tube by means of pressurized fluid, whereby the projectiles are fed one by one, in a direction generally transversal in relation to a firing direction, through an open side (16) of a housing (8) and into a chamber (15) in said housing (8), whereupon a pressurized fluid source (50) is brought into communication with the chamber for ejecting the projectile from the chamber (15) and for inserting the same into said pipe or tube through a nozzle, and whereby the housing (8), when a projectile has been fed into the chamber (15), is pivoted

from a loading position to a firing position for bringing its open side (16) in line with the nozzle and thereby to coincide with the firing direction, **characterized in that** the housing (8) is floatingly supported and **in that** a force is applied to the housing (8) in the firing direction, when the housing has been pivoted to the firing position, for sealing the inlet (19,20) and outlet (16) of the housing.

[...]

7. The apparatus according to claim 6, **characterized in that** the recess (34) in the sealing box (28) communicates with a pressurized fluid connection (31) in the housing of the cylinder (26) through an inner channel (29) in the piston rod (27) of the cylinder (26).

Figure 5.1: Example of two claims, from patent application SE9800621-1999 (with the permission of the Swedish Patent Office)

In almost all patent applications, we may find one or more *images* (see Figure 5.2, below). The goal of using images is to illustrate the technical details of the invention. Depending on the subject area, these images are considered very important or just an illustration of the invention. In some cases, the image is highly crucial in the assessment of whether the invention can be approved. The image gives the reader/assessor a better understanding of the idea behind the invention. Examples of image types are chemical structures, circuit diagrams, and flowcharts.

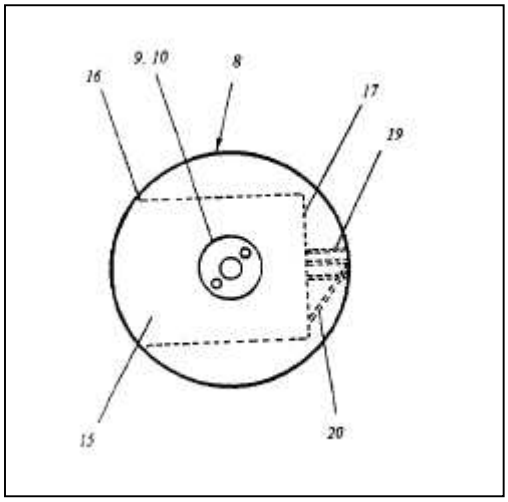


Figure 5.2: Example of an image, from patent application SE9800621-1999 (with the permission of SPRO)

Figure 5.3, below, is an example of an English *summary* from a patent application in a Swedish application with the filing data of *SE9800621 - 1999-08-28*[39] from esp@cenet[40]:

> The invention relates to a method for internal cleaning of pipes or tubes by inserting and propelling a projectile into and through respectively said pipe or tube by means of pressurized fluid, whereby the projectiles are fed one by one, in a direction generally transversal in relation to a firing direction, through an open side (16) of a housing (8) to a chamber (15) of the housing (8), whereupon a pressurized fluid source (50) is brought into communication with the chamber for discharging the projectile from the chamber (15) and for inserting the same into said tube or pipe through a nozzle, whereby the invention is characterized in that when a projectile has been fed into the chamber (15) the housing (8) is pivoted from a loading position to a firing position for bringing its open side (16) into alignment with the nozzle and thereby to coincide with the firing direction.

Figure 5.3: Example of a summary, from patent application SE9800621-1999 (in English from US6082378, in the same patent family) (with the permission of SPRO)

[39] The full reference for the patent application is (in Swedish): Title: Sätt och anordning för invändig rensning av rör eller slang; Inventor: SCHEF EDDIE, Applicant: EUROCOMP AB (SE) Publiceringsinformation: SE9800621 - 1999-08-28 IPC: B08B9/04.
[40] http://se.espacenet.com/.

5.4 General types of patent search

In connection with the patent handling process, there is a set of patent-search-related concepts, usually used within the patent domain but more recently also in academic settings describing the patent domain and search activities. Therefore, it is necessary to mention them here. The goal with some of the searches is just to test whether it would be worth the effort to write an application related to the idea in question, while other searches are concerned with the technology field and yet others are performed in different phases of the patent handling process. It must be noted that both the applicant (or a representative thereof) and professional engineers are involved in performing patent searches in general. It was not our intention to categorise the searches we encountered on the basis of these categories, since this would have required that we analyse the type of patent application itself and ask the patent engineers to point out when and where these different search types are present. Below is a categorisation of some (but not all) important search types (WIPO[41], 1998). More specifically, there are searches that can be performed by both the applicant and the patent department of the company to which the inventor belongs, if any, as well as by a patent bureau and a patent engineer within a national patent office. For example, prior art searches can be performed at different stages in handling of patent applications of different types within the patent information handling process.

Pre-application searches:
A pre-application search (PAS) is a type of search that is performed to determine whether an idea is patentable or how the general technological field looks. A pre-application search can be done by the applicant, an inventor, or a patent bureau. Academic papers, technological reports, and other public knowledge are of interest here. This is one kind of 'prior art' search performed before a patent application is filed with an official patent office. This search type is not represented in our study.

Search for the state of the art:
This is another kind of 'prior art' search. The goal of the search is to see what the state of the art is for a given technical problem. For prior art, usually all information that has been made available to the public before a certain date is relevant for the interpretation of the claims for a certain patent. So if the solution to the problem has been described in an earlier document, then the whole, or a part, of the invention is not valid. Another reason for such a search may be for assessing a specific technology that may be offered for licensing. This type of search is done by the national patent office.

Novelty search:
In a novelty search, the goal is to identify the novelty, or lack thereof, as regards the proposed solution claimed in the patent application. This search tries to decide and inform the applicant whether one should continue in developing the invention further. Usually, these searches are performed by the national patent office. A novelty search

[41] WIPO = World Intellectual Property Organization.

is problematic and difficult in that the claimed ideas are described in a very general and unspecific way. This search is sometime also called a patentability search. This search could be performed by both patent bureaux and patent departments before the patent application is filed or by patent engineers at the national patent offices during the patent handling performance.

Names search:
A name search can be done separately but is also interwoven with the overall search. The goal of this kind of search is to find documents involving specific names such as company or personal names, applicants, and investors. In our study, this search was performed by SPRO, for example, in the process of finding out whether a certain company filed applications in the same technological field that are similar to the current application.

Technological activity search:
The goal of a technological activity search is to identify companies or other organisations that develop or have considerable knowledge in specific fields of technology.

Patent family search:
A patent application usually belongs to a 'family' of applications. A family is defined by criteria such as the countries in which a specific application has been filed (published) and a list of documents with 'references cited' that would be an indicator of importance. This search may in some cases substitute for the search of patent categories and searching through classification. However, the patent applications within a family may contain references to patent documents from other technological fields or classes. These searches are frequently performed in our study at SPRO.

5.5 The patent document: Relevance aspects and criteria

When judging a document for relevance, the PE uses a very specific set of relevance criteria. Judging a patent document takes its starting point from several aspects of the document and its environment. In general, there are two important categories of relevance judgements made during the patent handling process (EPO, 2009, Part B).

Documents (references) of particular relevance:
References of particular relevance are marked 'X' or 'Y'. A category-X document is a document in view of which the claimed invention cannot be considered novel or to involve an inventive step. There may be several X-class documents for one application. Category Y is used when a document involves a claim that cannot be considered to contain an 'inventive step when the document is combined with one or more other documents of the same category' (EPO, Part B, p. 56). These aspects of relevance are summarised in the search report (Akers, 1999). There usually are two or more Y documents.

Category indicating cited documents (references) of relevant prior art:
The next main category of relevance judgement aspects (not considered to belong to either X or Y, above) is

- Category 'A': A document that is not considered to be of particular relevance and defines the general state of the art (aspects that are missing in the X and Y documents may be pointed out in an A document) and
- Category '&': Documents in the same patent family or whose contents have not been verified by the search examiner but are believed to be substantially identical to other documents the patent engineer has inspected (EPO, 2009, p. 53).

Most common is the usage of X and Y documents, and each category is treated differently but could also be combined with other categories of relevance judgements. One may say that each relevance category describes some kind of 'state' of the patent application. A patent application containing a long list of claims may end up in a long and complex combination of categories. For example, applying a strategy of using one X document plus one Y document, or a strategy of using one Y document plus another Y document, is a matter of judgement that is related to the current information need and the problem at hand. This way, each relevance degree has different *aspects* to be accounted for.

To some extent, the relevance judgement categories used by the patent office can be characterised as a 'top-down' scale beginning with the X/Y documents (see Figure 5.4, below). However, the lower the level is on the scale, the more it is used in combination with other categories in order for one to arrive at a complete assessment.

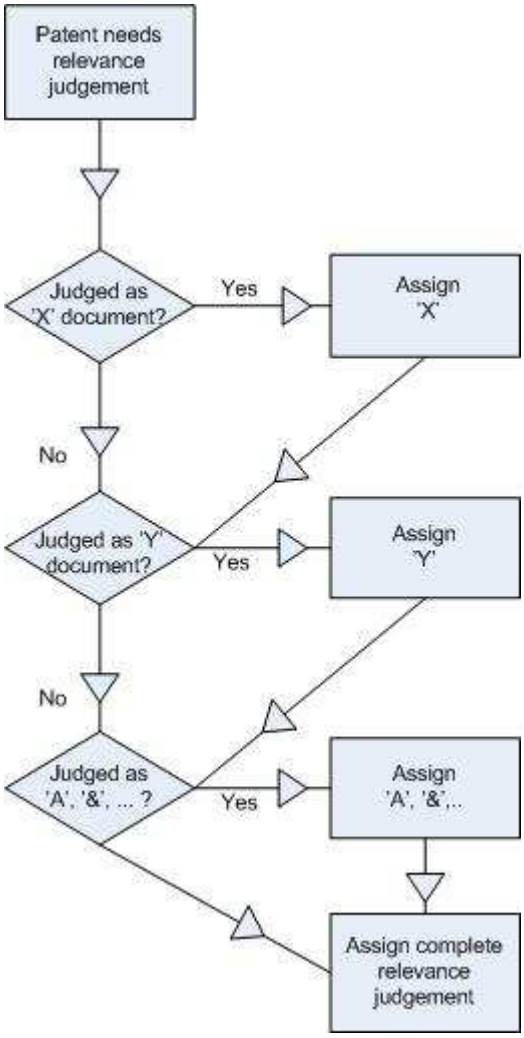


Figure 5.4: Simplified process model for the relevance judgement procedure

5.6 The patent classification system

A patent classification system is not only for uniform classification system of patent documents but also an important tool of effective search, to retrieve patent documents or related documents. One of the treaties for classification of Intellectual Properties that the World Intellectual Property Organization (WIPO) administers is the IPC system (WIPO, 2009), now in its eighth version. This IPC system has a formal hierarchy (illustrated in Figure 5.5):

a) *Section*: Eight main sections, with symbols ranging from capital letters 'A' through 'H' (e.g., D - *TEXTILES; PAPER*).
b) *Class*: Each class has a two-digit code, such as '21', used in combination with the section letter (e.g., D 21 - *PAPER-MAKING; PRODUCTION OF CELLULOSE*).
c) *Subclasses*: The third level in the hierarchy consists of one or more subclasses to each main class. Each subclass is designated with a single capital letter (e.g., D 21 F - *PAPER-MAKING MACHINES; METHODS OF PRODUCING PAPER THEREON*).
d) *Group*: Every subclass is broken down into groups, of which there are two types, *main groups* and *sub-groups*.

a. Main groups (the fourth level in the hierarchy) consist of a letter and number, followed by a slash and the code '00' (e.g., D 21 F 11/00 - *Processes*)
b. Sub-groups (a lower level related to a main group) have the same code as the main group and then an additional number counting from '01' onward after the slash (e.g., D 21 F 11/12 - *Making corrugated paper or board*)

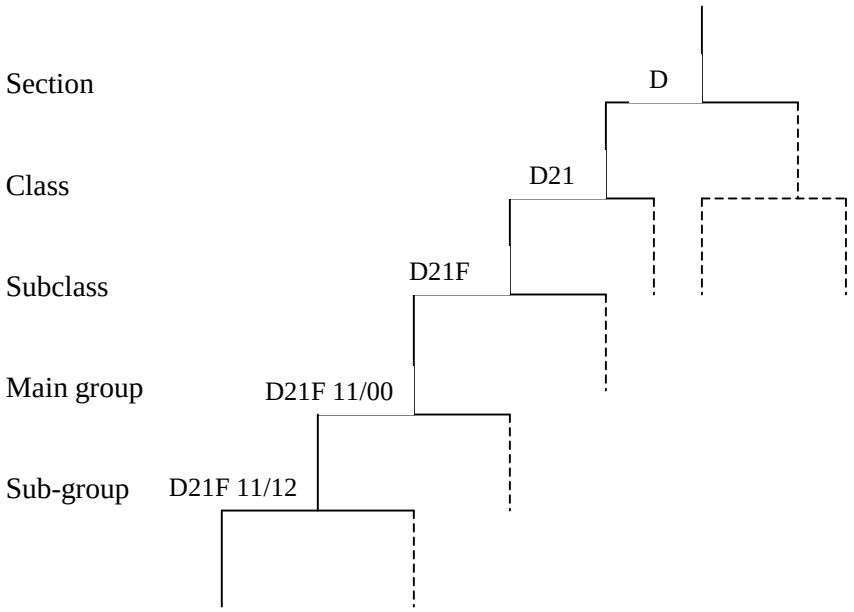


Figure 5.5: IPC classification system

Each patent document is assigned one or more appropriate IPC symbols for subject classification. This classification is highly important for the retrieval of, for instance, 'prior art' patents. One may often find a patent indexed in two or even three sections ('multi-classification'). An invention may involve, e.g., a large apparatus/system and a complex construction. Specific details of this system or construction may be classified with different codes.

Another important classification scheme used by patent engineers in Europe is the ECLA[42] classification scheme used by the European Patent Office[43]. This system also serves to facilitate prior art searches of patent documents. ECLA, commonly seen as an extension of the IPC system, has more than 135,000 subdivisions (about 60,000 more than the IPC) and is also considered more precise than the IPC system. One recommended classification search strategy is to combine queries in the ECLA/IPC fields with queries in the abstract field. Other important classification systems frequently used by the patent engineer are the United States Patent Classification system (USPC[44]) and the classification system used in Japan, JPO.

[42] http://v3.espacenet.com/eclasrch?locale=en_EP.
[43] http://www.epo.org/.
[44] http://www.uspto.gov/.

5.7 The patent application handling process: A general model

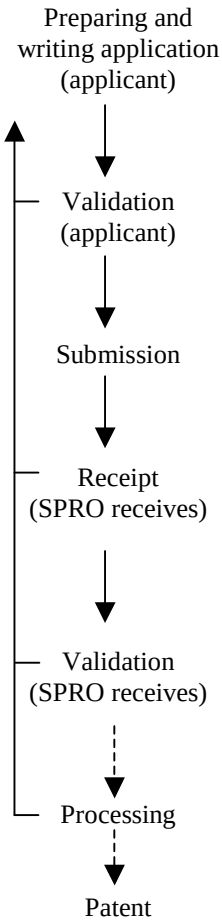
The patent handling process is a very information-intensive and focused work task, which involves extensive information seeking and retrieval activities and highly complicated problem-solving procedures. A patent is usually seen as an agreement between the applicant and the country where it was filed first. Generally, SPRO items are protected only within Sweden. If the invention is to be exploited outside Sweden, the patent must be issued in other countries too. There are two ways to seek protection outside Sweden: through the PCT[45] – the Patent Cooperation Treaty, which allows one to file an international application – and EPC[46], the European Patent Convention, which is an agreement between about 20 Western European countries. According to this convention, a patent application is submitted to the European Patent Office (EPO), which is located in Munich. In this case, a patent can be granted for all signatory nations at once. One major obstacle for the general patent handling process is the processing time for each application. Efforts to shorten the average processing time have resulted in issuing patents in an average of 2.79 years in 2004 (SPRO Annual Overview, 2004, p. 7).

Generally, the initial patent task performance process at SPRO is formally well structured and involves a certain sequence of stages

Initially, the patent application arrives at SPRO and is *registered*. The Patent Law Division first investigates the incoming patent application, in terms of whether the formal requirements for the application have been met. Before the application reaches SPRO, it must have been written by the applicant or, if written by a patent bureau, validated by the applicant. The preparation may include a so-called pre-application search. The applicant/investor needs to do this search in order to determine whether the idea is already patented. This search can also give the applicant/inventor a picture of other patents in the relevant technological field. From this point, a patent engineer processes the patent application. As noted above, the application may have been filed as either a national or an international application. There are some differences between these two types of applications (see Figure 5.6, below).

Figure 5.7, is a general conceptual model of the IS&R process, within the patent handling work task. The figure is based on a general understanding of the domain as well as on initial discussions and interviews prior to the study. The following is a general description that may apply for a national application.

[45] The Patent Cooperation Treaty – see http://www.wipo.int/pct/en/welcome.html.
[46] The European Patent Convention – see http://www.european-patent-office.org/legal/epc/.



(The application handling process starts at the Patent Office)

Figure 5.6: The data flow process in pre-processing of a patent application at SPRO

The PA undergoes a first formal *review* and a *classification process* and is then *assigned* to one of the technical divisions and to a specific patent engineer with expert knowledge in the area of the PA. An engineer responsible for a specific technical area investigates whether the invention is patentable or not (the 'novelty examination') – that is, whether certain requirements are met, such as novelty and inventiveness. The invention must be of a technical nature, have technical effect, and be reproducible, and each invention must differ significantly from that which was known before.

The *preparation* and *planning* task is then started, and the PE *reads the application* in order to *identify and define the problem(s)* and the possible approach to solving the problem(s). The patent application is a highly structured document, as noted above (see section 5.3 above for a detailed description of the document). Further information is collected and reviewed, and *requirements* for the handling process are decided. The goal is to *identify the information need* in order to perform and resolve the patent application handling process. The *information need formulation* stage involves the process of describing the identified need and specific conditions for further processing.

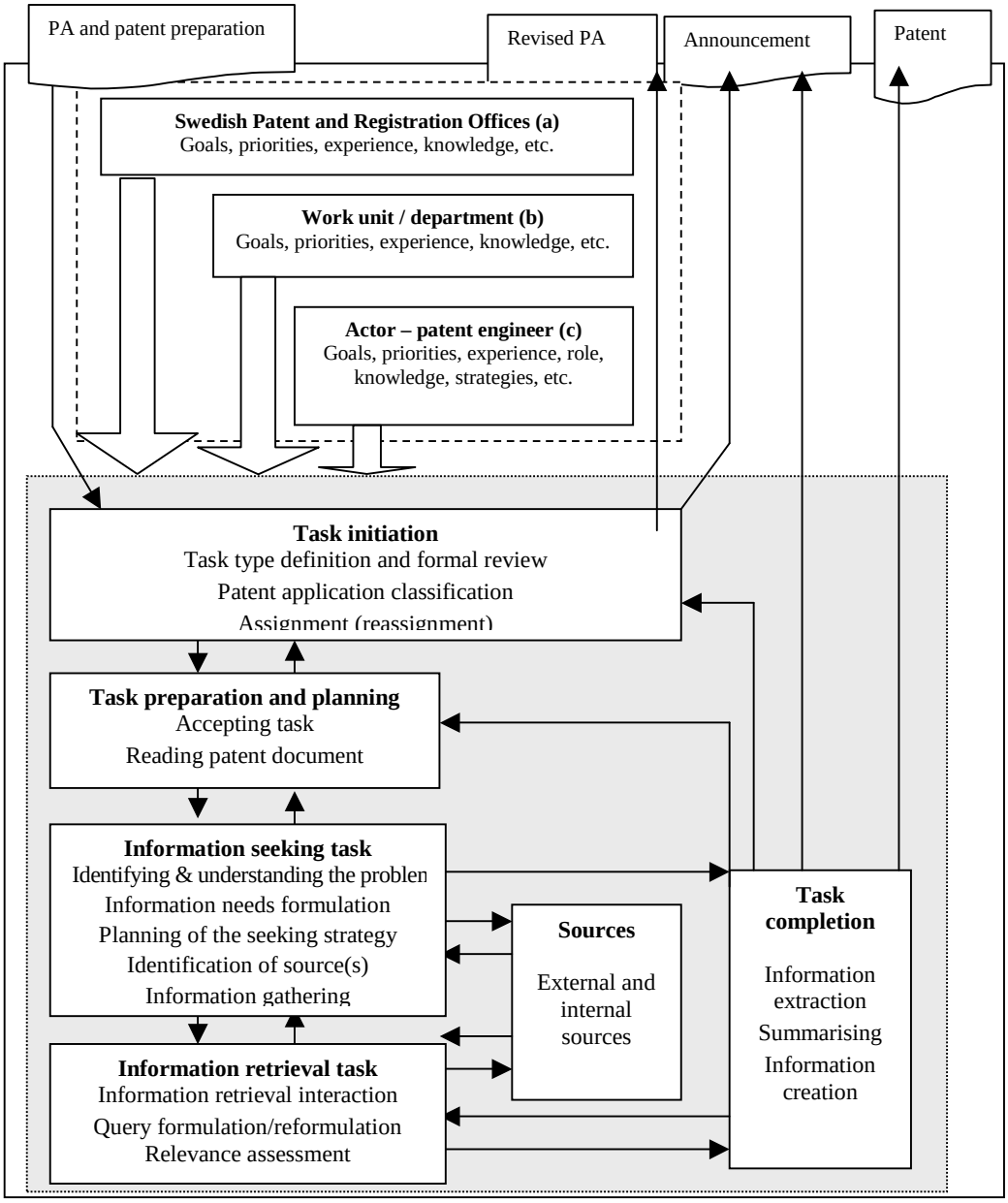


Figure 5.7: General conceptual model of the patent handling process at SPRO

The next phase is *planning the seeking strategy* and *identifying relevant and appropriate sources* to be used. When an electronic source is selected, an IR task process is initiated. For this purpose, the PE has both internal sources (databases created in-house) and external (commercial databases) ones. The IR task involves various *interactions* with *electronic IR* systems through different user interfaces, which enables feedback to the user when one performs searches in the system. *Query formulation and reformulation* involve the iterative formulation and reformulation of the identified information need through construction of query sequences.

The search outcome then undergoes *relevance assessment and judgement*. When a satisfactory set of documents has been retrieved and judged to be relevant, the next phase of the process begins, involving the use of these documents. At this stage and in

the previous ones, collaborative activities are common, for sharing knowledge and experience.

In the next stage, the retrieved documents are used in various ways as components in the *task completion stage*. From the documents retrieved, information may be extracted and summaries may be written for the reports of various types that will be sent to the applicant. Depending on the type of PA, the PA may be referred back to the *applicant for revision*. When the revised version comes back to the patent office, the PA undergoes an *inspection*.

Finally, when the applicant and the patent office have reached agreement, a *public announcement of the patent* is made and the patent is *filed nationally and internationally*. These reports are information created during, and as part of, the task process and may be used in later patent tasks. What are not explicitly shown in this description are the long time intervals between the patent handling phases. As described, there is a set of decision points at which answers to questions are required and judgement must be made of those answers before the application can be accepted.

The tasks investigated in this study are the IS&R tasks (in the shaded box in Figure 5.7 above) and start when the patent engineer receives a classified PA and ends with judgements and decisions on whether the retrieved documents, or parts thereof, should be used for creating the final report. Outside the shaded area in Figure 5.7, we identify three factors (a, b, and c) that may influence the PA process. First, there is the overall SPRO organisation that acts in line with specific goals and priorities. The second level is the unit or the work group. The patent engineers are divided among 12 technical patent departments or units that perform certain duties. The knowledge level and their collective knowledge will influence the overall performance of this group. Also, at the work unit level there is a set of specific goals that guide the members of the group. The final level is that of the actor. Each work unit member belongs to a specific team corresponding to his or her topic area of knowledge (e.g., mobile phones). The actors also have personal goals, based on their knowledge and experience of work tasks and topic. Other specific personal characteristics are personal search and assessment strategies. Each group/team member also has specific roles within the group and in the organisation.

In this study, we focus on the task level. Therefore, group- and actor-related behaviour and search patterns are not at the heart of the study.

6

DESCRIPTIVE ANALYSIS OF THE WORK AND IS&R TASK PROCESSES

In this chapter, we present the results from the analysis of the different categories of variables in view of the general framework depicted in Figure 3.1, in Section 3.2. Furthermore, each category of variables is related to one of the research question as described in Section 3.3. The chapter functions as follows:

- Sections 6.1.1–6.1.11 deal with research question 1: Work task features
- Sections 6.2.1–6.2.8 address research question 2: Information need
- Sections 6.2.9–6.2.11 deal with research question 3: Source
- Sections 6.2.12–6.2.17 deal with research question 4: Query formulation
- Sections 6.2.18–6.2.21 consider research question 5: Relevance judgement
- Sections 6.3.1–6.3.2 deal with research question 6: Information use
- Section 6.1.12 deals with research question 7: Collaboration

This chapter does not explain the relationships; rather, each variable is defined and described and is related to the patent work task with respect to the work task level, the information seeking level, and the information retrieval level (see Figure 6.1). For each variable, values are assigned and the distribution of the values is described and, in some cases, illustrated. Tables for all of the variables described in this chapter can be found in Appendix G.

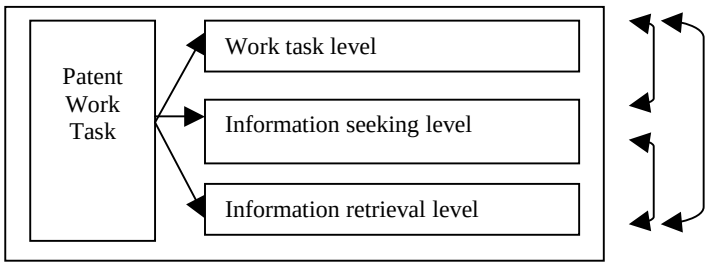


Figure 6.1: Analysis framework

This chapter covers the work task level (in Section 6.1); the task performance process stage, involving the information seeking and retrieval process (6.2); and the task completion stage – information use (6.3).

6.1 Work task performance

This section is concerned with our research question 1, on the effects of the work task features on the work task. It is addressed through sections 6.1.1 to 6.1.11. In this section, we also address research question 7, considering the manifestation of collaborative activities (in Subsection 6.1.12).

A work task starts with a task initiation phase, and, as a first step in the task performance process, the *incoming patent application* arrives at SPRO. Even though the incoming PAs have all been prepared in line with the SPRO recommendations, they encompass different characteristics, in areas such as type of applicant and type of application preparation. After arrival at SPRO, the PA undergoes a *registration* procedure at the registration office. This is an important step, since it is here that the application is given the official date stamp indicating when the patent handling process starts. Next, the patent application undergoes a first review and also is subjected to an internal *classification process*. The main subject is set for the new incoming patent application. The PA may be national or international, and there is a further set of predefined and formal *types* of patent application (see Section 5.2).

In this section, the results concerning domain goals (in Subsection 6.1.1), type of work tasks (6.1.2), type of applicant (6.1.3), and type of application preparation (6.1.4) are presented. Note that the variables of task constraints (6.1.5) and time to complete the task (6.1.6 and 6.1.7) have been placed here since we analyse them as overall work task aspects.

Results having to do with the preparation of the work task, such as perceived task difficulty or *task knowledge*, are presented (in Subsection 6.1.8), as are task structure (6.1.9), problem formulation (6.1.10), and the domain knowledge of the patent engineer (6.1.11). Most of these are closely related to the task performer – the patent engineer.

6.1.1 Domain goals

During the pilot interview (c.f. Appendix B) with the expert patent engineers and the pre-task interview with all participants, it was revealed that there exist several levels of perceived goals where the patent work tasks are concerned. The analysis of the data resulted in the categorisation of three levels of goals of patent work. We categorised (in Table 6.1) the different goals in terms of these main levels: organisational, group/team, and individual level.

Table 6.1: Categorisation of domain goals and their frequency

Domain level	Domain goals
Organisational level	a) Supporting the development and growth of Swedish industry and further development of the patent domains (7)[47] b) Providing applicants with high-quality searches and services (5) c) Protecting ideas (4) d) Helping applicants (2) e) Disseminating information and knowledge (1) f) Supplying quick answers (1)
Group/team level	a) Creating and developing praxis and consensus within the work team with regard to judgements, education, etc. (3) b) If necessary, providing information, knowledge, and protection of applied invention (2) c) Processing as many applications as possible (2)
Individual level	a) Giving each application a good qualitative judgement (6) b) Finding what is already known and therefore not accepting redundant applications or parts thereof, thus identifying patents that really are unique and therefore possible as applications (5) and c) giving the applicant strong protection for his or her ideas (5) d) Providing patent search as a service (4) e) Supporting the development and growth of Swedish industry (1)

At the organisational level, 20 comments were made, in total, and at the individual level there were 21 comments, while at the group/team level, only seven comments were made. This lack of opinions at the group/team level may be due to lack of a clear picture of the role the group may play in the patent work process within SPRO. One of the respondents said: 'This is a forgotten aspect.' One of the wishes most often expressed at this level was to build consensus on practices. There were many opinions expressed about what the goals are on an individual and organisational level. At the organisational level, three distinct aspects stood out: supporting the growth of Swedish industry, providing the applicant with a quality service (search), and providing protection for ideas. Another important goal is to provide the applicant with high-quality information about the possibilities for filing an idea. On the individual level, we find a large number of comments regarding giving the applicant a good qualitative judgement, giving the applicant strong protection, and exploring the patent space and not accepting redundant applications in whole or in part. This last comment carries a sense of a personal quest and challenge.

These goals – for example, those at the individual level – most probably have some implications. However, since the focus is not on user behaviour explicitly, this will not be investigated further in the present study.

6.1.2 Types of formal patent work tasks

In this study, when performing the observation, we investigated six *main formal* patent application types (work tasks) that the patent engineers handled. These were national (A) or international patent applications (A+ITS, PCT1, PCT2, and C). Furthermore, the third group of applications, called assignments (AS), involve a kind of 'pilot search'. These categories, or types of work tasks, were extracted from our

[47] The numbers in brackets, as in '(7)', mean that, for example, seven out of 10 professional patent examiners made comments on this particular aspect.

pre-task interview (q22, Appendix B) and used as the main categories of patent tasks in our study. Other types of patent tasks were encountered; however, most of them did not include any IS&R activities, so these were excluded from the analysis. The total number of tasks investigated was 54. These were broken down by task type as follows: 13 A tasks, 11 PCT1 tasks; nine PCT2 tasks and also nine A+ITS tasks, five AS tasks, and seven C tasks.

6.1.3 Types of patent applicants

The patent applications investigated were submitted in the following categories: from a *private* applicant or by a *company/organisation*. Private applicants may write the application themselves or leave it to an external *authorised representative* such as a patent bureau. A company/organisation may create its own application through an in-house patent department (often the choice of large companies) or by using external authorised representatives. Other applicants may, for example, be invention organisations and research institutions.

Out of the 54 applications followed, 51 could be categorised, through the electronic diaries (Appendix D), as being either privately or organisation-originated; 47 (92%) of the applications were submitted by a company or other organisation and involved task types A (8), A+ITS (9), AS (5), PCT1 (11), PCT2 (7), and C (7) , while four applications (8%) were submitted by a private person and all of the latter belonged to task type A (national task type).

6.1.4 Types of application preparation

Before being submitted to SPRO, the patent applications need to be *prepared* (which includes the applicant writing the application; doing some prior searching of the area; and pointing to previous related applications, in order to provide arguments for the application). In our study, the preparation was done mainly by companies and by patent bureaux. The preparation of a patent application might differ between applicants.

From, in total, 49 patent tasks, 13 patent applications were prepared by a company's patent department and 34 applications by a patent bureau. Data were collected through the electronic diary. Only two of the four privately submitted patent applications were prepared by the applicant him- or herself. The distribution shows two groups: Group 1 has a large number of applications prepared by bureaux and usually involves task types A, PCT2, AS, and C, while Group 2, involving PCT1 and A+ITS tasks, contains mostly company-based PA preparations (see Appendix G and Table 6.2). In four cases, the information was not sufficient for extraction of the type of application preparation.

6.1.5 Task constraints

We found that, while performing a work task, the PE encountered different types of *task constraints*. All told, 74 constraints were reported across the 54 task processes (see Appendix G and Table 6.3). The most often encountered and reported constraints were time limitations and problems related to interruptions such as visitors, internal meetings, and courses. They correspond to 47% of all constraints reported (data were

collected via the electronic diary and on-site observations). If we add the IT-related problems to these, they correspond to 67% of all constraints. It may be noted that interruptions due to visitors/colleagues coming into the room with some task to be discussed accounted for 24%. Other constraints were costs such as paying for database access per minute/hour.

6.1.6 Task completion time for observed tasks – scale of days

Using the electronic diaries, we measured the task completion time for observed tasks as a number of days and hours. It must be noted that a complete patent work task may take even 1.5–2 years before being completed and filed and that our observations cover only a small portion of this entire task duration. The 54 task units observed were grouped into two-day intervals, except for day 1 and day 2 (see Table 6.2, below, which can also be found in Appendix G, as Table 6.4).

Table 6.2: Distribution of completion times (scale of days) by number of tasks

Completion time (days)	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*	*Σ*
1	1	3	4	0	3	0	11
2	10	2	4	2	2	7	27
3–4	2	5	0	3	0	0	33
5–6	0	1	0	2	0	0	17
7–8	0	0	0	2	0	0	14
9–10	0	1	0	0	0	0	9
Total no. of days	29	37	12	40	7	14	139
Average no. of days per task	2.23	3.08	1.50	4.45	1.40	2.00	2.57
% of total days	21	27	8	29	5	10	100
n = tasks	13	12	8	9	5	7	54

Task types A, C, and PCT2 have a large number of tasks finished within two days. Task type PCT1 has a high score for 3–4 days. Task types A+ITS and PCT1 had the highest average numbers of days used per task (4.5 and 3.1, respectively). The AS and PCT2 task types had the lowest average scores (1.4 and 1.5). These findings are snapshots of IS&R within much longer tasks. The entire task from beginning to end (i.e., to a filed patent application) takes much longer. For the AS task type, the short duration is understandable since this task has an internal time limit (~4-6 hours). For a PCT2 task, this is also explained by the fact that a PCT phase 2 task is a continuation of PCT phase 1 with additional argumentation and so is processed more quickly

Table 6.3: Distribution of completion times (scale of hours) by number of tasks

Completion time (hours)	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*	*total*
1–4 (2.5)*	4 (1**)		8 (2)	3 (1)	8 (3)	12 (6)	3 (1)
5–8 (6.5)	12 (2)	22 (3)	33 (6)	6 (1)	6 (1)	5 (1)	14 (1)
9–12 (10.5)	54 (5)	12 (1)	0	9 (1)	0	0	7 (1)
13 –16 (14.5)	46 (3)	14 (1)	0	28 (2)	0	0	6 (1)
17 –26 (21.5)	44 (2)	186 (7)	0	90 (4)	0	0	13 (1)
Total hours	160	234	41	136	14	17	602
Average no. of hours/task	12.31	19.50	5.12	15.11	4.67	2.43	11.14
n tasks	13	12	8	9	4	7	54

* = Class midpoint, ** = Number of tasks

6.1.7 Task completion time for observed tasks – scale of hours

Secondly, we measured the number of hours used to complete the task (making use of the electronic diaries). In the present study, task type A (national application) has an even distribution but a peak at nine to 12 hours. PCT1 and A+ITS had a high score of 17 hours and higher, while PCT2 tasks had a high score of 1–8 hours. AS and C tasks have a low score (1–8 hours) (see Table 6.3, above, which can also be found in Appendix G, as Table 6.5). We also found that the PCT1 and A+ITS task types have the highest average scores (19.5 and 15.11 hours per task, respectively). Again, it must be noted that the observed numbers of hours apply only to the IS&R activities observed.

The task initiation is also the actual start of the work task performed by the individual patent engineer. The application is classified and *assigned* to a department and then to a specific patent engineer with knowledge in the specific area of the patent application, for handling of the task. In the present study, we do not look at the users on an individual level and their individual seeking behaviour; instead, we describe the behaviour of the patent engineers as a group, not comparing individual actors. However, some user-related aspects are presented.

After the assignment of patent applications, the patent engineer initiates *preparation* of the task. First, the PA is *reviewed* and the PE *reads the application* in order to get acquainted with the specific task at hand, its topic, and the scenario of the application. The reading phase is to *identify the problem(s)* to be addressed and to understand the claims made in the patent application. When the reading phase is completed, the patent engineer initiates the *planning and structuring* of the work task, which also involves the formulation of *requirements* for the handling process and how to proceed.

6.1.8 Perceived overall work task difficulty or task knowledge

The PE's perceived *work task difficulties* are related to the perceived level of difficulty of the work task. This involves a complex advance estimation by the PE of what needs to be done, how to do it, and what resources one needs to use in order to accomplish the task. We asked the patent engineers, before they engaged in a task, to assess the difficulty of the task at hand. We distinguished three different degrees of perceived work task difficulty as shown in Table 6.4, below (this table can also be found in Appendix G, as Table 6.6). Data were collected through the electronic diary.

Table 6.4: Distribution of perceived overall work task difficulty (task knowledge) by task type (n = 52)

Task knowledge	*A*		*PCT1*		*PCT2*		*A+ITS*		*AS*		*C*		Σ	
Easy (E)	7	58	6	54	7	78	6	67	1	25	7	100	34	65
Difficult (D)	3	25	5	46	2	22	1	11	3	75	0	0	14	27
E/D	2	17	0	0	0	0	2	22	0	0	0	0	4	8
%		100		100		100		100		100		100		100
total	12		11		9		9		4		7		52	

Tasks were perceived as easy, difficult, or both easy and difficult simultaneously. Tasks perceived as having elements of both the easy and the difficult were rare. PCT1 and AS tasks were seen as most difficult. For AS tasks (assignments), the reason can be that the application is written as a 'pilot' and is generally a short and freely written application that offers very little information for use by the patent engineer. When the work task was perceived as partially easy and partially difficult, the reason was often that the task involved two or more topics, with different level of difficulty.

6.1.9 Task structuring

The data for the task structuring variable came from the electronic diary, observations, and post-task interviews. As a first step in problem-solving performance, the PE needs to get an overview of the task. Tasks are carefully structured and planned on the basis of formal guidelines; legal considerations; effective work processes, involving the use of specific relevant sources, collection of appropriate reference material, etc.; and personal experience. Task structuring can be considered the first necessary step if one is to be able to identify the information needed on a detailed level when interpreting the actual text. Since the data analysed come from the patent engineers' own diaries, the answers are based on subjective understanding of how each task was handled. The majority of the tasks observed (50 tasks) are structured; however, in some specific cases, the planning of the task was unstructured. In Table 6.5 (which can also be found in Appendix G, as Table 6.7), two main groups of tasks can be observed.

Table 6.5: Distribution (percentage) of task structuring by task type (n = 54)

Type of task structuring	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
Structured	62	73	100	78	80	100
Unstructured	23	9	0	22	20	0
Not classifiable	15	18	0	0	0	0
%	100	100	100	100	100	100
n	13	11	9	9	5	7

Group 1 contained 20% or more unstructured tasks (23% for A, 22% for A+ITS, and 20% for AS), while Group 2 contained highly structured tasks (73% for PCT1 tasks[48]; and all PCT2 and C tasks).

6.1.10 Problem formulation

A patent application always involves a problem description and a section that describes the solution to that specific problem, proposed as a claim. We also analysed the patent engineer's ability to formulate the overall problem to be handled on the basis of the reading of the patent application at hand. We asked the patent engineer to describe (with the diary protocol) the formulation of the problem and, if possible, to categorise it as a clear or muddled formulation (see Table 6.6, which can also be found in Appendix G, as Table 6.8). Data for this variable were collected through the electronic diary, during observations, and through post-task interviews.

[48] The low score is due to a high level of unclassifiable tasks.

Table 6.6: Distribution of problem formulation clarity for information needed across task types (n = 53)

Problem formulation clarity	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
Clear problem formulation	8 (67)	9 (81)	7 (78)	3 (33)	4 (80)	7 (100)
Muddled problem formulation	3 (25)	2 (19)	2 (22)	5 (56)	1 (20)	0 (0)
Clear/muddled problem formulation	1 (8)	0 (0)	0 (0)	1 (1)	0 (0)	0 (0)
Total	12	11	9	9	5	7

The data showed that in Group 1, task type *A+ITS (33)* had a rather low degree of clear problem formulation, and Group 2 (*A (67), PCT2 (78), AS (80), PCT1 (82), and C (100)*) involved clear problem formulation. One reason the A+ITS task type had a low degree of clear problem formulation could be that part of the ITS application requires a more in-depth search.

6.1.11 The patent engineer's domain knowledge

An important factor that may affect the patent handling process is whether the topics of the patent application are fully or partly within the scope of the patent engineer's knowledge domain. We found that the problem statement and solution suggested by the applicant may encompass several topic areas. This requires the patent engineer to have knowledge in several subject areas or to consult colleagues or other sources to cover these aspects before being able to complete the task.

We classified the data into three levels of domain knowledge: within one's topic area, outside one's topic area, and partially within one's topic area. In general, most of the tasks (see Table 6.7, below, also reproduced in Appendix G, as Table 6.9) were not entirely outside the patent engineer's area of knowledge; an obvious reason for this is that the engineers have been assigned the application on the basis of their subject knowledge in a specific area.

Table 6.7: Distribution of patent engineer domain knowledge by task type (n = 54)

Engineer domain knowledge	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
A within one's topic area (a)	39	36	100	89	40	100
Outside one's topic area (b)	46	28	0	11	40	0
Combination of within and outside one's topic area (c)	15	36	0	0	20	0
%	100	100	100	100	100	100
Wholly or partially outside (b and c)	62	64	0	11	60	0
n = 54	13	11	9	9	5	7

In the A task type, 62% of the tasks were considered partially or wholly outside the PE's subject area. Tasks of type PCT1 also had a high degree (64%) of being partially or totally outside the PE's subject knowledge, with 19 out of 54 tasks (35%) requiring some degree of knowledge outside the engineer's subject area. This partial or total incompleteness or lack of domain knowledge will affect IS&R performance. For this variable, the data were collected through the electronic diary. When incompleteness in data was recognised, we collected the missing data in a later interview.

Summary of the work task:
In this section, we summarise the results related to research question 1: What were the effects of the work task features on the work task? This section focuses on important variables (described as variables 1a–1h in Section 3.3) and the values identified for them.

The type of patent task (1b in Section 3.3) was identified to six different task types for examination in our study.

The applicants (1a in Section 3.3) could be private persons or companies / other organisations. However, the applicant may not always be the one who prepared and wrote the application. The preparation of an application may be done by a private person, a patent bureau, or a patent department within a large organisation.

There are also, in real life, a series of work constraints of various types (1g) that may arise during the task performance process, such as interruption, shortage of time, and costs, among other things. The number of interruptions observed indicates that the work environment is dynamic and multitasking.

Two time-related aspects (1h) (days and hours) were observed, but they should be considered in the light of the fact that the observations were specific units, snapshots, of the full process. We identified differences between task types and number of days as well as hours. The average number of days was 1.5 to 4.45, depending on task type. The some patterns can be found in the hours spent on tasks. On average, each task took between 2.43 and 19.5 hours, depending on the task type. These differences between tasks call for some consideration when one is planning for each type of task.

Our observations revealed that a work task involves a multifaceted perspective regarding the requirements for preparing and planning the task. The procedure of handling the patent application is rather formal and follows certain steps. We found that a work task could be perceived not only as difficult or clear but also as partly difficult (1f). This means that some parts of the work task may be performed or resolved without difficulty and another portion with some degree of difficulty. This may not be reflected in work tasks judged to be difficult. When this is recognised, efforts could be focused on enhancing handling of the difficult parts.

The patent engineers also structure the work tasks in different ways (1c). This seems to some extent to be connected to the procedural aspects of the task resolution performance rather than, for example, to topical procedure.

Next, we found that the formulation of the overall problem (1d) to be solved was done in a clear way, except for one task type (A+ITS), in which the problem formulation tended to be unclear, which may be caused by the fact that this task type involves no national patent applications.

When investigating the patent engineer's topic knowledge (1e), we found that tasks might be either within or outside the PE's subject area. However, we also found a third group, including knowledge both within and partially outside the PE's topic area. This is because a task may include several topic areas, some known and some

not well known. This result is, of course, important, since this will acknowledge that we may need to differentiate between separate portions of a more complex topic.

We also recognised that the goals of the work task guide how work tasks may be interpreted and performed. When examining the goals, we could group them into three different levels: organisational, group, and individual level. On the individual level, the most important goals are to do with personal services in the form of high-quality searches and giving customers fair judgements. At the group level, more strategic issues become important, such as providing information and knowledge to the applicant, as well as protection. At this level, the patent engineers also gave voice to the necessity of working to develop praxis of performing the judgements. At the organisational level, the goals are very formal and have political and socio-economic aspects. How these levels of goals interact and affect the patent handling process is important. However, the focus we chose for the thesis is not on user behaviour and cognitive elements, though these may be a focus of future studies.

6.1.12 User effort – collaboration

The data from the on-site observations showed that collaborative information handling activities were performed during the task performance process. Nine participants performed one task each, while participants 2, 3, and 4 performed two tasks each. In total, 12 tasks were observed. Each individual task usually took 1–5 days to complete[49]. As to how collaboration activities occurred during IS&R processes, we found two main categories of collaborative activities: document- and human-related. The former means collaborative activities that are based on documents and textual information and may involve creating or (re)using documents (electronic or paper-based), such as 'working notes' that may contain information about search strategy, query terms, and classification codes. Human-related collaborative activities involve sharing knowledge between humans directly; examples are asking colleagues internally and externally for advice and expert opinions.

A mean of nearly 13 collaborative activities per task was observed (see Appendix G's Table 6.10). Both document-related (mean: 8.3 per task) and human-related (4.6 per task) activities were performed. This is a fairly high number of events. In tasks 2, 5, 6, 8, 9, and 10, in total, 15–20 collaborative events were observed, which points to a dynamic and interactive information handling process. Tasks 1, 4, 5, 6, and 8 show a 50% or greater share of document-based collaboration in comparison to the human-related, while tasks 2, 3, 9, and 11 show a minimal but significant bias toward human-related CIR events.

Summary of the collaboration:
These results are related to research question 7: How are collaborative information retrieval activities manifested within and over the course of the IS&R task performance process? We uncovered a very important aspect of the IS&R activities in that collaborative information handling activities were detected, shown as being frequent and important for task performance. Chapter 8 addresses our question 7 in more detail.

[49] Details of this methodology, analysis, and more comprehensive description of the study are presented by Hansen and Järvelin (2005) in 'Collaborative information retrieval in an information-intensive domain', in *Information Processing and Management* (IPM), 41(5), pp. 1101–1119 (Sep.).

6.2 Information seeking and retrieval task performance

After the initial structuring and planning, the PE starts the *preparations* for the *information seeking task performance* process. The first decision might be to *collect complementary information*. For example, the PA contains many references to other patents, research articles, state-of-the-art documents, books, images and figures, etc. to which the applicant relates the present application and its claims as motivation for them. Most of these documents are necessary to shape the content of the problem at hand.

When the information is collected, the PE reads the documents to enhance his or her *understanding of the problem* and problem area and to identify the problem(s) and claims made. The goal is to *identify the information need* in order to resolve the patent application task. The *information need formulation* stage involves the process of describing the identified need and specific conditions for further processing. This also includes the *representation* of the information need. The next stage is to *identify relevant and appropriate sources* for use.

This phase involves a set of categories of variables related to the two main stages in the task performance phase: information seeking and information retrieval. These categories of variables include information need (8 variables), source selection (3), query formulation (6), and relevance judgement (4). Each category will be described in detail below.

Information need

The category of information need proved to include complex aspects of IS&R. We recorded eight separate aspects during the patent handling performance. This section is concerned with our research question 2, involving the effects of the work task features on the information need utilised. It is addressed through subsections 6.2.1 to 6.2.8 as follows: the perceived and planned information need (6.2.1 and 6.2.2, respectively), the aspect of information need change (6.2.3), and the decomposition of the information need (6.3.4). Other variables were expression of the information need as single or multiple needs (6.2.5) or in narrative form (6.2.6); PA document components needed for information need formulation (6.2.7); and, finally, the type of information needed (6.2.8).

6.2.1 Perceived information need

The aspect of perceived information need is related to the level of clarity and structure of the perceived information need in resolution of the task – that is, what information to search for. The data for this variable were collected through electronic diaries and complementary post-task interviews. The perceived information need was categorised in terms of the following two binary variables: structured/unstructured and clear/unclear (see Appendix G's Table 6.11). Thus variables were divided into four value pairs: *structured/clear* (sc), *structured/unclear* (su), *unstructured/clear* (uc), and *unstructured/unclear* (uu). These pairs were grouped into two sets: a) perceived information need that was clear and structured (sc) and b) a second set, containing all statements involving some kind of unclear or unstructured perceived information need.

The first group, with its three task types (A+ITS, AS, and A), involved a rather low level of clarity. The second group (task types PCT1, PCT2, and C) had the opposite distribution: a high percentage (79% of all structured/clear tasks) of structured and clear perceived information needs. The reason for this could be that an assigned task may be somewhat 'muddled' in its characteristics and not very focused or well described. In the case of the A+ITS tasks, it might be because the applications are internationally filed, which may mean that the initial part of the patent handling process may have been performed at SPRO, rendering it potentially problematic to 'catch up with' the problem-solving task.

6.2.2 Planning related to information needs

The next information need variable has to do with the information need structuring. Through the diaries from the patent engineers, we categorised the statements about what the structuring of information need was about, according to a) the purpose in grouping of topics and selection of sources and b) more procedural issues (see Appendix G's Table 6.12). Task types PCT1, PCT2, and C had a high level of structured planning activities, while A and A+ITS had a very high level of unstructured planning of the information need. In general, this result means that there was not a great difference between what the engineers anticipated and what they actually did. The reason may be that the work tasks in general follow a routine procedure at that stage.

6.2.3 Change in information needs

When describing their search activities, patent engineers occasionally reported that their information need had changed during the process. As can be seen in Table 6.8, below, changes in information need were frequently made in two task types (A and A+ITS) while in four types of tasks there was no information need change at all (see also Appendix G, where these data are presented as Table 6.13).

Table 6.8: Distribution of change of information need by task type (n = 47)

Information need change	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
Change	38	0	0	75	0	0
No change	62	100	100	25	100	100
n = tasks	13	11	9	8	5	1
%	100	100	100	100	100	100

Most of the changes were found with A+ITS tasks (75%) in the course of the task process. In total, 11 tasks, or 23% of all tasks observed, featured information need changes. The stability of the information need may depend on a complex and multifaceted problem/topic to be solved. The observation here shows that the searchers clearly make changes to their information need during the process in real life. This should be considered in interpretation of the results of experimental research in the IS&R field.

6.2.4 Decomposition of information needs

It was observed rather frequently that the patent engineers often *broke down* their information needs. This is discussed next. In this study, we have identified two types of information need decomposition: one is related to such elements as the structure of the information need (terms, classification codes, and topic(s)), and the other is related to the overall search task process (e.g., the performance of the task at hand). The data for this variable were collected primarily through the electronic diary and log files.

In the former case, the decomposition can be expressed as pre-designing several subsets of search sessions containing specific combinations or separations of topics and terms/codes (example: 'For component X, I will use the following keywords and classification codes with distance operators and truncation […] and for component Y I will only use classification codes').

Information need decomposition related to overall search task performance refers to the planning and structuring of the search process and may involve the order in which to execute certain searches, in what order certain sources should be used, etc. (example: 'First I will read all the previous documentation and then maybe ask colleague X since she is an expert regarding component Y. Then I will search for the X component in sources 1+2 and then I will search for the Y component in sources 1+3'.). In Table 6.9, below (also in Appendix G, as Table 6.14), we can see that the distribution of information need decomposition was related to the specific task type in question.

Table 6.9: Distribution of information need decomposition by task type (n = 36)

Information need decomposition	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
% of information need decompositions related to overall search task performance	100*	27	0	50	50	0
n = tasks	12	11	2	8	2	1
% of information need decompositions related to the information need	100	100	100	100	100	100
n = tasks	12	11	2	8	2	1

Legend: * = Percentage

In all tasks, the information need was itemised in least one of the two observed alternatives. All tasks were broken down in relation to structured information need. In relation to search task performance, this was observed mainly in the A task (100%) and A+ITS task (50%). In the PCT2 task, no decomposition related to the search task was found.

6.2.5 Expressed information need as single or multiple needs

We found (through electronic diaries) that PEs' expressed information needs could be categorised as a) a single expressed need or b) a set of (multiple) needs (see Table 6.10, below, also found in Appendix G, as Table 6.15). Multiple information needs were either stated in the beginning or expressed as a sequence during the information handling process. When multiple information needs were expressed in the beginning,

they were related to different identified aspects and problem areas of the application that needed to be resolved. When multiple information needs were expressed sequentially, that was usually either a result of a newly identified problematic aspect or a consequence of the resolution of the previous information need. The expression of multiple information needs could be viewed as one aspect of the complexity of the task.

Table 6.10: Distribution of information need expression across task type in terms of single or multiple information needs stated (n = 44)

Expressed information need as single or multiple need	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
Single	0	30	0	11	0	100
Multiple	100	70	100	89	100	0
n = tasks	13	10	2	9	3	7

Only a few tasks included just a single information need. All but one task type (the C task) featured information needs as multiple needs. This is a very interesting result, as not only are there several information needs expressed within a single search task but also, when these different information needs have been satisfied, their outcome must be integrated and merged into a final solution of the problem.

6.2.6 Expressed information need as a narrative

The information need was described in two ways in the electronic diary during the task performance process: in a narrative manner and in non-narrative form. By 'narrative', we mean a fluent description containing key words and concepts integrated in one or more sentences. These sentences could describe the overall problem and sub-problems that need attention. The non-narrative expression is basically elements such as single terms and codes not necessarily connected to each other or to any descriptive context.

The only task type that showed high use of a non-narrative way of describing the information need was the A+ITS task. This implies that there is a need to embed the keywords, facts, and concepts within a larger context in order to be understandable.

6.2.7 PA document components needed for formulation of the information need

To make a satisfactory formulation of the information need, the patent engineer must make use of the information contained in the patent application (see Appendix G's Table 6.16). The data for this variable were collected through electronic diaries. In all of the cases observed, the whole PA was used throughout the process of reading the application. In the next step, specific parts of the document are considered, depending on the application. The components pointed out as important were sections (e.g., abstract, description, and claims), references, terms, classification codes, and images.

For this variable, internal analysis (within all six types of tasks) was done, since the variable covered several values and each task could include one or more of these values. In task type PCT1, all components except the reference component are important for the formulation of the information need, and in task type PCT2, terms

were not important. The classification code component is the component used most in five of the six task types.

6.2.8 Type of information needed

Next, during the observations and through the electronic diaries, we found that there were three types of information needed. The answer may essentially be found, firstly, in the text (T); secondly, in an image (I); or, thirdly, both as text and in images (T/I).

Depending on a) the type of application and b) its topic, as well as c) the type of problem to be solved, the information needed could be found in either text or images, or both. On the basis of our limited units of observation of the entire patent task, we may identify one group of tasks that involved a high requirement of both text and image elements (namely, PCT1, PCT2, and AS). For the A task type, text was important, while task type A+ITS had an even distribution among all of the various alternatives of required information types.

Summary of the information need:
In this section, we present important findings related to research question 2: What are the effects of the work task features on the decomposition and formulation of the information need? Careful structuring (2b in Section 3.3) of the task and consideration of the types of information needed are performed.

The perceived information need (2a in Section 3.3) can be both structured and clear or unclear plus unstructured, as well as combinations thereof. This finding is related to formulation of the problem (1d in Section 3.3) and shows that several degrees of perceived information need are to be considered and, therefore, this may also have an effect on overall task performance.

Most interestingly, we observed that there were quite a large number of changes (2c) of the information need during the course of the search task: 23% of the searches featured information need changes. That the information need actually can change is seldom accounted for in IR experiments.

Another very interesting finding was that the information need was itemised (2d). Further to that, we found that the patent engineers performed different types of decomposition. One was related to a) the overall search task performance and the other b) to the information need itself.

We also noticed that the information need could be expressed as either a single need or multiple needs (2e). The information need could also be stated in the beginning or as part of the ongoing information handling process. Therefore, we can state that in some situations, there is an *evolving* information need. This is clearly a situation that deviates from the traditional viewpoint from which the information need is stated as a stable entity at the start and remains stable throughout the search session. Within the patent domain, we found that the information need could be expressed as single terms or codes (2g) or as a narrative (a description) (2h).

We also recognised that the patent engineers' requirements, in the type of information (text, image, or a combination) they needed to satisfy their information need, varied (2g).

Source selection:
In this section, we are concerned with our research question 3, on the effects of the work task features on the source utilised. It will be addressed from Subsection 6.2.9 to Subsection 6.2.11. The information retrieval process begins, in this case, when an electronic source is selected and accessed. For this purpose, the patent engineer utilises both internal/in-house-created databases and external sources such as commercial patent databases (and other commercial sources – e.g., STN and Dialog).

Three elements were observed and analysed: the numbers of sources selected (see Subsection 6.2.9) source type(s) (6.2.10), and source content type(s) (6.2.11). The data for these variables were collected through the descriptive fields in the electronic diaries and through the attached log files.

6.2.9 Number of sources selected

After the first preparations, the patent engineer starts to select a source. We observed that usually several sources were used. We may distinguish two patterns: one group involving a low number of sources selected (PCT2 and C) and a second group showing a high number (see Appendix G's Table 6.17). Task types A and A+ITS showed an especially high number of sources used. For example, the patent engineers use 5–6 information sources in order to solve the problem.

6.2.10 Source types and their combination

In their IS&R task performance, the patent engineers used three types of information sources: paper-based (P), such as lexicons and dictionaries; human sources (H); and electronic information sources (E), such as IR systems and classification code schemata (see Table 6.11, below, also in Appendix G, as Table 6.18).

Table 6.11: Distribution of source type selection by task type (n = 54)

Source type	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*	
P			3			7	10
H							0
E	3	3	5	1	1		13
PH			1				1
HE	2						2
PE	2	2		3	2		9
PHE	6	6		5	2		19
Total/avg. source types per task	29/2.23	25/2.27	10/1.11	22/2.44	11/2.20	7/1.00	
# tasks	13	11	9	9	5	7	54

Legend: P = Paper-based source; H = Human source; E = Electronic source

These may be used in combination or as individual sources. When two source types are indicated (e.g., 'PH' or 'PE'), this means that two source types were used in the task, and 'PHE' denotes that all three source types were used in a single task. Each

source type is counted as '1'. So, for example, in the case of the A task type, the sum 29 is the result of 3 x 1(E) + 2 x 2 (HE) + 2 x 2 (PE) + 3 x 6 (PHE).

The mainstream IR viewpoint usually only takes into consideration the use of one (electronic) source. However, real-life searching proves this to be invalid when electronic but also human and paper-based sources are available in the information seeking situation. Task types C and PCT2 show little or no combination of source types, while A, PCT1, A+ITS, and AS have a higher number of combinations.

6.2.11 Source content type

The different sources used at SPRO had different types of content. Electronic sources might be, for example, patent databases, IPC classification systems, full-text applications, and image search. Paper-based sources may contain, for example, patent applications, images, and other patent-related documents, while human sources contribute with expert knowledge of search strategies, knowledge of specific topic areas, an awareness of specific content in certain patent documents, etc. PEs performing tasks of types PCT2, AS, and C used a low number of source content types, while task types A, PCT1, and A+ITS showed a high level of use of different content types (see Appendix G's Table 6.19).

Sources with 'full text' are clearly the most important content type, followed by the classification code, abstracts, and then images (for PCT1 and PCT2). Furthermore, we observed a strong relationship between content types. When classification codes were considered important, the abstract co-occurred as important. This was observed in 23 of the 54 tasks observed. The same relation was observed between classification codes and images (in 20 of the 54 tasks). It shows how important different document components are in the final part of the patent handling process.

Source summary:
We now consider important findings related to our research question 3: What are the effects of the work task features on the types of sources and source content used? The patent engineers utilised multiple sources (3a in Section 3.3) as well as various types of sources (3b) during a work task performance; they made use of various types of sources – paper-based ones (lexicons, dictionaries, books, etc.) and both human and electronic information sources (such as IR systems and classification code schemata) – in order to accomplish their problem-solving task. The number of sources used per task ranged from one to six. Patent engineers also combined sources throughout the task performance process. The mainstream IR viewpoint almost always assumes that only one source is used in the search process, while several are used in reality. Thirdly, we found that in addition to selecting different sources, a patent engineer used different types of content: classification codes, full text, abstracts, and images, depending on task type.

Query formulation
In this section, we are concerned with our research question 4, on the effects of the work task features on the query formulation. It is addressed through sections 6.2.12 to 6.2.17. Generally, the information retrieval process starts when the source is selected and a query is formulated. The *query formulation and reformulation* phase involves the iterative formulation and reformulation of the information need identified. In the

study, we covered six aspects of query formulation. The first two are the number of expressed terms (6.2.12) and of query elements (6.2.13), The number of synonyms (6.2.14) and that of classification codes (6.2.15) used were observed. Finally, we considered the number of terms per query (6.2.16) and combinations of search elements within a query (6.2.17). Data related to these aspects of query formulation were gathered through electronic diaries and via the attached log files.

6.2.12 Number of unique terms expressed

For execution of the problem-solving task, the information need was expressed, for example, by means of terms. The number of terms may indicate the complexity of the task at hand. For task types PCT2 and C, there were no data reported. First the terms were calculated for each type of task (see Appendix G, Table 6.20).

The A task type had 13 tasks, and, all told, 80 unique terms were expressed for these tasks. The average number of expressed search terms per task was then calculated. Two task types have a rather low score (4.3–4.7) for the number of unique expressed terms (PCT1 and AS), while another two (A and A+ITS) feature a higher number (6.2–8.1) of expressed unique terms per task. The A+ITS task has almost twice as many terms expressing the information need prior to the search as the assignment and the PCT1 tasks do (again, it must be stressed that we monitored only sub-units of the entire task process of each task). The reason for this may be that in the case of the AS task type, prior search terms may already have been identified, with the patent engineer only using them because of the character of the AS task type, while the A and A+ITS tasks need more careful background work before entry in the search system, on account of the complexity of the task types.

6.2.13 Number of types of query elements

We found that not only terms were used. Codes, document IDs, dates, etc. were used too. The most frequently used query elements were terms[50], synonyms, classification codes, and document ID. They all were used in task types A, PCT1, A+ITS, and AS. Furthermore, year, country, and structure search were used in A, A+ITS, and PCT1.

No data exist for task types PCT2 and C, since it was not possible to collect log information from these searches. Not surprisingly, when counting co-occurrences of elements, we found a strong relationship between terms and classification codes. In 30 of the 34 tasks, they co-existed. This may indicate that a multifaceted approach is used in the patent domain.

6.2.14 Number of synonyms and terms per session

Synonyms were often expressed in diaries in connection with the terms used. Task type A+ITS shows an interestingly high level of synonym usage (see Appendix G, Table 6.21). With these added to the terms used, the average is almost 20 search expressions per session. This is more than twice as high as within A tasks (8,25). One explanation might be that the A+ITS is a more complex task, involving multiple

[50] Here, the use of the word ‘term’ covers both terms from a thesaurus and free keywords. The two have not been separated in our study.

topics, and that the ITS applications come from international applicants so the patent engineer needs to recheck prior searches and investigations. For practical reasons (the investigator did not have access to the log data), the data from PCT2 and C tasks could not be collected and analysed.

6.2.15 Number of terms per query string

Furthermore, we investigated the average number of query terms per query string (see Appendix G's Table 6.22). Said data were collected through the attached log files; however, no data for task type PCT2 or C could be obtained. A query string is here defined as the number of terms (composed of characters). First, we calculated average terms used per query string and then the average number of query strings per session. Task type A+ITS showed, overall, a higher number of terms used, on average, during each session. The reason for this can be found in that the A task has an ITS added and that a second, more thorough and in-depth, search must be done.

6.2.16 Combination of types of search elements within a query

In a domain-specific setting, the use of different types of search elements is common, as is *combination* of these elements. In this section, we deal with the latter. We found that the patent engineers did combine different types of elements. The data were collected from multiple sessions from within each of the 34 individual tasks, and the distribution of the combinations can be seen in Table 6.23 (in Appendix G).

The figures were calculated in the following way: First, we defined different types of query elements. For example, reference to single elements indicates use of a single term, and double terms or codes mean that two or more terms or classifications codes have been used in a query. We then calculated all occurrences of all elements in a search session. The occurrences of the elements were then divided by the number of tasks; for example, for the use of two (double) terms or codes in task type A, we found an average of 12.8 (10 tasks). Finally, the range of occurrences of the most important elements is presented. PCT2 and C do not have any data, the reason for this being that it was not possible to acquire those data from SPRO.

We found that the patent engineers did use different query components and that they were combined in different ways for different task types. We observed a high level of usage of double terms or codes in queries for A, PCT1, and A+ITS tasks (49%, 42%, and 37%, respectively). The AS task type showed a low level of double term/code usage. One possible interpretation of the distribution of AS tasks is that these patent applications require thorough investigation. There may have been no prior investigation of these; therefore, some queries need to cover more than one topic area.

Task type AS (29%) showed a rather high level of combination of terms and classification codes. In this case, the patent engineers wanted to cover both a specific area and a term used in that specific topic area. However, the dataset was too small for any significant conclusions. Document ID in combination with classification code was used in 20% of PCT1 tasks. Furthermore, cited documents were commonly used in task type AS (29%). Finally, a combination of terms and structures (e.g., chemical structures) was used in 20% of the A-ITS tasks. The distribution points to chemical

structures usually occurring in patent applications that have international coverage (ITS). However, these results are only indicative, as the dataset was small.

If we look at only the use of a single term or code, A tasks had 166 occurrences in 10 tasks, for an average of 16.6 per task; PCT1 had an average of 15.6 terms or codes per task; and on the low side we find task type AS, with six terms or codes per task.

6.2.17 Number of unique classification codes

Usually, search terms or keywords are used as representations of the information need. However, in certain domains, other types of representations may be used. When analysing data collected from the log files, we found that within the patent domain, classification codes were widely used in the early phases of search performance (see Appendix G's Table 6.24).

Task types A, PCT1, A+ITS, and AS all fell into the range of having 4–5 unique classification codes in the queries per task. The average was 4.4 for all tasks. It must be noted that the same classification code may be used several times in combinations with other classification codes. This could mean that the use of classification codes is a general tool for capturing relevant sub-domains for further in-depth searching. The PCT2 and C tasks did not result in any data. Use of a large number of classification codes may indicate that the task to be resolved should be considered multifaceted.

Summary of query formulation:
In this section, we present findings related to research question 4, considering what the effects of the work task features on query formulation are.

We found the number of unique expressed terms (4a in Section 3.3) per task to vary between 4.3 and 8.1 across task types. The group with 8.1 terms was more varied in topic and complexity (truncation etc.). The average number of terms divided by terms per session varied between 3.3 and 11.8. Again, there was a wide range observed. This shows that real-life query formulations are complex and involve multiple unique terms in task completion.

Besides unique terms, synonyms were frequent. The number of synonyms used (4c) varies between 0.7 and 8.1 per session across different task types. This shows that, in some cases, different aspects of the terms used were sought.

The number of terms per query string showed variation between 1.6 and 4.1 (4d), and the average number of search strings per session was 6.8 to 11.4.

We also found large variation in the combination of query elements used, by task type (4e). Most often used were double terms or double codes, term and code together and cited materials.

Finally, classification codes (4f) were often used. This indicates that several topics were involved or that some topics needed to be checked within other topic areas.

Relevance judgement and search outcome

In this section, we are concerned with our research question 5, on the effects of the work task features on the relevance judgement utilised. This issue is addressed through sections 6.2.18 to 6.2.21.

The outcome of the search is *judged for relevance*. Relevance judgements are usually related to the retrieval stage and measured in close connection with a search session. We wanted to investigate how and when relevance is judged in a real-life patent handling situation. Four angles were observed, covering the type of relevance judgement applied (6.2.19) and when the RJ was applied (6.2.18), as well as what was judged for relevance (6.2.20) and how it was judged (6.2.21). The data for describing these aspects were collected from electronic diaries as well as from the attached and annotated log files.

6.2.18 Relevance: Relevance judgements in TPP stages

Relevance judgements (RJ) are one focus in IR research, but we may also find RJs in information seeking, outside the IR domain. Usually, relevance judgements made in the task initiation (TI) and information seeking task stages are not recognised (see Table 6.12, below, also in Appendix G, as Table 6.25). When the RJ is made during the TI or IST stage, the patent engineer judges articles that were suggested as readings or documents belonging to the area of the current patent application that had been recommended or assigned to the current patent document as relevant reading or as updates.

Table 6.12: Distribution of the stage of RJ by task type in terms of numbers of relevance judgements made during task stages

Relevance: Relevance judgements in TPP stages	*A task* %	*PCT1* %	*PCT2* %	*A+ITS* %	*AS* %	*C task* %	*Total* %
TI stage (e.g., task initiation and preparation)	6	13	9	7	13	0	9
IST stage (e.g., information gathering (such as human–human information need formulation and source selection)	15	9	9	7	20	0	11
IRT stage (e.g., query formulation or relevance judgement)	79	78	82	87	67	100	80
TI+IST total	21%	22%	18%	14%	33%	0	20%
n (judgements)	33	23	11	15	15	7	104
n (tasks)	13	11	9	9	5	7	54

Legend: TI = Task initiation stage; IST = Information seeking task stage; IRT = Information retrieval task stage

Not surprisingly, the relevance judgement is mostly done in the information retrieval task stage (67–100% for all task types)[51]. The 54 tasks observed also revealed that RJs also were made in the TI stage and the IST phase. In total, 11% of the RJs were made in the IST stage and 9% at the TI stage; that is, 20% of the RJ work was done outside the traditional IR phase. This shows that there are variations in the RJ

[51] Note that in one task, relevance judgements can be made in all three phases and, therefore, the total number of phases in which judgements have been made can be 3.

performance, depending on the tasks. Relevance judgements made at the start of a work task and even during the task performance process affect the overall process.

6.2.19 Relevance: Applications of relevance judgements

Based on the data from on-site observations, we could identify two categories of relevance judgement tactics. We denoted them as the 'sequenced relevance judgements' and 'aggregated relevance judgements'.

Sequenced relevance judgements
By our term 'sequenced', we refer to a document or set of documents being judged for relevance continuously in the search task process. In this situation, the document was read at each stage when retrieved in more or less detail, for decision on whether it would be used in the end. This activity also includes the possibility of the relevance judgement being saved and used to give the actor new information, such as new keywords or information about classification codes, for use in a subsequent query.

Aggregated relevance judgements
In 'aggregated relevance judgement', one or more sets of documents are saved by the actors (PEs) during the information search process. This tactic usually involves a two-step process. Typically, the document sets are saved/stored after each search session in order to be judged a second time in a later, final relevance judgement session. During the search process, each query could in fact (but not always) create a set of documents that at the end of the search process are judged for relevance. Usually, when the actor makes a query and retrieves a set of document objects, this set may be saved in at least one of the following three graded ways:

a) It is saved/stored as it was retrieved, if small enough (1–2 documents)
b) The set is saved/stored, containing a selection of 5–10 documents with only a less close inspection (often only title)
c) The set of documents is saved/stored on the basis of closer inspection of the resulting set (usually with full bibliographic information, including the abstract)

These three aspects of aggregated relevance judgement were found in our study. At the end of the search task, all documents saved/stored were read and judged for final relevance, for decision on whether any of the documents could be used against the patent application.

Table 6.13: Distribution of relevance judgement strategy application by task type (n = 52)

	A	*%*	*PCT1*	*%*	*PCT2*	*%*	*A+ITS*	*%*	*AS*	*%*	*C*	*%*
Sequenced relevance judgement	13	100	4	28	0	0	7	100	4	80	0	0
Aggregated relevance judgements	12	92	11	100	9	100	7	100	3	60	7	100
n = tasks	13		11		9		7		5		7	

Table 6.13, above (also Table 6.26 in Appendix G), shows that, in some cases, the PE used both relevance judgement tactics during the performance. Task types A (100%/92%), PCT1 (28%/100%), and A+ITS (100%/100%) especially evidence combination of the two relevance judgement tactics. Aggregated relevance judgement was made mostly in PCT1, PCT2, and C tasks (100% in each of them). One category

of patent application tasks required only one kind of relevance judgements before being completed, while others required more steps of filtering procedures. When there was a multi-session task, relevance judgements were usually made for each session.

6.2.20 Relevance: Elements judged to be relevant

During the relevance judgement, the PEs pointed to different elements in the document that were important for the relevance judgement process, such as the summary, figures, claims, description, terms, classification codes, references, and bibliographic information, as shown in Table 6.14, below (Appendix G's Table 6.27).

Table 6.14: Distribution (%) of types of document elements judged for relevance across task type (n = 52)

Types of document elements used	*A %*	*PCT1 %*	*PCT2 %*	*A+ITS %*	*AS %*	*C %*
Summary	16	24	17	18	17	5.5
Full figure or part of figure	21	17	26	18	17	11
Claim	16	13	20	14	17	5.5
Description	16	24	20	21	25	5.5
Term	10	11	0	11	0	0
Classification code	14	6	2	14	17	39
Reference	7	2	15	4	8	34
Bibliographic	0	2	0	0	0	0
	100	100	100	100	100	100
n = 221	71	46	46	28	12	18
Average number of components judged per task	5.46	4.18	5.11	4.00	2.40	2.57
n=52 (tasks)	13	11	9	7	5	7

We found that different elements of patent documents were used. In total, 221 elements were recorded. Task types A, PCT1, and PCT2 showed a varied usage of elements. The national task type, A, also had the highest average number of elements used per task (5.46), and the lowest score was seen for task type AS, with an average of 2.40. Overall, these results imply varied and complex usage of what is judged relevant or on what basis the RJ is done. Summaries, figures, claims, and descriptions were involved especially frequently in the relevance judgements. The analysis shows how important these document elements are in the final part of the patent handling process. It also shows that the types of text judged when a PE is resolving a task are rather diverse. Usually, in a general IR evaluation context, only one type of text is used for RJs. One consequence for general IR evaluation would be that both the dataset design and the evaluation criteria and measures should be commensurate with such a situation. Further, we found that certain elements are frequently used for judgement: for PCT1, 'summary' and 'description'; for PCT2, 'claim' and 'description'.

6.2.21 Relevance: RJ degrees in the various types of tasks

The final aspect we observed when dealing with RJs concerns the combination of categories or 'degrees' of RJ. The purpose was to see how many instances of RJ were made. The patent engineers judged retrieved documents against a predefined set of criteria (x, y, a, z, and &, among others). See Subsection 5.4.1 for a more detailed description of the relevance judgement categories. These criteria correspond to certain relevance definitions according to a) how well the document at hand matches the

information need and b) in what way the document can be judged to be part of the solution and of completion of the task. We calculated the average number of RJ degrees used for a task. Task types A and A+ITS showed a high number of relevance judgements per task (9.2 and 9.00).

Summary of relevance judgement:
This section presents findings related to research question 5: What are the effects of the work task features on relevance judgement performance?

The most interesting aspect observed and analysed was the different relevance judgement 'tactics' the patent engineers employed: '*sequenced relevance judgements*' and '*aggregated relevance judgements*' (5b in Section 3.3). Many tasks involved both of these tactics and therefore a more complex relevance situation arose.

We also found that relevance is judged not only in conjunction with the information retrieval phase; rather, it is judged, more or less, in all phases of the IS&R performance process (5a). Relevance judgements are mostly done in the information retrieval task stage (between 67 and 100% for each task type). However, the empirical study revealed that 20% of the relevance judgements were made outside the traditional IR phase. This means that in real-life situations, certain judgements have already been performed before those normally considered the 'real' relevance judgements.

On average, 1–5.3 relevance judgements were made for each task. In cases where only one relevance judgement was made, this might be an aggregated relevance judgement.

We also found that different document elements (5c) were judged for relevance. Most judged were 'summary' and 'description' (PCT1); 'claim' and 'description' (PCT2); and, finally, for task type C, 'classification code'. This finding also points to the fact that different portions of a document are being judged, depending on the task type.

The average number of relevance judgement degrees (5d) per task ranged from two to 9.2, indicating that some tasks are more varied than others. This empirical finding shows that real-life tasks involve more variation in relevance judgements that anticipated.

6.3 The task completion stage - information use

This section addresses research question 6, on the effects of the work task features on the information use judgement. Subsections 6.3.1 to 6.3.2 deal with this issue.

A major aspect of a work task is the *task completion* phase. This phase may involve compiling and extracting information from the relevant retrieved documents. On the basis of the relevance assessment procedure, the patent engineer makes a decision or creates a solution regarding the present patent application. The decision is formally written as an answer to the applicant in the form of a report. The retrieved documents are used in *preparation of various reports, according to the type of PA*. Depending on the type of PA, the PA may go back to the *applicant for revision*. When a revised version is submitted to the patent office, the PA undergoes an *inspection*. Finally, when

the applicant and the patent office have agreed with each other, a *public announcement of patent* is performed and the patent is *filed nationally and internationally*.

Two variables are reported on here: the types of (electronic) information sources used (6.3.1) and the types of information components used (6.3.2). Detailed description from the electronic diaries and annotated log files provided data for these variables.

6.3.1 Types of information sources used

For completion of the task, different types of information were used. In our study, we found the following types of information sources used: patent applications, journal articles Web pages, images, classification code schemata (such as the IPC), and bibliographic information (see Appendix G's Table 6.28, which presents the percentage of tasks of each PA type using each source type). Often several information source types were used for finalising the task. Six types of electronic sources were identified. These were used on 137 occasions during the 54 patent tasks monitored. It must again be noted that these numbers account for only the specific periods during which the investigator made observations.

However, this does provide an indication of how electronic sources were used. For example, the A+ITS task type showed the greatest variety in the information sources used (3.7), followed by the A tasks (2.9) and PCT1 tasks (2.6). The information components most frequently used are, of course, all of the various sections of the patent application that must be read. We may also note that classification codes and images are of great importance. For example, in 62% of the 13 type-A tasks, 75% of PCT1 tasks, and 87% of the eight A+ITS tasks, at least one image was used. This great use of different information types might indicate a complex task, which highlights another factor that generally is not taken into account in laboratory IR experiments. In addition, both human- and paper-based sources were used, as reported for the 12 tasks subject to on-site observation. In total, 26 instances in which paper notes were used were recorded.

6.3.2 Types of information components used

On a lower level, looking at the specific document components used in one of the source types, the patent document, we found that different patent application 'components' were used in creation of an end product (a report). The patent application itself is first read as a whole; then, in the next step, specific parts of a document are marked as useful. The distribution of elements used is shown in Table 6.15 (and also in Appendix G, as Table 6.29).

By 'useful' we mean that these components are pointed out, referred to, or mentioned in the report as evidence supporting the final decision made by the PE. As an example, in tasks of type A, one or several images from a patent document were selected 54 times (in 13 tasks) and referred to as important or necessary for the final outcome of the task performance. This corresponds to 12% of the full 54 tasks.

Table 6.15: Distribution of information components used by task type in terms of average percentage of components used (*n* = 54)

Patent document components	*A %*	*PCT1 %*	*PCT2 %*	*A+ITS %*	*AS %*	*C %*	*Total %*
Image	12	19	17	21	27	25	76
Paragraph[52]	18	19	21	18	18	25	79
Abstract	16	17	19				57
Section[53]	7		21	15		21	52
Reference	11		17		18	21	46
Terms	16	15					41
Classification code	20	17		15	27		59
	100	100	100	100	100	100	
	431	473	478	367	220	400	
n tasks =	13	11	9	9	5	7	
Avg. component no. & *n* tasks	4.3	4.6	4.4	4.1	2.2	4.0	

By far the most used patent application components across all task types are the paragraph texts (used in 79% of all tasks), images (76%), classification codes (59%), and abstracts (57%). Therefore, images seem definitely to be an important element for consideration.

Summary of information use:
Finally, we present important findings related to research question 6, concerning the effects of the work task features on information use.

We found that different types of information sources (6a) are used and that often several types of information sources (1.3–3.7) were used for completion of the task. In laboratory IR experiments, this is, in general, not taken into account.

It is also commonplace to use information components of several types (6b) when finalising the end product. For example, images (used in 76% of all patent tasks) were an important component alongside the textual components of the task performance process. These two variables show that there is extensive and interesting work performed after the actual relevance judgement stage and before the task is considered to be completed. Implicitly, the RJ was made with the final product (the final PA report) in mind.

[52] A paragraph is a piece of text within the PA. It may be one sentence or a couple of sentences.
[53] In contrast to a paragraph, a section is defined as an entirely separate piece of text with a heading.

7

CROSS-TABULATION AND RELATIONSHIPS

In Chapter 6, we presented a descriptive analysis of observed variables and their extracted indicators and values. In this chapter, we present further reasoning concerning the relationships and possible dependencies between these variables. The reasoning will be tied in with research questions 1 to 6 (question 7 is dealt with separately, in Chapter 8). For each section, we address the relevant research question.

We report on dependencies on single variables at work task level (in Section 7.1) with the categories of knowledge level (Subsection 7.1.1) and task planning (Subsection 7.1.2). These sections correspond to research question 1. Furthermore, at the levels of information seeking and information retrieval (Section 7.2), research questions 2 to 6 are addressed in more detail. The category of information need (Subsection 7.2.1) involves research question 2, while that of source (Subsection 7.2.2) is related to research question 3 and query formulation (Subsection 7.2.3) to research question 4. Finally, the category of relevance judgement (Subsection 7.2.4) is considered in view of question 5 and task completion (Subsection 7.2.5) in view of question 6. All details for the figures and their reference to tables in this chapter can be found in Appendix H.

7.1 Work task level

7.1.1 Patent engineer and knowledge types

In this study, we distinguish between two knowledge types among patent engineers that illuminated interesting results. Domain knowledge is knowledge about the topic of the task, while task knowledge involves the task procedure and how the task needs to be performed.

Domain knowledge:
We found that domain knowledge affects the number of sources used. When the patent engineer possesses little domain knowledge, this leads to use of a large number

of sources (χ^2 = 6.75, p < .01; (see Table 7.15 in Appendix H and Figure 7.1). Low domain knowledge also leads to a low number of query terms per session / string used: χ^2 = 6.01, p < .025 (see Table 7.16) and rho -1.00 (see Table 7.33 and Figure 7.1). Domain knowledge also has a weak but noticeable effect on perceived task knowledge: p < .10, χ^2 = 3.076 (see Table 7.1 and Figure 7.1), which means that a patent engineer with high domain knowledge also tends to have high task knowledge.

Knowledge types affect the strategy of relevance judgements. High domain knowledge (HDK) often leads to utility of an aggregated RJ strategy, and low domain knowledge (LDK) leads to a combined (aggregated and sequential) RJ. However, while this difference, taken on its own, is not statistically significant by the χ^2 test (p < .10; see Table 7.8 and Figure 7.1), it is reported here since in combination with the other results it is part of the characteristics of the search strategies studied.

Task knowledge:
Task knowledge as viewed in the study refers to both problem knowledge (PK) and problem-solving knowledge (PSK) as described by Byström and Järvelin (1995). These knowledge types may vary; therefore, they can affect other variables. Knowledge about the task to perform often leads to the PE having a clear perception of the information needed to perform the task (χ^2 = 4.840, p < .05; see Table 7.17 and Figure 7.1), and a low level of task knowledge correlates with a high number of relevance judgements in all task performance process stages (rho = 1.00; see Table 7.34 and Figure 7.1).

Worthy of mention is that there was not a significant dependency between relevance judgement types (aggregated/sequential (AS) and aggregated (A)) and task knowledge (level of perceived task difficulty). However, we observed high perceived difficulty to be linked to an AS relevance judgement strategy and low perceived difficulty slightly to A relevance judgement. Again, this is not at all significant (χ^2 = 2.00, p < .50; see Figure 7.1).

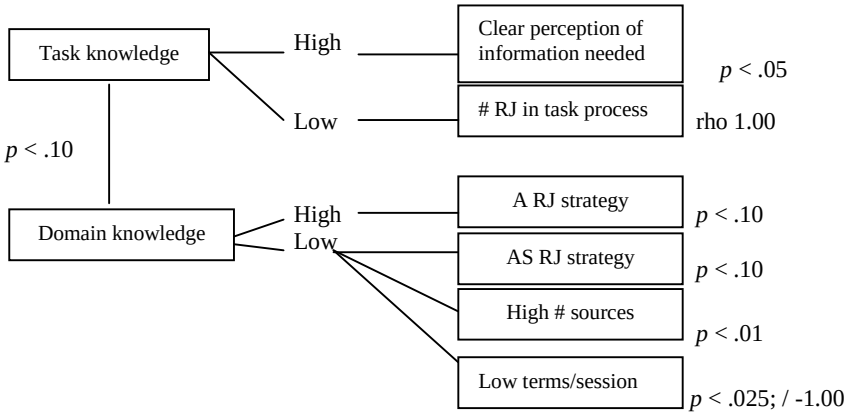


Figure 7.1: Knowledge types

7.1.2 Task planning

When one begins a patent task, it is normal to do some sort of task structuring. This planning activity is linked to how well the patent engineer knows the task procedures. In our study, the task planning could be perceived as both clear/unclear and

structured/unstructured. The distinction between clear and structured planning is that with clear planning, the PE has a clear sense of how the task should be performed, while structured planning may, for example, involve the order in which the patent engineer decides to execute different subtasks (time) and the people to ask at certain moments during the task performance. A significant dependency emerged in that when a patent bureau has prepared the patent application, planning of the task is often clear and structured ($\chi^2 = 14.89$, $p < .001$; see Table 7.24 and Figure 7.2), and when a company submits the application, the planning of the task is clear but unstructured.

We found significant dependencies between clear task planning and the expressed information need ($\chi^2 = 4.11$, $p < .05$; see Table 7.21 and Figure 7.2) and between clear task planning and the relevance judgements ($\chi^2 = 7.34$, $p < .01$; see Table 7.23 and Figure 7.2). Clear planning leads to the use of multiple expressed information needs ($\chi^2 = 4.11$, $p < .05$; see Table 7.21 and Figure 7.2). Clear planning of the work task results more often in a high number of relevance judgements being performed; by contrast, clear *and* structured planning leads to a low number of RJs in the task performance process. This could mean that if the PE has a clear perception of both how and when to execute the different activities in the task performance process, one result is a lower number of RJs being made.

When the planning of the task is seen as clear, the PE often uses a combination of sources ($\chi^2 = 3.91$, $p < .05$; see Table 7.22 and Figure 7.2) as well as a high number of RJs ($\chi^2 = 7.34$, $p < .01$; see Table 7.23 and Figure 7.2) throughout the task performance process. If the planning of the task is perceived as both clear and structured, only a single source type is used ($\chi^2 = 3.91$, $p < .05$; see Table 7.22 and Figure 7.2) and few relevance assessments are done ($\chi^2 = 7.34$, $p < .01$; see Table 7.23 and Figure 7.2). Finally, there is a dependency between tasks that have been perceived as clear and structured and short task completion time ($\chi^2 = 9.36$, $p < .005$; see Table 7.25 and Figure 7.2).

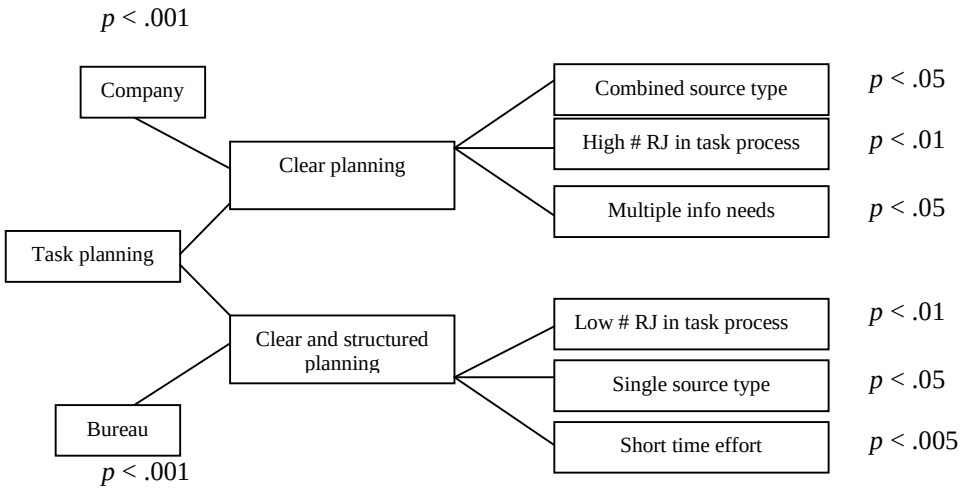


Figure 7.2: Work task planning

Summary:
This section presented findings related to research question 1, on the effects of work task features on the work task. When a work task is about to be initiated, we observed, the patent engineers' knowledge levels had significant correlations with

a) The information need (clarity),
b) Information sources (number of sources used),
c) The query terms used, and
d) The relevance judgement strategy used.

Furthermore, planning for a structured approach to the work task may, for example, involve the order of completion of the subtasks, whom to ask and when, and other things. The planning of the patent task will affect the IS&R process, depending on, e.g., the type of applicant. The patent engineer can plan or adapt as the situation dictates. Clear and structured planning also has dependencies with expressed information need and the use of a single source type or multiple types. We may also expect the number of relevance judgements made during the task performance process to be related to whether the work task is perceived as clear or as both clear and structured.

7.2 The IS&R task

7.2.1 Information need

We have showed that, in real situations, the information need is not always stable; instead, it is changing. Furthermore, the information need is not always a single need – multiple needs may occur during the search process. In the previous chapter, Chapter 6, we discussed seven variables related to information needs, of which the following variables showed interesting dependencies: the change of information need, the expressed information need, and the type of information needed.

Stability of information need:
The stability of information need during IS&R task performance showed dependencies with three aspects of the IS&R process: combination of source types, use of unique classification codes, and the number of query terms.

The correlation data ($p < .05$, as seen in Table 7.10, and rho .950; see Table 7.28 and Figure 7.3) show that when there is a stable information need, either a single source or a combination of sources may be used. When a change in the information need occurs, then only one source is used – e.g., an electronic source.

Not surprisingly, stability of information needs also showed a significant correlation with the number of query terms. A change in the information need usually results in use of a low number of query terms ($p < .025$, as seen in Table 7.11, and rho .950; see Table 7.28 and Figure 7.3). This might indicate that before complicated and advanced query formulations are made, the source is tested with fewer and more 'explorative' query terms. A weaker, but interesting and slightly unexpected, dependency relationship was found between change in the information need and a low number of unique classification codes ($p < .10$; see Table 7.12 and Figure 7.3).

Expressed information need:
An information need may be expressed by means of single or multiple terms or codes. There is a significant dependency between use of a low number of sources and the usage of single terms/keywords ($p < .001$; see Table 7.13 and Figure 7.3). We also found, not surprisingly, that the information need was expressed with multiple terms/keywords when the PE used a large number of sources and utilised a wide variety of source types ($p < .005$; see Table 7.14 and Figure 7.3).

A higher number of expressed search terms suggested for use prior to search correlates with probable change of information need during the task performance process (Spearman's rho .950; see Table 7.35 and Figure 7.3). The use of a large number of suggested search terms may then reflect uncertainty on the part of the patent engineer. This uncertainty may be caused by the current topic, or it may reflect that several topics are involved in the specific patent task at issue.

Clear or muddled perceived information need:
A muddled need for information leads to a high number of search strings (Spearman's rho 1.00; see Table 7.30 and Figure 7.3). A muddled need also leads to the use of a combination of source types ($\chi^2 = 6.38$ and $p < .025$, as shown in Table 7.18 and in Figure 7.3, below).

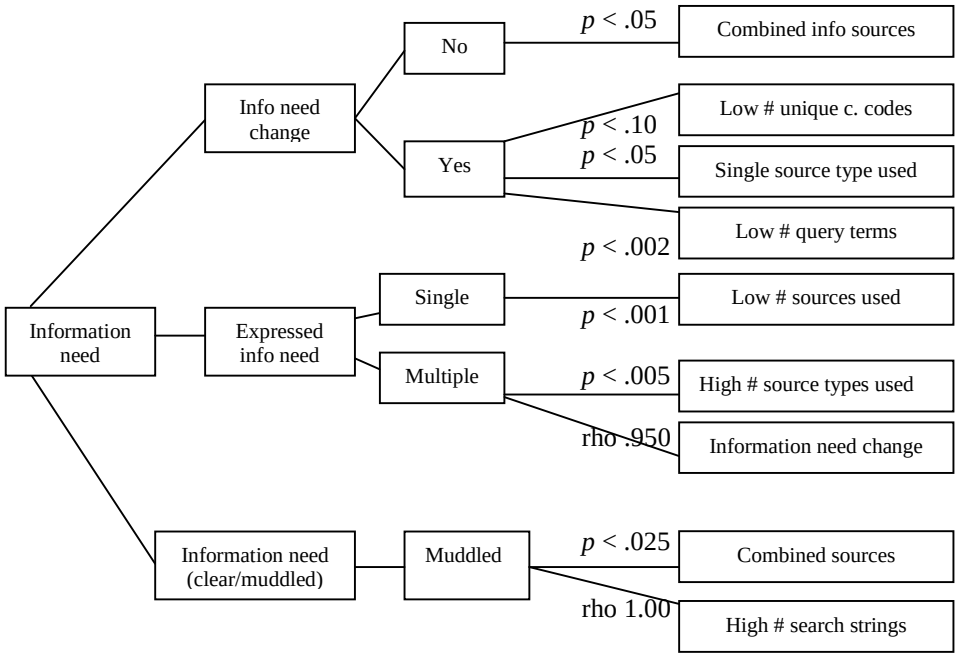


Figure 7.3: Information need variables

Summary:
This section presented findings related to research question 2: What are the effects of the work task features on the information need?

We found that two variables in the information need category showed dependency relationships with other aspects of the IS&R process: change in the information need and expressed information need (as terms). An information need is not merely a stable statement of what is needed and thus the foundation on which the search is

based. In real-life patent search situations, it is common for the information need to change during the course of the search process, for any of various reasons. A change in the information need affects the search process in that fewer query terms are used, few source types are used, and not as many classification codes are used. When the information need is represented by multiple terms, more sources are used, different source types are used, and eventually circumstances also lead to a change in the information need.

Finally, the expressed information need being muddled leads to use of a combination of all three sources (HEP) and to a high average number of search strings being used.

7.2.2 Source

We found two aspects of the source category that significantly affect the task performance process: the use of single/combined types of sources and the number of sources used.

Source type:
The categories of source types used were paper (P), electronic (E), and human (H), grouped as follows: as single source type (P, H, or E) or combinations thereof (PHE, PH, PE[54], or HE), as noted above. In the case of a single source type, a different source of the same type may have been opened and closed once or several times.

A significant dependency was found between the manner of use of source types and the clarity of the perceived information need ($p < .025$; see Table 7.18). When the patent engineer had a muddled view of what information was required (for example, 'I need to use databases X and Y, I need to ask my colleague, and I need to look in the old Nordic patent files'), the patent engineer usually used all three types of sources (i.e., HPE) for resolving the task. If the perceived information need was clear, only one of the source types was used (E, H, or P). Thus uncertainty leads to ensuring coverage through the use of all source types available.

The source types used also are related to the relevance judgement process. When only a single source (P, H, or E) is used, only the aggregation relevance judgement strategy is used ($p < .01$; see Table 7.2). Sequential evaluation of a single source could be done through performance of several searches with one source. For example, search #1 ('potato') might be followed by an assessment and then search #2 ('tomato'), followed by another assessment, with, finally, a search #3 ('salad') followed by a third assessment. It is statistically significant that the greater the number of sources used, the more frequently a combined relevance judgement strategy is used. This is in line with the variable for the use of a high number of sources, described below. The difference is that one type of source (for instance, electronic) may have been used several times during the task performance process.

Number of sources:
The number of sources affects the task performance process. We found three dependencies here.

[54] Here, 'PE' refers to paper and electronic sources.

The number of sources used is dependent on who wrote the application. When a company-based patent department wrote and submitted the patent application, usually the number of sources used was high (4–7) (see Table 7.19, where $p <= 0.025$, as in Figure 7.4). Those written by a patent bureau were almost evenly split between low (1–3) and high (4–7) numbers of sources used. A weak dependency was seen between a high number of sources and the use of a combined aggregate/sequential RJ strategy ($p < .10$; see Table 7.7 and Figure 7.4), but rho showed a stronger dependency (rho .943; see Table 7.29 and Figure 7.4). Finally, when the PE used a large number of sources (4–7), the amount of time spent on the task (6–25 hours) was high too (see Table 7.20; $\chi^2 = 9.56$, $p >= 0.025$ – as shown in Figure 7.4, below).

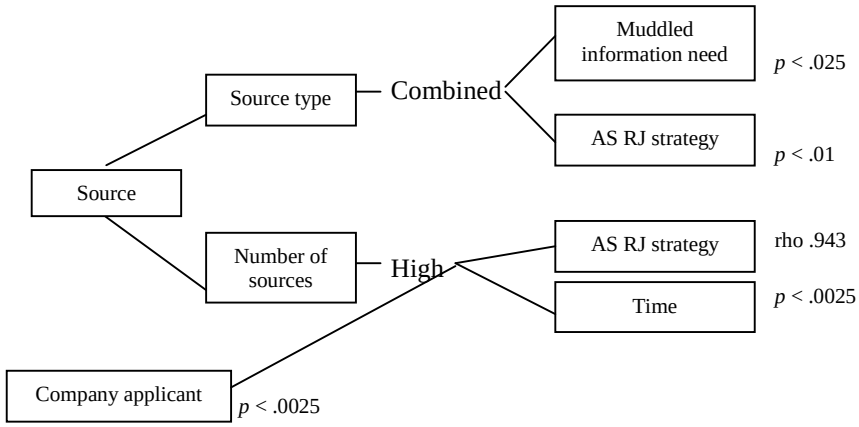


Figure 7.4: Source types

Summary:
In this section, we present findings related to research question 3: What are the effects of the work task features on the types of sources and source content used?

The type of applicant may affect the number of sources used. In our case, if an application was written by a patent department within a company, this led to the use of a higher number of sources. The reason may be that these applications are written in a more complex way or with more references that need to be checked.

Traditional IR research is usually concerned with one electronic source and one relevance judgement activity. In real-world search tasks, the situation may be different. In our case, we found that a high number of sources often leads to the use of a combination of aggregated and sequential relevance judgement strategies. This is obviously in contrast to traditional IR experiments, since in our findings, a) several sources are used; b) different types of relevance judgement strategies are being employed; and, c) not just one relevance judgement action is taken – several are. This should, in fact, have an impact on traditional measurements of precision and recall. This is a similar result to that with the correlation of combined source types and relevance judgement. Usually, when a single source type is used, the PE has a clear perception of what information is needed for completion of the task. This might include several sources of the same type being used but not several source types.

7.2.3 Query formulation

In the previous sections of this chapter, we have already highlighted some interesting correlations, but we will briefly repeat them here.

Number of query terms per string/session:
A significant dependency was found between domain knowledge and the number of terms per session and query string ($\chi^2 = 6.01$, $p < .025$; see Table 7.16 and Figure 7.5) used by the patent engineer. Patent engineers with high domain knowledge used a high number of search terms.

If the patent engineer expressed a high number of search terms prior to the search, it was more likely that the PE also used a higher average number of search items per string (rho 1.00; see Table 7.30 and Figure 7.5) before accessing the IR system.

A stable information need leads to the usage of a high number of query terms per string/session ($p < .0025$; see Table 7.11 / rho .950; see Table 7.28 and Figure 7.5). Finally, use of a high number of query terms per string and session will affect the number (high) of information sources used for finalising the work task and product (rho 1.00; see Table 7.36 and Figure 7.5).

Combination of query types:
Calculations with Spearman's rho showed a correlation between the average number of query types combination with a) number of search sessions per task (rho 1.00; see Table 7.31 and Figure 7.5) and b) number of elements judged in a single task and final product (rho 1.00; see Table 7.32 and Figure 7.5). A large number of combinations of query types often led to a high number of information components being used in the final product (rho .975; see Table 7.37) (see Figure 7.5).

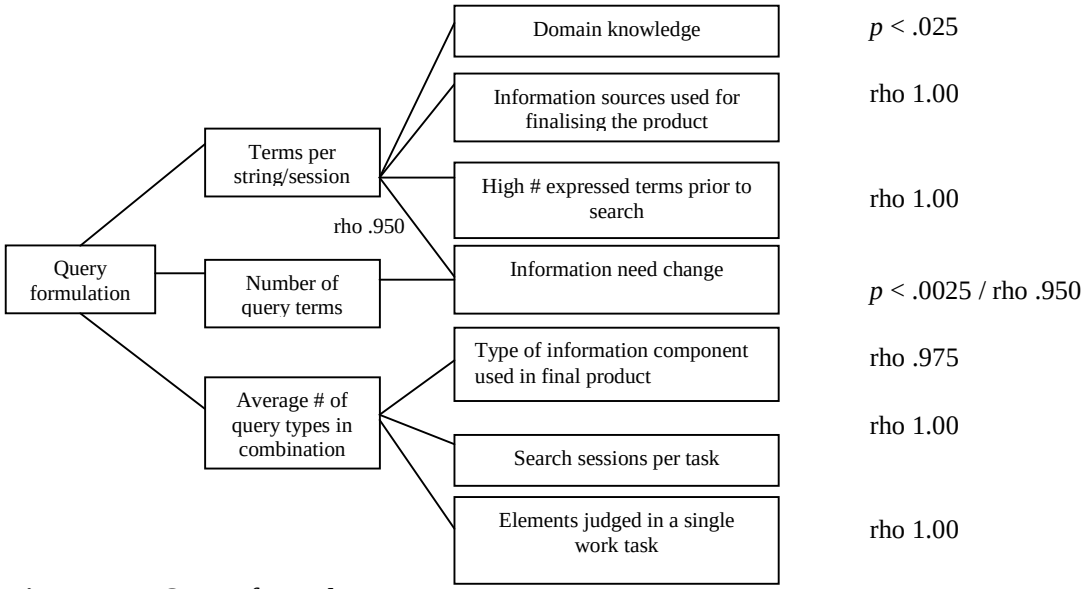


Figure 7.5: Query formulation

Summary:
In this section, we present findings related to research question 4, dealing with the effects of the work task features on query formulation.

We found that insufficient domain knowledge leads to the use of a low number of query terms per session. Greater domain knowledge can give the patent engineer a better starting point for query term usage. Furthermore, patent engineers with low domain knowledge might be supported by a thesaurus or similar.

Another interesting finding is that when the information need is stable, the PE uses a high number of query terms per session, while a change in the information need leads to the use of either a high or a low number of query terms. This may again reflect differences in domain knowledge. The level of complexity can be seen in the number of combinations of query types (those based on terms, codes, date, document number, etc.) used during a search session. In this study, we found that a large number of search sessions also results in a high number of distinct query types being used and, in judging of information, a large number of document elements being judged as well.

7.2.4 Relevance judgement

Two aspects of relevance judgement in particular showed dependency relationships with other factors within the IS&R process: the relevance judgement strategy used and the number of elements judged.

Relevance judgement strategy – aggregated or sequential:
Two separate RJ strategies were identified in performance of a single IS&R task: aggregated and sequential strategies. Only aggregated (A) RJ and a combination of aggregated and sequential (AS) RJ are considered, because of the dataset being too small for sequential relevance judgements.

There are dependencies among three variables within the information need category and the RJ strategies chosen. First, the use of multiple terms for expressing the information need led to the use of an AS RJ strategy ($p < .01$; see Table 7.3 and Figure 7.6). Secondly, when the users had a clear and structured perception of the information needed, they often used the aggregated RJ strategy, while a structured but unclear perceived information need resulted in a combined RJ strategy ($p < .01$; see Table 7.4 and Figure 7.6). Thirdly, a high number of required document components (abstract, claims, etc.) having to be investigated for the formulation of the information need often meant that both aggregated and sequential RJ were used ($\chi^2 = 10.72$, $p < .01$; see Table 7.6 and Figure 7.6). This is discussed in Section 6.3.7.

Statistically significant dependencies were not found between PEs' level of any kind(s) of knowledge and other IS&R variables, but this issue will be reported upon here since it forms a vital part of the landscape of the search strategies studied. The choice of RJ strategy correlates with the actor's level of domain knowledge (topical knowledge). High domain knowledge (HDK), leads to an aggregated RJ strategy, and low domain knowledge (LDK) leads to a combination of aggregated and sequential RJ. However, this difference when taken on its own is not statistically significant by the χ^2 test ($p < .10$; see Table 7.8 and Figure 7.6). Neither did any dependency emerge between RJ strategy and task knowledge. However, an indicative result shows

that high perceived difficulty fosters the AS relevance judgement strategy and a low perceived difficulty weakly fosters the A relevance judgement process. Spearman's rho showed a significant correlation (.929; see Table 7.34 and Figure 7.6).

Two variables related to source category statistically affect the type of RJ strategy. When only a single source type (P, H, or E) is utilised, the patent engineer engages in the aggregation relevance judgement strategy. It is statistically significant that the more source types that are used, the more frequently combined relevance judgement strategies (i.e., AS) are used ($p < .01$; see Table 7.2 and Figure 7.6). Secondly, when using a large number of sources (6–10, 33%) to resolve a patent task, the patent engineers preferred to use an AS relevance judgement strategy ($\chi^2 = 23.85$, $p < .10$; see Table 7.7 and Figure 7.6).

An indicative result ($\chi^2 = 14.18$, $p < .10$; see Table 7.9 and Figure 7.6) is that aggregated RJ is done for classification codes and that combined AS relevance judgement is applied when terms and bibliographic information (document numbers etc.) are considered.

Finally, not surprisingly, using the aggregated relevance judgement strategy leads to short time (1–4 hours) spent on IS&R subtasks within the work task, while a combined relevance judgement strategy correlates with 5–60 hours spent on IS&R task completion ($\chi^2 = 4.01$, $p < .05$; see Table 7.5 and Figure 7.6).

Average number of elements judged for relevance:
The average number of individual elements judged for relevance during the task performance process also has important dependencies on the IS&R process. When, on average, 1–2 elements are judged for relevance, only one information source is used for finalising the end product. When 3–8 elements, on average, are judged for relevance, two information sources are used for the final product ($\chi^2 = 8.60$, $p < .005$; see Table 7.26) (rho .943; see Table 7.29 and Figure 7.6). Not surprisingly, the more sources used, the greater the number of elements judged for relevance.

Furthermore, a high number of source types used led to a high number of elements judged for relevance ($\chi^2 = 9.11$, $p < .005$; see Table 7.27 and Figure 7.6). When all of the various source types are used, usually 2–10 relevance judgements are made during the task performance process. We also found that a high average number of query types used in combination affect the number of elements judged (positive correlation) for relevance in a single task (rho 1.00; see Table 7.32 and Figure 7.6).

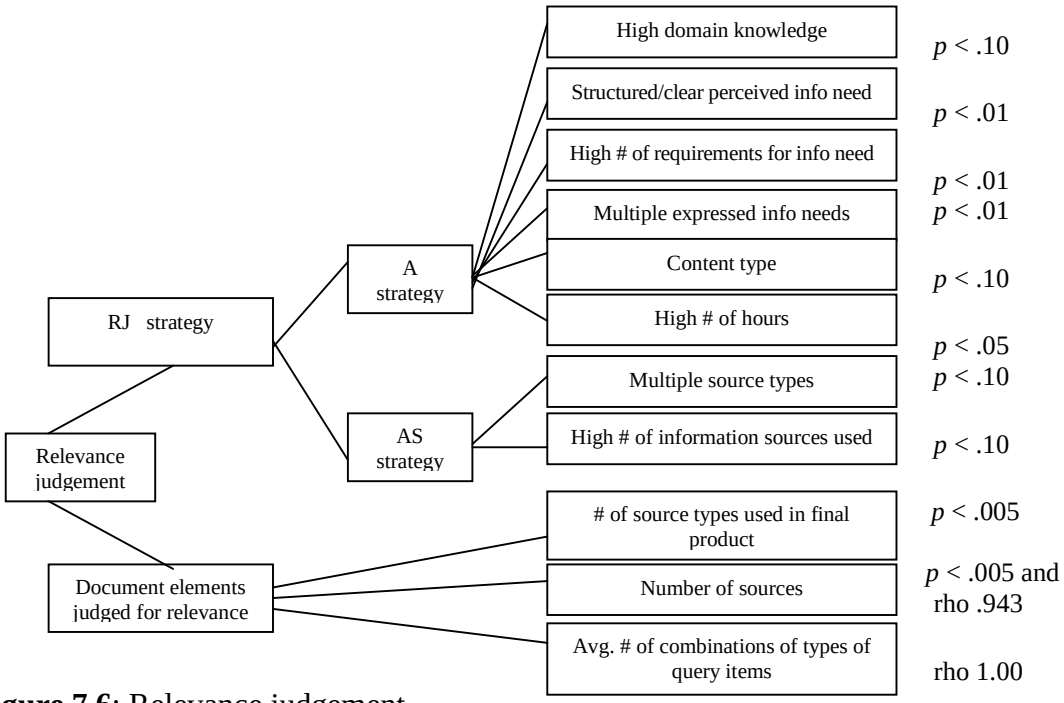


Figure 7.6: Relevance judgement

Summary:
This section presented findings related to research question 5: What are the effects of the work task features on relevance judgement performance? In Chapter 6, we reported on the finding that two RJ strategies were used – aggregated and sequential RJ strategies – when a PE was performing a single IS&R task.

Dependencies exist among three variables in the information need category and the RJ strategies chosen. Using multiple terms for expressing an information need affects the RJ strategy. Use of many terms usually leads to an AS relevance judgement.

Furthermore, the perceived clarity of the information needed also affects the RJ strategy: a perceived clarity leads to the use of an aggregated strategy. In both cases, the reason for this may be the level of knowledge of the patent engineer. However, no statistical evidence was found of a relationship between the patent engineers' knowledge level and RJ strategy, even though we found a weak correlation between a person having strong domain knowledge and the use of an aggregated RJ strategy, while a low level of domain knowledge led to an AS relevance judgement strategy.

Two variables in the source category statistically affect the type of RJ strategy. There is a statistically significant positive correlation between the number of source types being used and use of combined RJ strategies. Time also affects the RJ strategy, in that aggregated RJ strategy leads to spending less time on the search task. Lack of familiarity with the field/domain may also have an affect on the number of sources used and RJ strategy selected.

Finally, source elements also affect relevance judgements. As more sources are used, more document elements are judged for relevance. When the aggregated strategy is applied, usually the abstract and terms are in focus for relevance judgement.

7.2.5 Patent task completion

With respect to task completion, we found two elements relevant for answering our research question: the information source used for the final product and the type of information component used in the final product. When the PE used a large number of expressed terms, a large number of information sources will be used for finalising the product (rho 1.00; see Table 7.36 and Figure 7.7). If the patent engineer used a large number of information components for completion of the final product, it is likely that many search strings were used (rho .975; see Table 7.37 and Figure 7.7).

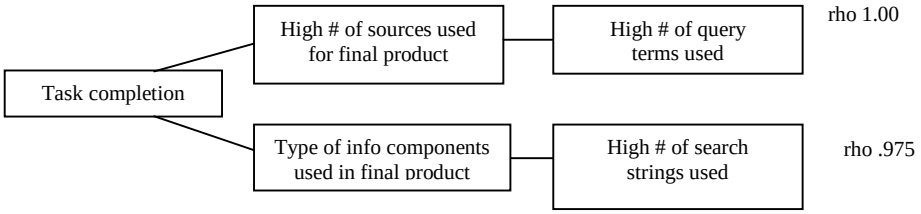


Figure 7.7: Task completion

Summary:
This section presented findings related to research question 6, dealing with the effects of the work task features on information use for completion of the task.

Interestingly, we found that some variables showed a relationship with variables outside the traditional IR model: a) the number of information sources used for finalising the product and b) the type of PA components used in the final product. For example, use of a high number of sources for the end product indicates a more complex search situation in the patent domain. In order for the final product to be comprehensive, more effort is put into the search process.

Furthermore, if the number of expressed search terms to be used is high prior to the search itself, it is more likely that the information need will change during the information seeking activity (.950). The variables addressed here show that the task completion phase is an important part of the interactive information retrieval process.

7.2.6 Connecting relationships

Next, we examine whether any of the significant dependencies described could be related, further, to a third variable. First, we examined the variables within the work task level and we found the following relationship: Low domain knowledge (LDK) leads to the use of a high number of sources (see Figure 7.8), which, in turn, leads to a) a combined (aggregated/sequential) RJ strategy, and b) long time to complete the IS&R task.

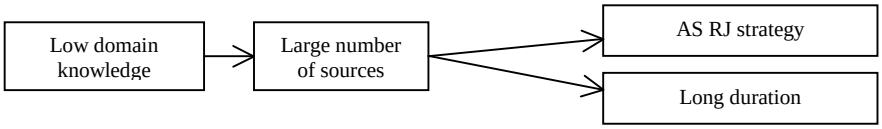


Figure 7.8: Extended relationship of domain knowledge

High task knowledge (HTK) leads to clarity of the information need (see Figure 7.9), which then results in an aggregated relevance judgement strategy. This indicates that the aggregated RJ is used when the actor is fairly confident with respect to his or her problem-solving task.

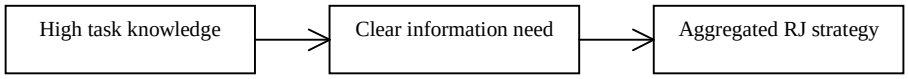


Figure 7.9: Extended relationship of task knowledge

Next, we examined the variables in the IS&R process. Whether the applicant was a company with its own patent department or a patent bureau affects the planning stage (see Figure 7.10, below). This, in turn, leads to combined (for a company) or single (for a bureau) source types being used, which then leads to an aggregated (single) RJ strategy or aggregated/sequential (combined) RJ strategy.

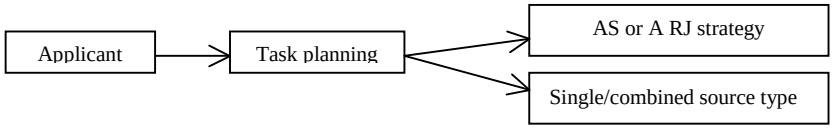


Figure 7.10: Extended relationship of applicant

The change in information needs leads to the use of a combination of information sources (see Figure 7.11), which, in turn, results in a combination of aggregated and sequential RJ strategy (also shown in the figure). The change of information needs also has a relationship with the information need as expressed with single or multiple terms.

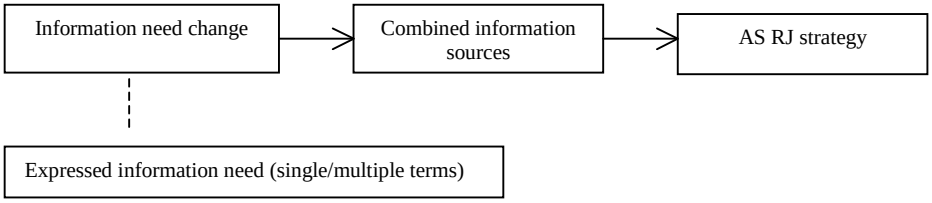


Figure 7.11: Extended relationship of change of information need

Expression of an information need via multiple terms leads to a high number of sources being used (see Figure 7.12). This relationship has implications, related to the use of an AS RJ strategy and long duration.

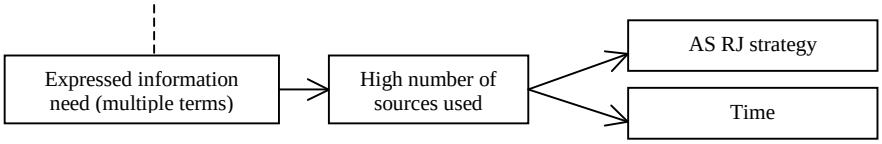


Figure 7.12: Extended relationship of expressed information need

Finally, we found an extended relationship connecting the use of a large number of combined query types (on average) to a large number of document elements being judged for relevance (see Figure 7.13). This can, in the final stages of the task process, affect the type of information components used in the final product.

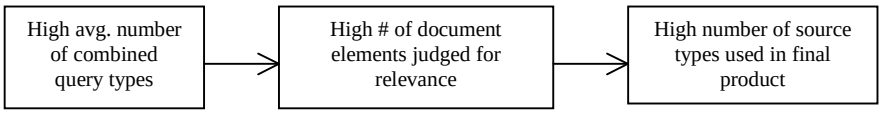


Figure 7.13: Extended relationship of document elements judged for relevance

Summary:
Above are six, different examples of extended relationship dependencies between different variables. Three of them arise from the work task level.

When the patent engineer has low domain knowledge, this may lead to the use of many sources, which, in turn, affects both the relevance judgement strategy and, not surprisingly, the time spent on a task. The complexity of the search increases with the use of several sources. Furthermore, task knowledge may also affect the IS&R process. High task knowledge usually affords a clear information need, which also leads to the application of an aggregated relevance judgement. In these cases, there is no need for relevance judgement during the IS&R process; rather, it is done at the end of the search task. Finally, depending on the applicant (e.g., company or bureau), this affects the clarity or muddledness of planning of the task, which, in its turn, influences the choice of relevance judgement strategies (AS or A strategy) and the number of individual or combined source types used. For example, we found that when the applicant is a patent bureau, the planning of the patent task by the patent engineer is both structured and clear and usually leads to the usage of a single source type for searching.

Information need instability (or change) is connected with the use of a combination of information sources, which, in turn, leads to the use of the AS relevance judgement strategy. Furthermore, when the information need is expressed through multiple terms, it is likely that a high number of sources will be used, which then leads to the use of an AS relevance strategy and is connected to time (longer duration).

Finally, the data show that complex use of query terms (high average number of combinations of query types) in combination with a large number of documents judged for relevance leads to a high number of information sources being used in the final product.

8

COLLABORATIVE SEARCH ACTIVITIES

When observing our patent engineers during the patent handling procedures, we detected collaborative information handling activities. This was a very interesting finding and will be treated separately in this chapter. Put basically, two separate groups of collaborative activities were found: document- and human-related collaborative activities. In our study, document-related collaborative activities were characterised by creation or use of electronic or paper-based documents or objects, such as 'working notes', containing information about a person's search strategies, query terms, and classification codes, made for a specific task. The human-related collaborative activities are characterised by, for example, communication between colleagues internally and externally for advice and expert judgements regarding search strategies and assessment issues. This chapter is concerned with our research question 7: How are collaborative information retrieval activities manifested within and during the course of the IS&R task performance process? The research question is addressed through sections 8.1 to 8.4. Data for these variables were collected with search logs, electronic diaries, and observations.

8.1 Document-related collaborative activities

The first group of collaborative activities involves creating or using electronic or paper-based documents, such as the individually written working notes for specific tasks, with information about a person's search strategies, query terms, and classification codes (see the notes below on Task 111). Furthermore, the working notes may contain information on how a specific search was done in a specific patent case. The notes were written down by the engineer and stored for later perusal by the engineer him- or herself and for colleagues, primarily within the work group but also others within the organisation.

> *(PE) I also take notes on classes, search terms, and document numbers in order to use them later or for others to use them for similar tasks.* [Task 111, diary notes]

The example of Task 111 shows the collaborative aspect in a note that is written down and stored (in paper or electronic form) by the patent engineer. Some of these notes, if found useful, are stored with notes made by other patent engineers in the work team. The working notes can then be reused, by other people, within/outside the work team. The retrieval aspect of the notes taken is that they are both classified and ordered according to subject area and problem description. Colleagues then are able to, for example, use for their own query formulation query terms or search phrases that have been found successful.

We also found other relevant document-related activities, such as the use of reports, reference documents, and notes accompanying the main patent application. In Task 52 (see below), the references in the application are made to serve at least two purposes: a) they are used implicitly as evidence to show the patent office that the applicant has the necessary underlying knowledge of the area in general and the topic of the application specifically and b) they are used as pointers to the patent engineer examining the application in some specific direction or recommending these references as means for further understanding and discussion. The patent engineers then usually search for these pointers when an application arrives.

> *In the incoming patent application, I* [the PE – PH] *looked for referred-to patent documents created by the applicant (in this case, by a patent bureau). I will now collect these documents and read them.* [These documents were then filed along with the patent application[55]; Task 52, diary notes – PH]

The references are a product of earlier searches made by the applicant that are intended to support the formal application and the patent engineer in the problem-solving. If the patent engineer finds the references interesting, the documents are retrieved.

In summary, the document-related collaborative information search activities identified and categorised totalled 100, which gives 8.3 document-based collaborative activities per task.

8.2 Human-related collaborative activities

We found 55 cases of activities in which people used other people's knowledge and experiences during the patent handling process. For example, one may ask colleagues internally and externally for advice and expert judgements. Other examples include asking individuals or groups within or outside one's team or department. The following is a transcribed example:

> *X gets a visit from a colleague. They both discuss what applications X will work on in the coming period. The applications should be within the subject area. Different strategies for searching and sources to be used were discussed* [Task 15, observation note]

The example above illustrates that the patent engineer is involved in synchronous collaboration including for example, specifying relevant information sources, suggesting appropriate search strategies, and sharing experiences.

[55] Author's note: PH = Preben Hansen.

8.3 Collaboration in IS&R processes

We also wanted to investigate *when* collaborative activities occur. Out of the 54 patent handling tasks studied, 12 were observed on the site by the researcher. The rest were reported upon through electronic diaries. When investigating the collaborative processes, we decided to use only *the 12 observed patent handling* tasks. These 12 tasks related roughly to three main task phases (the initial, searching, and completion phases), where the search phase involves both IS and IR. The 12 tasks were observed in different stages and showed the following distribution: we observed 10 tasks in the initial phase, six in the middle phase, and five tasks during the final (completion) phase (for a total of 21 observations). This distribution is explained by the fact that some of the 12 unique tasks were observed during more than one phase.

To visualise the task completion stages in an understandable way, we broke the work task level and the IS&R (IST and IRT) level down into stages. The WT level includes the task initiation stage, with activities of a) task classification and b) task assignment, and the task planning (TP) stage, with the activities of a) structuring of the task and b) formulation of requirements for problem-solving. At the information seeking task and information retrieval task level, the following sub-processes are involved: a) formulation of information needs, b) selection of source(s), c) query formulation, and d) relevance judgement. Finally, at the end of the IS&R task performance process, there is the task completion (TC) stage, involving activities such as information use and creation. The individual activities within the whole process were counted and categorised and then mapped to the corresponding stage in the IS&R process.

Table 8.1: Activities by collaborative category through the IS&R process stages

The IS&R process stage	*Collaboration categories*					
	Document-related		Human-related		Total	
	#	%	#	%	#	%
TI	10	10	4	7	14	9
TP	17	17	16	29	33	21
IST	43	43	24	44	67	43
IRT	17	17	8	15	25	16
TC	13	13	3	5	16	10
Total	100	100	55	100	155	100
n = 155						

Table 8.1 shows that the collaborative activities were observed in all work task stages in the ISR process. In all, 65% of the activities were *document-related*, which implies that the patent engineers experience a need to (re)use document-based information that they or others have created. This suggests that the patent system needs to be designed carefully to facilitate reuse of document-related information. Not surprisingly, the task planning stage shows a fairly high score for collaborative activities (17% document-related events and 29% human-related). Finally, the information seeking stage had the highest score, 43% document-related collaborative events and 44% human-related collaborative events. In order to check the reliability of the categorisation of the data, we performed an intra-categoriser reliability check. Reliability was found to be 95% after re-categorisation of the data three months later.

In the category of *human-related* collaborative information search activities, many activities were observed in the stages of planning the task (TP) and in the information

seeking stage, with percentages of 29% and 44%. Interesting indications emerged in the information retrieval stage, in which, in total, 25 (16%) collaborative information search events were observed for each of the categories. For the IRT stage, this is interesting. All of the tasks observed featured one or both of the two types of collaborative activities. This stage is normally believed to involve individual activities of information processing, but, as can be seen from Table 8.1 (above), collaborative events were frequent. A low number of human-related collaborative information search activities is found in the task completion stage (5%). Even though this score was low, the results indicate that collaborative events occur throughout the information handling process.

Next, we break the data down further, according to the three categories of knowledge proposed earlier, as described by Järvelin and Repo (1983) and further used by Byström and Järvelin (1995) and Järvelin and Wilson (2003): the categories of

a) *Problem-solving knowledge* (PSK), related to how problems should be treated and what knowledge is needed to solve them;
b) *Problem knowledge* (PK), describing the structure and properties of the problem at hand; and, finally,
c) *Domain knowledge* (DK), to do with known facts and theories in the domain of the problem.

A total of 222[56] identified knowledge activities was recorded (see Table 8.2, below). In order to check the reliability of the classification of the data by knowledge type, we performed an intra-classifier reliability check. Reliability was found to be 78–82% after reclassification of the data 2.5 weeks later.

If we look at the task stages, we see that, of the 222 activities, 46% (101) and 36% (80) respectively were performed within the information seeking task and the work task, while 18% (41 activities) were performed at the IR task level. Most of the 222 CIR activities were classified as document-related domain knowledge activities (18%).

The data show that on the work task level, patent engineers acquire both document- and human-based problem-solving knowledge, while domain knowledge is mainly acquired through documents. Problem knowledge collaboration is not needed at this stage, although problem knowledge in general is needed and the patent engineer got it from the patent application at hand. In the information seeking stage, domain knowledge is dominant, in both document- (39) and human-related activities (21), followed by document-related PSK. At the retrieval stage, collaborative activities are mainly document-based. These findings do not refer to the duration of the exchanges.

[56] Because several types of knowledge can be applied during a single collaborative activity, the numbers in Table 8.2 sum to 222 collaborative activities.

Table 8.2: Distribution of document- and human-related collaborative activities across knowledge types and main work task stages (n = 222)

	Work task (WT)		*Information seeking task (IST)*		*Information retrieval task (IRT)*	
	DOC.	HUM.	DOC.	HUM.	DOC.	HUM.
PSK	22	21	22	12	14	4
Total % of all doc.+hum. activities	10%	10%	10%	5%	6%	2%
DK	22	13	39	21	10	3
Total % of all doc.+hum. activities	10%	6%	18%	10%	5%	1%
PK	2	0	5	3	9	1
Total % of all doc.+hum. activities	1%	0%	2%	1%	4%	0,5%
Total	46	34	66	35	33	8
Total (doc.+hum.)	80		101		41	
Total % of all doc.+hum. activities	36%		46%		18%	

If we look more closely at the different knowledge types, we see that the use of document-related problem-knowledge activities increases from the beginning of the work task to the IR task. The reason for this may be that just before the final relevance judgement, different types of documentation are consulted. For activities related to domain knowledge, we find that both document- and human-related activities are greater in number at the IST stage than the WT stage, then drop again in number at the IRT stage.

8.4 Types of collaborative activities

The classification of the various collaborative activities was based on a) a series of predefined types of collaborative activities found in the literature and b) analysis of empirical data and the identification and definition of appropriate categories. In the literature, we found that the description and classification of information sharing suggested by Talja (2002) was useful. The strategic and social sharing were especially important for our study. Furthermore, O'Day and Jeffries (1993b) proposed a set of categories to handle different collaborative activities, such as: handling search requests made by others and archiving potentially useful information in group repositories.

Also, Talja and Hansen (2006) provide a description of a set of main dimensions according to which collaborative information handling can be classified. Some of the categories used in the present study[57]:

- *Asynchronous* and synchronous activities, such as sharing different types of contextual relationships
- *Co-located* and remote collaboration, such as communicating and sharing of personal and subjective opinions
- Loosely and *tightly* coupled activities, such as sharing the history of an information object and sharing search strategies
- *Planned* and unplanned collaboration of types such as work task co-operation

[57] The word in italics is represented by the example.

- *Intra-group or inter-group* collaboration (e.g., workload sharing and division of PA tasks)
- *Direct and indirect* collaboration (such as sharing external and internal domain expertise)
- *Co-ordinated* activities such as end product creation

Classification procedure:
The classification of the different collaborative events was done in a pragmatic way. For each of the work tasks observed, all available data (transcribed observational data) were analysed and marked. Then, each individual collaborative event was grouped with either document-related or human-related events. From these two large groups of descriptions of collaborative events (explicit or implicit, made by the patent engineer), a finer-grained classification was established, as can be seen below. This procedure was then repeated, in order to gather events that were overlooked the first time as well as to validate those already categorised in the first run.

In order for us to reveal the different activities of the collaborative information handling activities responsibly, the 155 occurrences were analysed and the following list of selected sample activities was created:

Document-related collaborative information search activities

- *Searching and sharing information objects/documents* such as articles, patent applications, working notes, and reports. The working notes created by colleagues may be searched by others and may reflect on previous processes involving a specific document and its connections and relationship to other documents and other information retrieval processes (26 instances[58]). Example:

  > *(PH) [The patent engineer writes down a series of search terms on a paper note.] […] these search terms could be used later, in another application task, by me or somebody else in my group.* [Task 15, observation note]

- *Sharing* various types of *contextual relationships* between individual, and sets of, information objects. Expressed means for describing the relationships of documents might be
  - *Annotations*, which are content-based comments, assigned to a document or specific sections of a document;
  - *References* that are assigned to refer to a topical or content-based relationship in a broader sense; or
  - *Citations* that are made to point out relationships to more specific parts of other documents and sections of documents.

  These described relationships are recorded in working notes or the actual patent applications that are filed. These are stored and can later be searched by colleagues (23 instances). Examples:

  > i) *(PE) [W]ithin the application, the applicant is referring to other relevant classes that might be relevant for the present application and should be judged for relevance. I order these documents and check the references.* [Task 86, observation note]

[58] Each collaborative search activity may result in more than one instance, hence the larger number of instances.

ii) *(PH) [The patent engineer writes down a series of search terms on a paper note.] […] these search terms could be used later, in another application task, by me or somebody else in my group.* [Task 15, observation note]

- *Sharing representations of information needs*. The representations of the current information need may be stored and then reused by colleagues. We found two kinds of representation: through classification codes, synonyms, query terms, and query structures and through a narrative description of the problem. Occasionally, these statements and descriptions are saved as working notes and later reused in information seeking and retrieval activities (11 instances).

- *Sharing information seeking and retrieval strategies*. Colleagues can share search strategies in two ways: a search history can be written down by the searcher, to be used in later work tasks, and log statistics can be saved, processed, and inspected for future use. In this category we also found sources used, statistics on time and number of sessions, documents inspected, etc. (21 instances). Example:

  (PE) I write down everything from a search on paper, all possible and used terms, classification codes, document numbers, etc. and combinations of them. This may become valuable in looking back on my previous searches, since an application could be interrupted or it could take longer than anticipated. This document could also be of use for my colleagues. [Task 34, observation note]

- *Sharing decisions and judgements* made for previous problem-solving tasks. Those that are of interest in the current work task can be recorded in working notes or on copies of patent applications that are filed for archival purposes. Other patent engineers can later utilise these working notes (21 instances). Example:

  (PE) [I]f the assessments have been made for this specific problem-solving task, parts of them can be used for a very similar task. [Task 111, diary notes]

- *Communicating and sharing personal and subjective opinions* in written form that, for example, reflect an immediate relationship between the document and its 'neighbourhood' (8 instances).

- *Sharing* the *history of an information object*. The history of an information object/ document may be of three types (36 instances), in the form of any of the following:
  - *Document* history in which the document belongs to one or more subject areas assigned by the PE to the document (on paper and/or electronically) in order to be identified and reused. Paper documents may also bear dates and stamps from previous investigations. Comments and decisions may show history for the document. For electronic documents, annotations in electronic or paper form pointing to the 'source' document may be added. The document then has some sort of history that the current user can use for specific purposes.
  - *Log* history. This history allows all databases and sources, search terms, concepts, query term structure, etc. to be stored and reused implicitly or explicitly by colleagues. They can serve as recommendations or as precise pointers to a problem-solving activity.
  - *Link history*, which refers to the document of interest having links to other documents ('out-links') and links from outside documents ('in-links'). The links made by the PE are related to activities before the relevance judgement of the document is recorded and can be used for

strategy and context-building activities. Also, the actual judgement activity can be recorded, as can information on how the document is used for an end product. Example:

(PH) [The patent engineer writes down number and terms.]
(PE) … [W]hen inspecting the documents from the result list, I write down the document numbers on interesting documents linked to the task at hand, and I am also looking at the reference sections and read the bibliographic information for later use. [PE writes down documents #7, 8, 17, and 27 from the first list. One of these, document #7, is really relevant, and a second document (#27) seems interesting, so the PE downloads them in full text for a closer reading. The PE then makes a 'ranked list' on the paper, based on a first relevance judgement performance.]
[…] Document #27 is also interesting, and I will examine it later in order to see what other document is citing the present one and what document this one is citing.
[…] (PE) When you judge the patentability, you need to look at a long chain of cited documents backwards in the past – that is, the history of a document that is developed through different people's judgements. [Task 15, observation note]

In patent work, identical information needs are rare. However, some overlap is rather common and patent engineers may share information representations where parts of a related information need could be reused. Furthermore, the data show that the activities in this section could be both explicitly and implicitly stated for sharing in collaborative activities.

Human- related collaborative information search activities

- *Work task co-operation*. Sometimes there is a need to share a patent application task, for any of various reasons. This could be done *sequentially* or in *parallel* (5 instances). Example:

  (PE) The current application is a collaborative handling of a complicated patent application. It deals with three specific areas and some of the reasons we share this workload. Before we started, all three of us sat down and discussed how to plan and divide the work and search task. [Task 49, observation note]

- *Work load sharing* and *division* of PA tasks. Colleagues discuss and decide how to divide the incoming patent applications among themselves in the subject group or whether it is necessary to assign the PA to another group in the organisation (6 instances). Examples:

  (PH) A colleague comes into the room and asks the PE I observed whether he would like to take over a patent application case. The reason is that the patent application is more closely connected topically to the other colleague. The PE I am observing inspects the application and accepts the request. [Task 15, observation note]

- Sharing of *search strategies*. The search process/strategy was verbally shared and used in a collaborative way if target documents were closely related. In this class we also find sharing of search terms and classification codes (21 instances).

  (PE) Some of the claims look difficult to assess. I need to consult a colleague. I do not understand some of the chemical figures that are described. […] I just got an answer from my senior colleague, with some clarification regarding relationships between chemical compounds and how new chemical compounds

could be used in new situations, and some hints on how a new, better sub-search could be formulated. [Task 49, observation note]

- Sharing of *external and internal domain expertise*. Patent engineers use both internal and external expertise to help with problem-solving. Colleagues can internally be asked for domain-specific knowledge as well as for information retrieval specificities, while external advice might concern clarification, IPR issues, legislation, etc. (20 instances). Example for external expertise:

  (PE) [S]ome claims in the application were written in a very unclear way, and since the writer of the application was a patent bureau, I will phone and ask them for clarifications of the topics. [Task 86, observation note]

- *End product creation*. In the final phase of a task, patent engineers may collaborate with each other within the group in order to finalise the end product of the task. This may be done through writing a report covering the outcome of the search and its applicability to the stated claims in the PA (4 instances). Example:

  (PE) The single patent document that came through my last search will be filed along with the three other relevant documents in the final report (usually it is regarded as a 'weak' decision when you need to use more than three documents). Two reports will be written. One report will be written in Swedish. […] Each of these four documents will be assigned an 'x', 'y', and/or 'a' judgement. Adding the fourth document, from the last search, could be valuable for colleagues searching similar topics at a later stage. [Task 15, observation note]

- Sharing of *internal experience*. Asking a colleague about experience with similar types of application is a category that also involves issues such as procedural, legal, and strategic issues (12 instances). Examples:

  i) *(PE) [I]t seems that the boundaries of different topics in this application are unclear. I need to consult my colleagues for advice and consensus regarding where the topic boundaries are.* [Task 86, observation note]

  ii) *(PH) [A colleague enters the room and starts talking with the patent engineer about issues regarding classification in a certain field and how to decide which level to begin with in a certain case.]* [Task 86, observation note]

Human collaborative activities involve asking colleagues, both internally and externally, about experiences, and search strategies etc. One collaborative activity that was not present but was expected in the data was the sharing of knowledge about source selection. One obvious reason for the lack of this is that the patent engineers have a rather high level of knowledge about the available sources so know what is there to be used.

O'Day and Jeffries (1993b) proposed four types of sharing of information in collaborative group situations: a) *sharing results* with other team members, b) self-initiated dissemination and broadcasting of interesting information, c) using other people's search requests, and d) storing potentially useful information in repositories for others to use.

The findings described above show that we could use two additional classes that might be integrated usefully into the framework proposed by O'Day and Jeffries (ibid.), below (see Figure 8.1). The difference between the two is that *case-building*

activity focuses on the overall activities related to a specific case, while *history-building* focuses on the various types of traces connected to a specific object

In Figure 8.1, these two new classes are in bold.

- *Case-building activity*: Here, the patent engineer collects, extracts, and adds knowledge such as internal and external documents, notes on external and internal people who have been contacted in this specific case, how the case is filed, and in what contexts this case has been used.
- *History-building for an information object*: In this class is one's collection of all information and *traces* generated during the task performance process. It involves gathering search terms, classification codes, documents viewed, decisions made, relevance judgements made, and how the relevant documents are used in the final reports and the report creation itself.

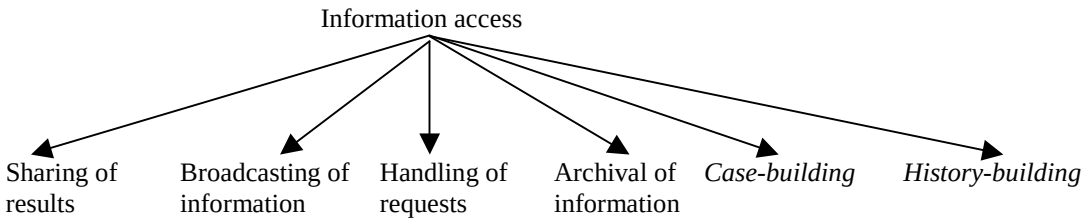


Figure 8.1: Classes of collaborative group activities according to O'Day and Jeffries (1993b), enhanced with two new classes

We now look more closely at the four classes suggested by O'Day and Jeffries (ibid.) and the two additional levels proposed in this thesis, in the context of comparison to the various activities found in our study. For each of the four classes of O'Day and Jeffries, we have extracted the important activities and marked them with '(O)'. Then we added to each class the information-related and human-related collaborative activities found in our study (marked '(H)') and described above, and mapped them to the overall structure suggested by O'Day and Jeffries (see Figure 8.2, below). With respect to the two new classes proposed in this thesis, some of the activities mentioned in classes 1 to 4 may also be valid for classes 5–6 and therefore are repeated since they also have a vital meaning in these contexts. We thus obtain the following categorised activities:

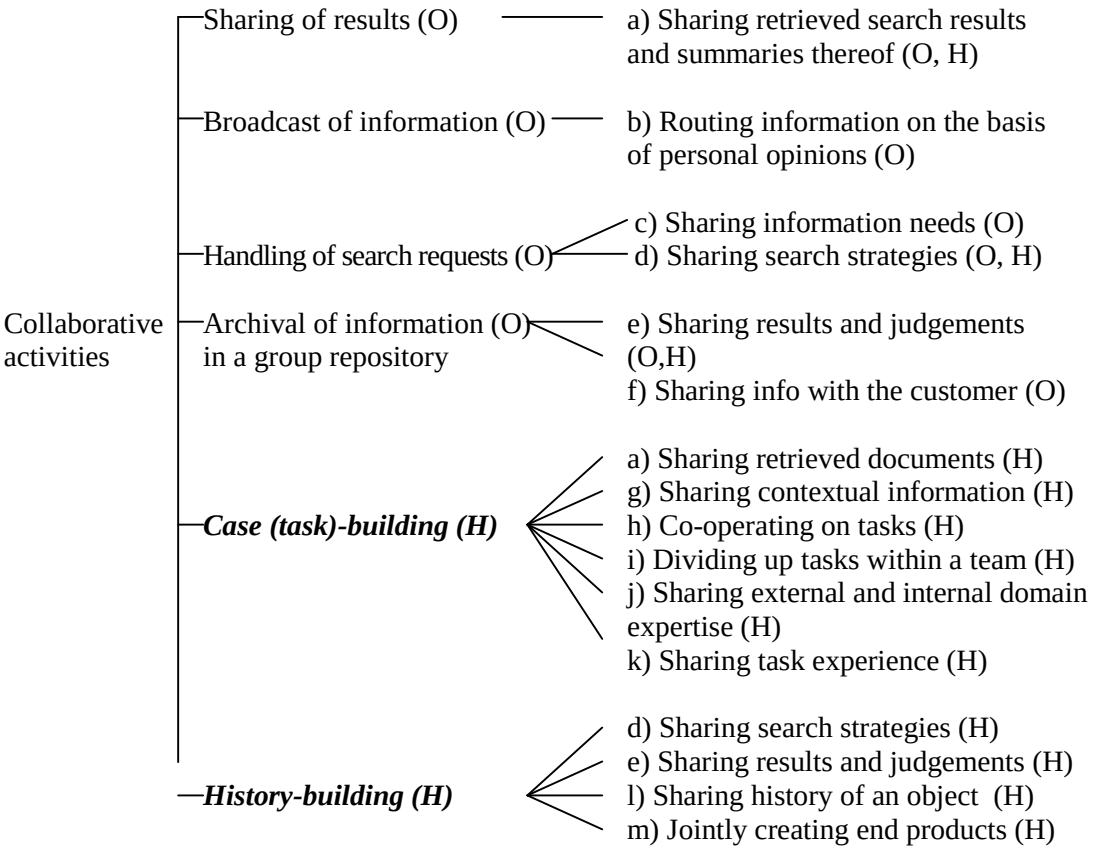


Figure 8.2: Detailed classes of collaborative group activities

Summary of collaborative activities:
In this chapter, we have addressed our research question 7, concerned with the manifestations of collaborative information handling activities during the interactive IS&R task process. We have shown that collaborative information handling activities are found in the general patent work tasks and IS&R processes. More specifically, these are of two types: document- and human-related. It is noteworthy that these collaborative activities belong not only to the information seeking stage but also to the information retrieval stage. Finally, we classified a set of different collaborative information handling activities for each of the two CIR types. Building on the framework of O'Day and Jeffries, we constructed two additional classes of collaborative information handling activities (see Figure 8.1). Further to that, on the basis of the analysis of the data, we extracted additional subclasses of collaborative group activities as depicted in Figure 8.2.

These findings clearly suggest that collaborative information handling activities are part of real-world interactive information retrieval situations in the patent domain and that the findings should be considered in the design of experimental IIR studies as well as when one is designing IIR systems.

9

A METHOD FOR ANALYSING AND DESCRIBING SEARCH SESSIONS IN INTERACTIVE IR

In chapters 6 to 8, we described IS&R as embedded in the work task and stated that there are relationships among various aspects of IS&R in the patent-work process. Our data were collected from the diaries and the attached search log files from patent databases. For every individual patent task, we analysed all log files, step by step and in detail; all of the separate activities; and the sources used, the queries and their reformulation, relevance judgements, documents inspected, and document used in the final products.

In this chapter, we present a method for capturing relevant data that may describe the interactive information retrieval processes. In Section 9.1, we briefly give background on search processes in general. In sections 9.2 and 9.3, we present a method for analysing and describing search sessions. The method has two steps: a) considering how the data from the electronic diaries and log files were used and b) the process of structuring said data into usable sets. The schematic visualisation in Section 9.2 and the schematic diagrams in Section 9.3 are representations of these structured datasets describing session-based information retrieval.

9.1 Search session processes

Marcia Bates (1979a, 1979b) proposed four categories, with 29 search tactics and 17 idea tactics, and suggested that these should be viewed as interactive components in an IR system. The tactics may also be used to analyse search processes and give the search behaviour an explanation. Later, in 1989, Bates proposed a principle or model called ‘berry-picking’. With this model, Bates criticised the traditional IR model as too narrow and neglecting the real user situation, with modelling of too stable information need as one example. Bates was convinced that a real search situation is characterised by the end users usually beginning with one feature and then moving

through a variety of sources. One of the main characteristics was that every time a user encountered a new object of information, this provided new ideas of paths to follow. These new ideas will then affect the next query and the information need. Bates called this an 'evolving search.' (p. 410):

- Search queries are not stable; they evolve
- Relevant documents are gathered in pieces and not from one single query
- Different search techniques are used during the course of the search
- A variety of sources are used (not only bibliographical ones)

Some findings and processes in the present study are also covered by Bates's berry-picking model (1989). However, there are some aspects that emerged from our study that are mentioned by Bates but not described in depth:

a) The source: Our framework, described in the present study, takes multiple information sources into account, not just one source.
b) The information need: The information need is mentioned but not explicitly described by Bates (1989). However, in the conclusions, Bates (ibid., p. 421) states that '[b]ecause information needs change in time and depend on the particular information seeker, systems should be sufficiently flexible to allow the user to adapt the information seeking process'. In our study, we show that the information need may change during the IS&R process. Finally, different needs may also be separately handled in different search sessions within a search task. The relevant document(s) resulting from each information need in each search session will then be merged for a final relevance judgement.
c) The relevance judgement: We discovered that the patent engineer performed at least two types of relevance judgements. In addition to sequential relevance judgement, aggregated relevance judgements were often utilised, separately and in combination.
d) The collaborative information handling: In this study, we found that not only were multiple sources used and multiple search sessions performed, but also the search task may be a collaborative effort.

With this in mind, we now describe a method for capturing information about the search performance on a detailed level.

9.2 A method for describing search processes

On the basis of the analysis of the log statistics and electronic diaries, it was possible to develop a method for extracting data and analytically describe the interactive patent IR processes in more detail. In this section, we describe our method for developing a schematic visualisation of session-based retrieval. Then, Section 9.3 describes the development of a schematic diagram for task-specific session-based search processes.

As has been described in Chapter 4, our work utilised different data collection methods, such as data from observations, electronic diaries (see Appendix D), and log files from each individual patent handling task monitored in our study. The electronic diary was designed to capture all data in a structured manner for later analysis.

The electronic diaries were filled in by the patent engineer and then returned to the researcher on a daily basis. Each diary described all the information seeking and

retrieval activities performed in the course of that day. This ensured that each diary and activity had a date stamp and timestamp. Each patent task could then be followed along a time line (see Figure 9.1, below). Included in the electronic diaries, or separately attached, we acquired all log files for each task. These contained the search logs for the various databases visited during a specific task. For any given patent task, we might receive, for example, five separate diaries with data to be analysed.

From each of the electronic diaries, we extracted information about the person who worked with the patent task, the dates when the PE worked with the task, the information need stated, whether collaborative activities were reported, and how the search task and the patent handling task were completed. The search logs (separately attached or pasted into the electronic diary protocol) included the (timestamped) iterative search sessions, the sources used, and when each of the search terms was used. The protocol also showed the number of relevant documents retrieved, inspected, and selected for each query. If something specific happened during the task performance, the PE made annotations either in the log files or separately in the electronic diary – for example, when and how the relevance judgements were made or how the process of selecting the final relevant set of documents was performed.

The data for each electronic diary, such as source(s), search session(s), information need(s), queries, retrieved document(s) and relevant document(s), and finally the document(s) used, were extracted and mapped into a separate table. From this table we constructed a general schematic visualisation of search sessions (see Figure 9.1) along a timeline. The schematic figure below could be described as depicting a search $task_1$ when one is using a source S_1. There were two search sessions (ss_1 and ss_2), at two specific points in time (t_1 and t_2). In search session 1, three separate queries were made, $q_{1–3}$, without any overlap. However, each query had a separate overlap with the stated information need n_1 (in the diary, in the field 'Information Need'). This tells us that the PE made three separate searches, based on three separate parts of the information need. These three searches resulted in an aggregated result set.

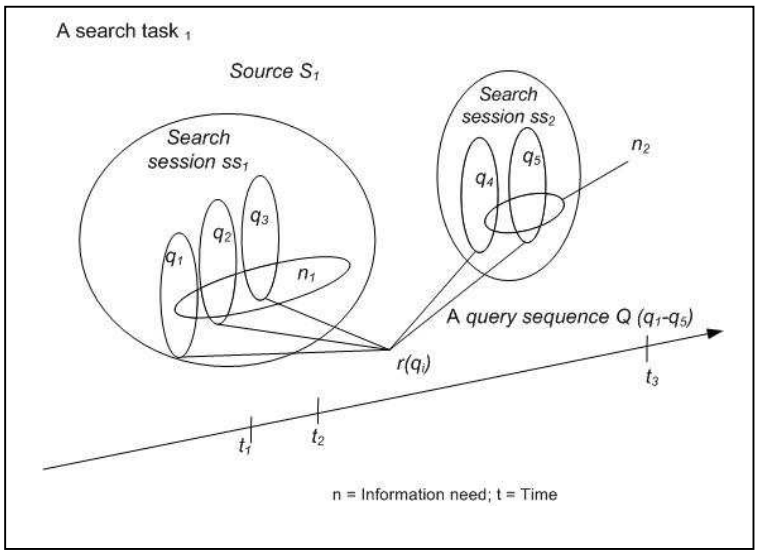


Figure 9.1: Schematic visualisation of a query sequence with two search sessions

Above, a method for capturing and classifying different features of search sessions is described. When using this step-by-step method, we can construct a schematic visualisation of session-based retrieval activities. Next, we present a formal description of session-based retrieval sequences.

A professional work *domain* **(Dom)** (for example, the patent domain) usually has different types of work-related *problems* **(P)** that need to be solved. Specific domains have their own specific problems that need to be solved. Within these specific domains, more or less specific procedures have been developed to address these problems. Different *work tasks* **(WT)** then encapsulate different types of problems. A work task may then involve one or more activities, of which one may be a *search task* **(ST).**

An *information need* (**n**) has to do with a recognised gap of knowledge experienced by the user in work to solve a problem at hand. Furthermore, the information need may, during the search task performance, *change* **($\mathbf{n_c}$)**. The change can be partial or total.

A *source* **(S)** is here a set of information objects from which one or more subsets can be selected. In our study, most of the sources are electronic systems. A source may be *mandatory* **($\mathbf{S_m}$)** or *secondary* **($\mathbf{S_s}$)** for the problem-solving. Mandatory sources are source that are used in every search task, while secondary sources are those needed in a specific situation or for a certain task or as a complement in order to enforce a judgement. The final solution to a problem may be found in either type of sources.

Each source contains *documents* **(O),** which can be of various types (text document, images, or other media). Thus, a *source* **S** is selected and a *search session* **ss** is initiated and a *query* **q** formulated. Several *search sessions* **($\mathbf{ss_{1–n}}$)** and *queries* **($\mathbf{q_{1–n}}$)** may be applied (see Figure 9.1, above, which places the features along a timeline).

For each *query* **q**, a *result set* **r(q)** is retrieved and returned to the user. This is a subset of the *source* **S** used. The returned query result set **r(q)** may be *inspected,* **$\mathbf{r_i(q)}$**, for relevant documents (objects). The inspected document set may then be judged for relevance and yield a *relevant set* **$\mathbf{r_r(q)}$.** For each individual *query* **q**, there is an *information need* **n** related to that query: **n(q).**

The inspected result set may, of course, also contain *non-relevant* **($\mathbf{r_{nr}(q)}$)** and *partially relevant* **($\mathbf{r_{pr}(q)}$)** objects. In our real-world study, one of the non-relevant documents was known since the patent engineer may have encountered it in a previous search task. In a controlled laboratory setting, such conditions can be controlled.

The query may be changed. A subset of the relevant document set might then be a *saved result set,* **($\mathbf{r_s(q)}$**. This subset can be saved and judged for final relevance (sequential relevance) or saved for judgements to be made later in the search process (aggregated). A *sequence of queries* is denoted by **$\mathbf{Q = (q_1, q_2 \ldots q_n)}$**, where **$\mathbf{q_i}$** represents the *individual queries*. The *aggregated result set,* aggregated from each query result, is denoted by **$\mathbf{\cup_i r_s(q_i)}$.** Based on the saved result set, there will be a *used result set,* **$\mathbf{r_u(Q)}$,** a *subset of all saved subsets* **$\mathbf{r_u(Q) \subseteq \cup_i r_s(q_i)}$,** containing the objects that will be utilised for closing of the problem-solving task. For example, a description of the query result, given the *query* **q**, may be expressed as follows:

$\mathbf{r(q) = r_r(q) \cup r_{pr}(q) \cup r_n(q)}$

The following aspects of the retrieved objects can be expressed:
$\mathbf{r_i(q) \subseteq r(q)}$**:** a subset of retrieved objects is inspected
$\mathbf{r_r(q) \subseteq r_i(q)}$**:** a subset of inspected objects is relevant
$\mathbf{r_s(q) \subset r_r(q)}$**:** a subset of relevant objects is saved
$\mathbf{r_u\ (Q) \subseteq \cup_i r_s(q_i)}$**:** a subset of all saved subsets is used

Figure 9.2, below, provides a schematic visualisation of the formally described features along a timeline.

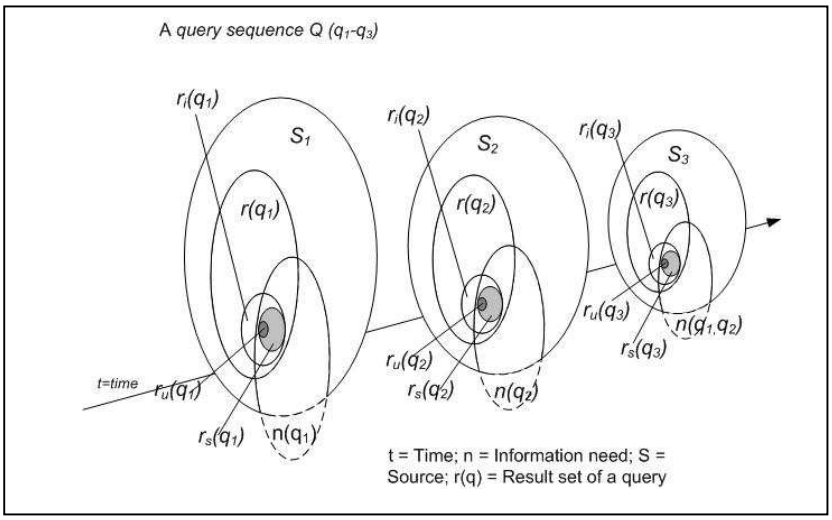


Figure 9.2: Schematic visualisation of a sequence of sessions

We have seen that a *collaborative information retrieval* **(CIR)** activity may take place during the task performance process. A query sequence involving such a CIR activity may be performed either synchronously or asynchronously. In Figure 9.3, below, we see an asynchronous CIR activity that encompasses two different sources and two different information needs (within the same problem-solving task). The figure depicts the use of another person's a) search query and b) inspected result set, the two cases use the same expressed information need but searching of two different sources (S_1 and S_2).

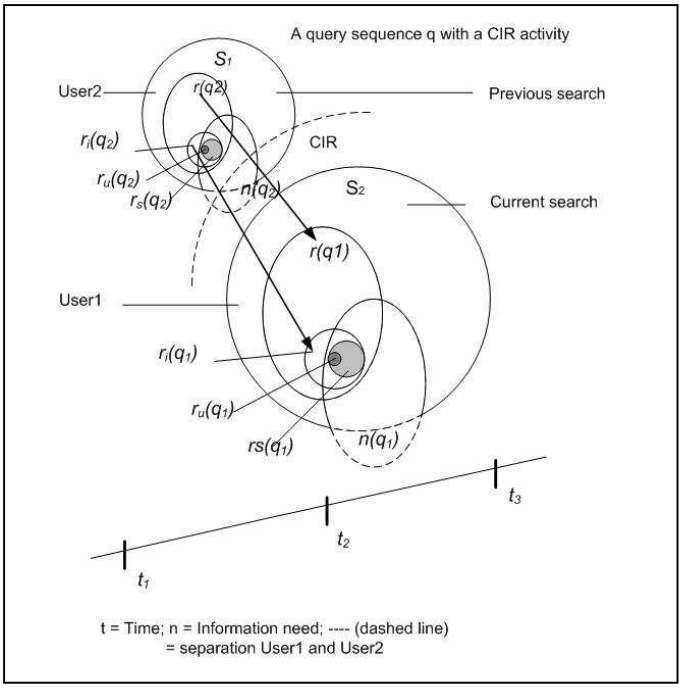


Figure 9.3: Schematic visualisation of a query sequence with CIR activity

Finally, Figure 9.4, below, shows a set of documents that actually will be utilised in the final product of the work task. By this we mean that one or more components from each of these objects have been used in the final product. In the upper circle, *two sets of saved objects* $\mathbf{O_{1-4}}$ have been recognised/identified. The first group of *saved documents*, $\mathbf{r_s(q_1)}$, comprises *documents* $\mathbf{O_1}$ *and* $\mathbf{O_2}$, and the second group, $\mathbf{r_s(q_2)}$, based on the two different queries within the same *source* **S**, contains the documents $\mathbf{O_1}$, $\mathbf{O_2}$, *and* $\mathbf{O_4}$. The lower circle shows the *result set* $\mathbf{r_s(,q_1,q_2)}$ as an aggregated document set and the *set of objects* used O: $\mathbf{r_u(O_1,O_2,O_4)}$, containing documents objects.

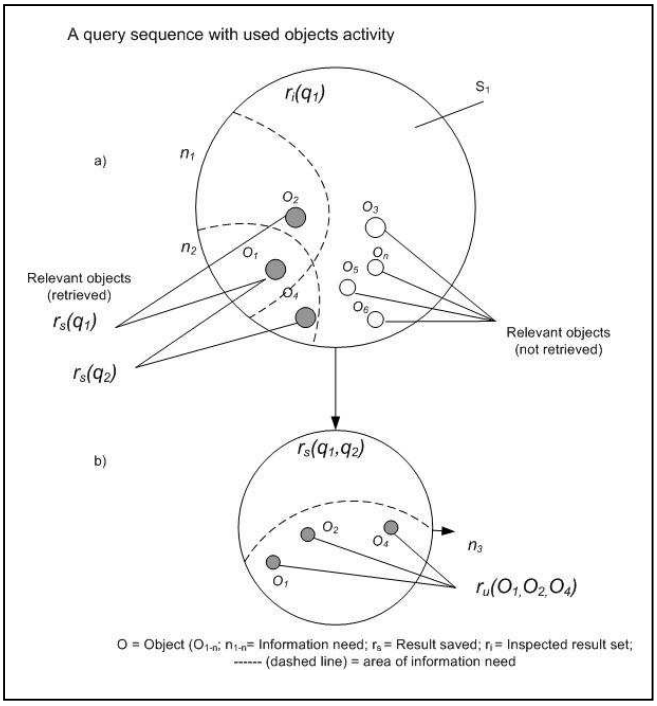


Figure 9.4: Schematic visualisation of a query sequence with used object activity

Summary:
We have described the development of a method for structuring data captured from electronic diaries and log files. Based on these structured data, schematic visualisations were created to represent the session-based search activities. This procedure may be viewed as a mechanism for analysing and describing sequences of actions in an information retrieval process as performed by one or several actors. The basic formal description outlined serves the purpose of describing a query session or a sequence of query sessions during the task performance process and shows important features. The features represent important steps and activities in the patent search process. Each session can be viewed as either a single instance or a set of sequential actions. With this schematic visualisation, the complexity, interactivity, and search dynamics may be teased out. Therefore, the method may be valuable not only for the patent domain but also for describing search processes in other domains. Finally, the features of the search activities could be used in generic analysis of searches, for interactive IR tasks as well as for analysis of real-life search tasks.

9.3 Task-specific search processes

Next, we show examples of how the features of the search activities can be described (see figures 9.5, 9.6, and 9.7) as task-specific session-based retrieval activities. The same data (from diaries and search log files[59]) used to outline the schematic diagrams described in Section 9.2 were used.

On the basis of the three patent search tasks, we identified and defined different features, then placed them on a timeline as they occurred during the real-life task

[59] This list is based on analysis of three two-task processes: 106, 107, and 109. Details are given further on.

performance process. This represents the horizontal axis. Each action was related to a certain time and, in its turn, started a vertical chain of actions. The schematic diagrams starts where the patent engineer interacts with a source. For each source, the interactive information retrieval features of session(s), queries, terms used, relevance feedback, etc. have been classified. This schematic diagram may bring out interesting features of real-life patent search tasks, such as source use; which documents were retrieved, inspected, saved, and used; different relevance judgement strategies; and also collaborative information retrieval.

In our analysis, when the different actions had been outlined, a unique value was attached to each action or event. This value is connected to a specific action in time. That means that, for each action point along the timeline, there is a corresponding feature (for example, q = query) and, in most cases, a numeric value attached to it. That value might, for example, reflect the information behaviour of the PE and how large a specific result set was at a certain point in time (e.g., $r_{(34)}$ meaning that the result set contained 34 documents). Denotation of source S with a numerical value attached represents a unique source; for example, source S_1 may refer to the source INSPEC. Another example is $r_{s(7)}(q)$; seven documents were saved from the result set from query q.

We will now provide some examples of what these schematic diagrams can look like when used systematically.

Example 1 (represented in Figure 9.5) is taken from Task 107 (an A-type task). The features of the information search process are plotted at different levels (or in channels). We can see that the task was performed on two (non-consecutive) days ($D_{(1–2)}$). In this specific case, two days passed between the first and second day of work with the task. We can also see that the majority of the search work was done on the first day and that the second day mainly involved the final judgement.

As can be seen in the example below, the scheme reveals that five distinct sources ($S_{(1–5)}$) were used (source 2 was used three times).

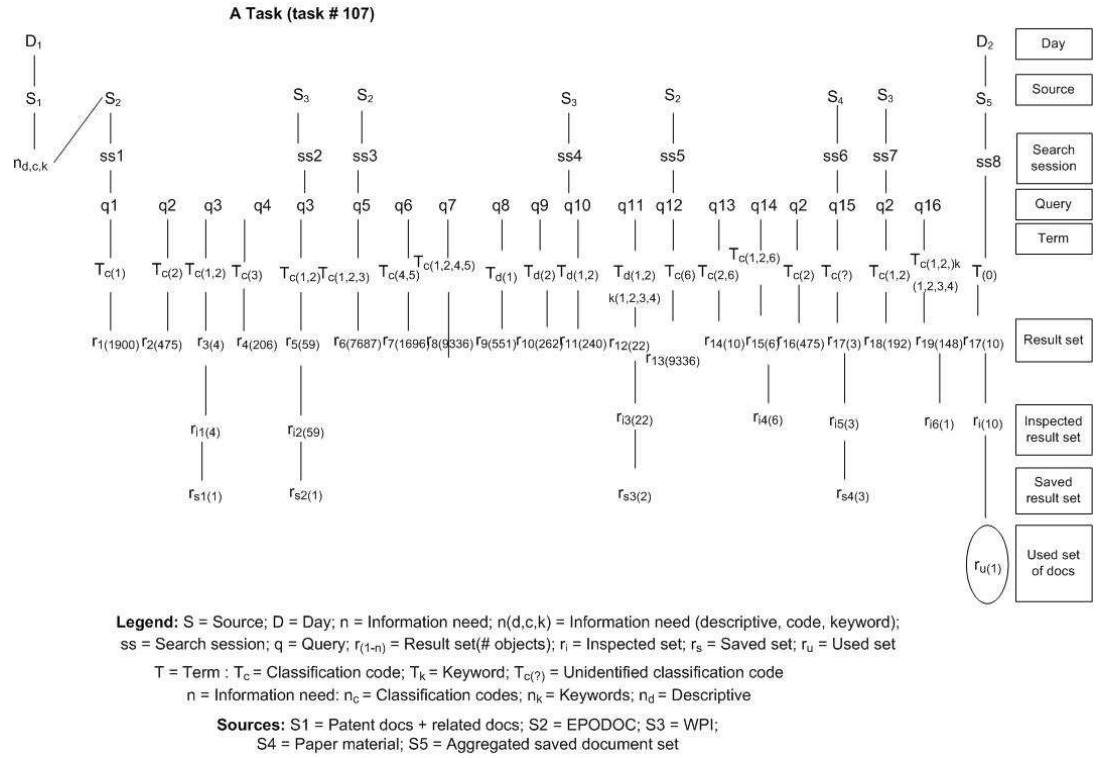


Figure 9.5: Schematic diagram of the process of Task 107 (A task)

The next level shows search sessions (ss), and at the level below we find queries belonging to a specific search session. Note that an identical query can be performed several times but in different search sessions and points in time. For example, query 2 (q2) has been performed three times: in ss1, ss5, and ss7. The query was performed exactly two times with source 2 (S_2) and one time with source 3 (S_3)[60].

For each query (q), various *terms (T)* were used. These could be of different kinds: classification codes, keywords, etc. Each unique term (code or keyword) or combination of unique terms is given a specific number. For example, $T_{k(1)}$ = plastic and $T_{k(2)}$ = bulk. A combination could look like this: $T_{k(1,2)}$ = plastic,bulk. The scheme gives us an overview, and we can, for example, follow that query 1 (q1) was performed in both search session 1 and search session 6 (ss1, ss6). Furthermore, the user applied a large number of terms in the query (in our case, in q11 and q16). This scheme could be used to illuminate, for example, how many times a certain term was used and when in the search process. In our example, term 4, $T_{k(4)}$, was used with source S3, in search sessions ss4 and ss7, and in queries q11 and q16.

At the level of query results, we find one level depicting each result set, including the number of retrieved documents – for example, $r_{2(475)}$, referring to the second result set, containing 475 retrieved documents. The next level shows how many of these were inspected, $r_{i(...)}$, and, at the next saved result set level, we can see the numbers of documents saved. For example, $r_{s1(1)}$ means that one document was saved in the first

[60] An agreement between PRV and the researcher prevents us from revealing details from search logs.

saved set of documents. Finally, there is the level of documents actually used, $r_{u(1–n)}$ for the final task completion.

The foregoing description made it possible to read the schematic visualisation horizontally. However, the search process could be described equally well in vertical terms. For example, in Figure 9.5, we see that at the point when source 3 (S3) was used, during the second search session (ss2), the patent engineer performed the third query, q3. In this query, two classification codes were used ($T_{c(1,2)}$). The query resulted in result set number 5 ($r_{5(59)}$), of 17 result sets in total. This result set contained 59 documents. In this result set, the second of seven, all of the documents retrieved were inspected ($r_{i5(59)}$). In the second saved result set, one document was saved ($r_{s2(1)}$).

In yet another example, Task 106's parts 1 and 2 (a PCT1 task), depicted in figures 9.6 and 9.7, we see a search process that is more complex and of a larger magnitude. This example includes a CIR activity. It also displays more variation and a certain degree of intensity in, for example, how often sources are changed and the ways in which terms or classification codes are repeated and combined with one another.

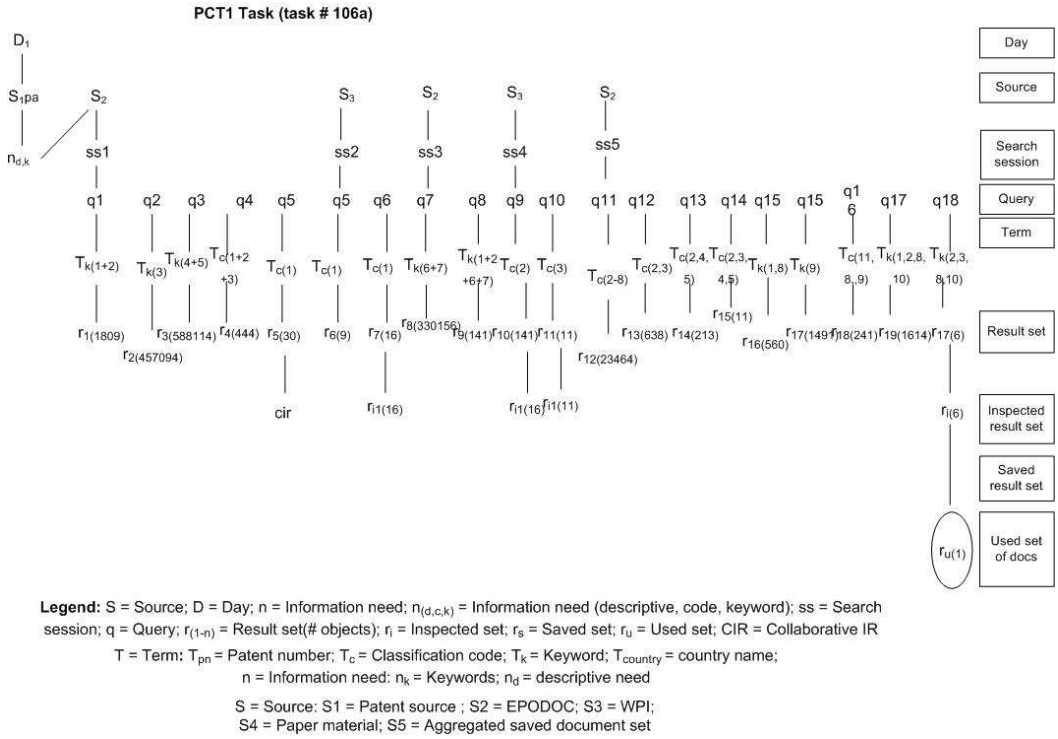


Figure 9.6: Schematic diagram of the process of Task 106 (PCT1 task) – part 1

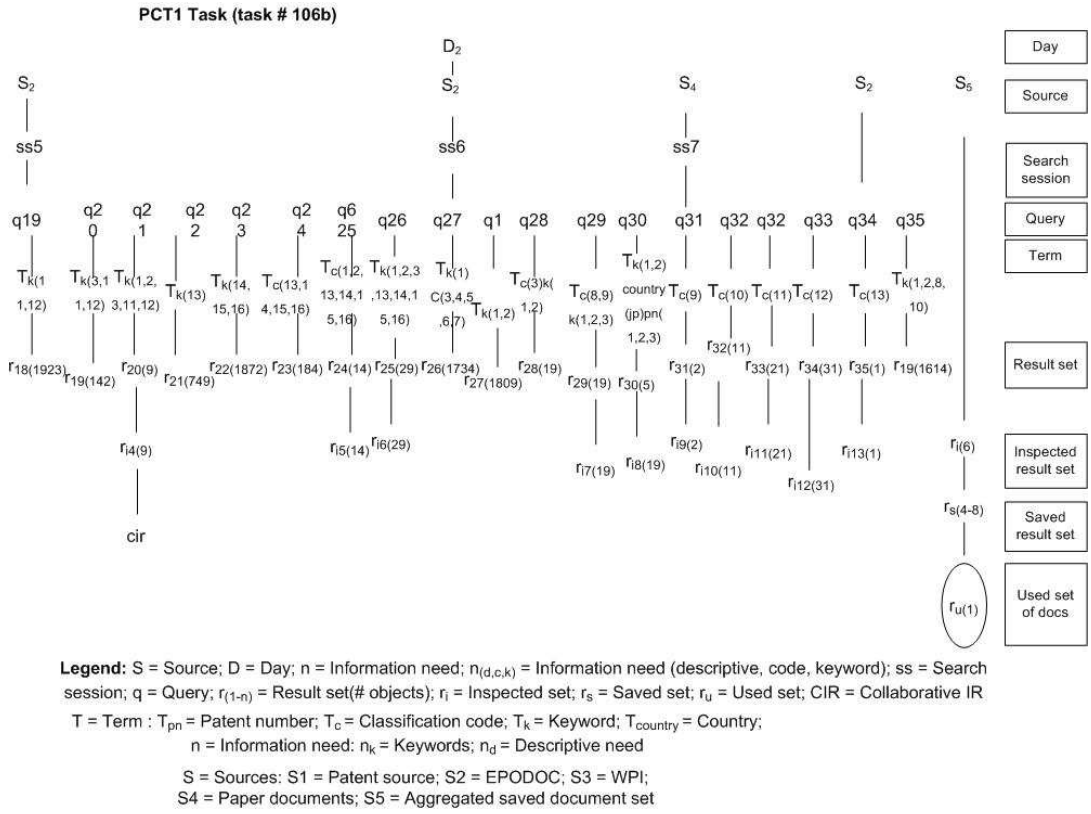


Figure 9.7: Schematic diagram of the process of Task 106 (PCT1 task) – part 2

We now present an example of how these features can be used in a comparative and systematic way (see Table 9.1). We take this method one step further, taking the features identified and marked in the scheme and systematically comparing data from three different task processes. The way in which the features are marked in the scheme makes it easy to collect them and tabulate them for comparison. In Table 9.1, we have selected data from three search tasks, representing different search processes: T106, T107, and T109.

Table 9.1: Comparison of data for three task processes: 106, 107, and 109

	106	**107**	**109**
Task type	**PCT1 task**	**A task**	**AS task**
Days	2	2	2
Sources	5	5	5
Search sessions	7	8	5
Queries	35	16	18
T_c = Term (code)	13	6	14
T_d = Term (publication date)	2	2	
T_k = Term (keyword)	16	4	4
Country	1		
r = Result sets	36	18	18
r_i = Result set inspected	15	6 (4+59+22+6+3+1)	5 (33+19+57+21+1)
r_s = Result set saved	12	4 (1+1+2+3)	3 (7+1+3)
r_u = Result set used	1 (1)	1 (1)	1 (4)
CIR = Collaborative search activities	2	0	1

All three search processes were completed in two days. All of them also used five sources. The number of search sessions varied between five and eight. Although tasks 107 and 109 feature different numbers of queries (16 and 18, respectively), they used the same number of keywords and differ with respect to the use of classification codes in the queries (6 and 14, respectively). Another difference can be seen in the use of terms (keywords). Task 106 used 16 unique keywords (in 35 queries), while the other two tasks each used four keywords only (in 16 and 18 queries, respectively).

In the next part of Table 9.1, above, we find the units that describe the retrieval part of the process. We see that Task 106 has 36 different result sets, in contrast to the 18 for each of tasks 107 and 109. We can also count the number of documents in the result sets inspected, saved, and used. For example, in Task 107, we can see that the patent engineer inspected, in six distinct result sets, the following numbers of documents: 4, 59, 22, 6, 3, and 1. The situation for task 109 is as follows: 33, 19, 57, 21, 1 for the five result sets inspected.

So what are the benefits of these two ways to analyse and represent data? The general scheme with the features across two axes and different levels visualising the process has the benefits of

- Showing how many days were used for task completion and when the different activities were performed,
- Displaying the sources visited and revisited and when in the process they were visited,
- Showing the number of search sessions and how many there were in relation to different sources,
- Showing queries across several sources,
- Revealing how queries and relevance judgements progress and evolve,
- Showing single events and their context both vertically and horizontally, and
- Showing the actual and final use of document as a result of the search process,

while a summary table may show

- Numerical comparisons between a large set of schemata,
- Patterns and anomalies within large numbers of search processes, and
- Patterns within and between domains.

Summary:
This section addresses the second research problem – methodology. We have presented and specified a task-specific session-based scheme of information retrieval action with two dimensions. This scheme allows us to follow the IR features and task performance both as query sequences, horizontally, and as per-query actions, vertically. We can see not only how a search process proceeds but also the state for a specific position and how it connects to related actions.

The depiction of the search processes in figures 9.5, 9.6, and 9.7 may be used for recording, extracting, and constructing an index – e.g., for language analysis and statistical analysis – of terms (e.g., classification codes or keywords). We can also see when and how terms were applied at different stages in the search process.

Furthermore, the different levels of result sets, inspected sets, saved sets, and sets used, with their individual values, may actually be considered as different instantiations or parts of the relevance judgement process. For example, the inspected result set is a subset of the retrieved result set. At the level of an inspected document set, the criteria for judging may have been based on the number of documents retrieved (for example, a retrieved set of 1,000 documents may not lead to inspection while 75 documents might). At the saved document set level, the criteria for saving of a document for further perusal may be more content-based than at the previous level. Finally, in the used document set, the documents have undergone thorough inspection and those saved reflect the most appropriate documents for completion of the task.

The schematic visualisation and the schematic diagrams offer different advantages for recording of data. The schematic visualisation has the advantage of showing the details of a single query sequence. Each event is shown in its context. Tabulation of several search processes (as in Table 9.1) has the advantage of revealing patterns, tendencies, averages, and percentages, giving support for statistical measurements.

Another advantage with the schematic diagram is that by utilising the skeleton described, one can see how better to collect data. For example, it could give guidance in what to look for, and how, when designing collection of log statistics and when inspecting a log file. The main purpose would be to guide the collection and capturing of data.

When one uses the schematic diagrams, it is possible to detect and pinpoint unusual patterns, anomalies, and query characteristics and so gain better insight into, and understanding of, information search processes. Finally, it is our belief that these schematic visualisations and diagrams for visualising the relevant features can be utilised not only within the patent field but also in other domains.

CLOSING

10

DISCUSSION AND CONCLUSIONS

The research described in this thesis investigated IS&R processes in a professional patent office work setting. The investigation focused on the relationship of work tasks and IS&R tasks in order to describe patent IR processes.

The overall research question about the *effects of work task features on the information seeking and retrieval process in the patent domain* has been addressed through answering a series of sub-questions.

In Section 10.1, the general research question (see above) is discussed and we categorise a set of descriptive features of the IS&R processes embedded in patent work tasks. The features are related to the first six sub-questions. Based on these descriptive features, a general framework for patent IS&R has been outlined (Section 10.2). In Section 10.3, six (of the seven) sub-question are addressed and the relationships between the features of the IS&R process are discussed. In Section 10.4, collaborative information search is identified and the seventh sub-question is addressed. Empirical results show that the patent IS&R task process involved highly collaborative activities throughout the task stages. IR processes have been described through development of a scheme for capturing features in search processes. In Section 10.5, the second research problem of methodology is discussed. Our first concern was to utilise a combination of data and analysis collection methods. Secondly, we describe a methodology for analysing the data of the task-based PIR studies and modelling session-based information retrieval. In addition, schematic visualisations and schematic diagrams provide examples of its application. Section 10.6 deals with limitations of this thesis, followed by conclusions (Section 10.7). Finally, in Section 10.8, we discuss future work. The specific empirical findings are discussed in the following sections.

10.1 The patent domain and patent IS&R phenomena

In order to answer our first main question, concerning '*effects of work task features on the information seeking and retrieval process in the patent domain*', we performed

in-depth exploration of the IS&R conditions within real-life patent work tasks. This empirical exploration included a description and classification of features of the patent IS&R process. We describe the conditions for real-life patent IS&R below.

At the work task level, we found goals related to the work tasks at three different levels: the organisational, the group, and the individual level. However, we decided not to integrate these into the study. It will, however, be interesting for future studies to investigate further how these types of goals may affect search activities.

The patent handling process involved six formal types of patent tasks, each with its own characteristics. The preparation of patent applications also had different backgrounds: they were private persons, skilled people from patents bureaux, and companies' own patent departments, which affects the content of the applications in a variety of ways.

As expected, constraints of different kinds arose during the patent IS&R task performance, and we established a list of work-related constraints, but we did not further pursue that path. We did find, importantly, that the duration of the IS&R task considerably exceeded the usual time frame given to users in experimental or simulated studies of IS&R.

Related to the first sub-question, the work task features of the engineer's perception of the task difficulty and task knowledge not only showed a binary value of high/low knowledge but could also involve perception of a task as both easy and difficult, depending on specific parts of the patent task handling process. This is another indication of a complexity inherent in real-world IS&R tasks (Byström & Järvelin, 1995). With regard to domain knowledge, we detected that the patent engineers perceived the domain knowledge needed as residing either inside or outside their domain of knowledge and, most importantly, it could also be found in both areas, partially outside. The reason for this could be that some parts of the 'problem area' are better known to the patent engineer than other parts, in view of the complexity of the patent application itself. The idea of different levels of knowledge was, in general, confirmed (Järvelin & Repo, 1983; Byström & Järvelin, 1995; Järvelin & Wilson, 2003). For example, Järvelin and Repo (1983) suggested three categories of knowledge: problem-solving knowledge, problem knowledge, and domain knowledge.

We found that the work task (including the IS&R process) was structured or unstructured as part of the task process and that it could be considered the first step in the information seeking stage, since it involved a combination of guidelines, legal aspects, work processes experienced, planning for the use of specific relevant sources and appropriate reference material, and personal experience. Moreover, problem formulation was identified as a natural aspect of the IS&R process in real-life patent handling, confirming, for example, findings by Kuhlthau (1991) and Byström and Järvelin (1995).

Our second sub-question concerned the decomposition and formulation of the information need. At the IS&R task performance level, different aspects of information need were found (see Figure 10.1, below). For example, the perceived information need was categorised as structured/unstructured and as clear/unclear. The

information need did change in the course of the patent handling process, which also confirms what Bates (1989) and Kuhlthau (1991) suggest.

Another feature of the information need was that it was occasionally broken down. It was itemised with regard to task performance procedure and, as well as to the formulation of the information need.

The information need sometime was expressed as a narrative as well as with individual terms/codes.

Further, an information need may be expressed as either a single need or a set of (multiple) needs. Multiple needs could be viewed as one criterion for complexity of the task (Byström & Järvelin, 1995).

Finally, in patent information search, specific document components, such as sections (e.g., abstract, description, and claims), references, terms, classification codes, and images, were pointed out as important and as required for the formulation of the information need.

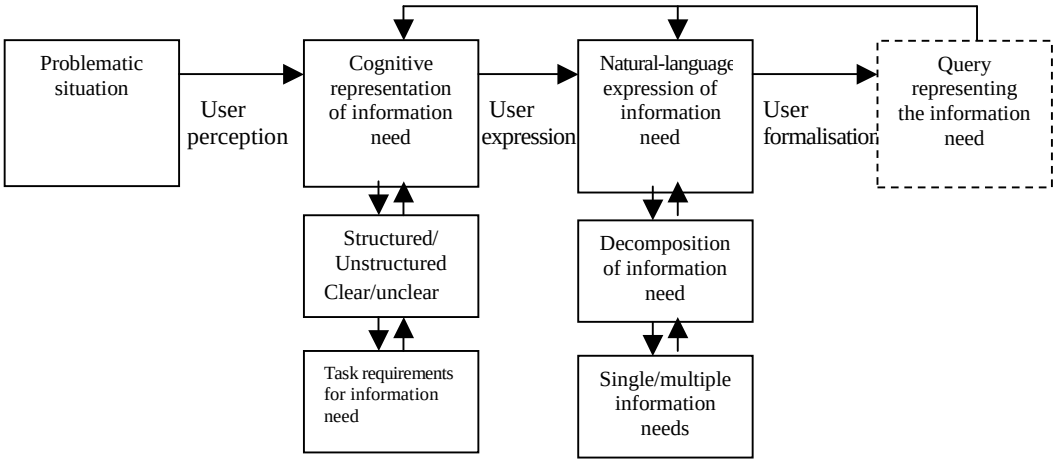


Figure 10.1: The information need process

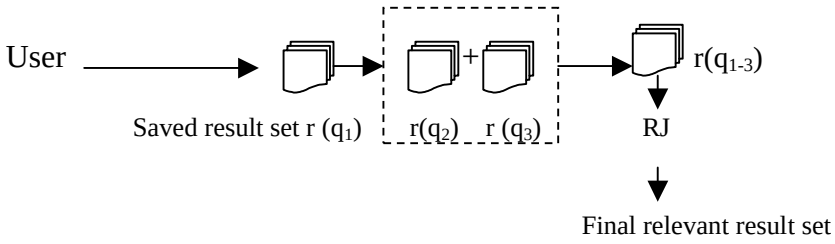
Thus information needs have several aspects, in confirmation of previous work (Taylor, 1968; Mizzaro, 1998). First, the patent application, when facing the patent engineer, may lead to a 'problematic situation' (Belkin et al., 1982a, 1982b) in which the patent engineer needs to identify and solve a problem (Taylor's 'visceral need' and Mizzaro's 'real information need', RIN). The user then *perceives* the information need and creates a *representation* of the problematic situation (Taylor's 'conscious need' and Mizzaro's 'PIN'). This is mainly a cognitive process. As noted above, this study found that the perceived information need could be structured/unstructured and clear/unclear. Thirdly, the perceived information need can then be expressed (Taylor's 'formalised need' and Mizzaro's 'request') in narrative form or as terms, usually in natural language. In addition to these confirmatory findings, we found that the information need could be expressed in relation to topics (structuring) and source selection. Furthermore, the information need might be broken down and expressed as single or multiple needs. Subsequently, the information need expressed is formalised, as a query representing the information need.

The sources used (related to our third sub-question) in this study are those used in the patent engineers' normal professional work tasks. We found that both multiple sources and different types of sources were used. This clearly shows the complexity involved in a real-life IS&R work task process (Byström & Järvelin, 1995). The patent engineers were also involved in reading, handling, interpreting, and assessing different types of content in these sources of different types and numbers. This also contrasts against the general assumption, often applied in laboratory IR experiments (e.g., TREC), that searching involves only one source.

Query formulation (addressed in the fourth sub-question) involves several findings, and we identified patent engineers as using a variety of types of query elements such as keywords, codes, dates, document ID, and country codes. Furthermore, it was common that both terms and classification codes co-existed in the search sessions and this may point to a multifaceted search approach in the patent domain. This handling of a variety of search keys makes the query formulation a complex task. Finally, on average, three to eight unique search terms were used in a problem-solving task. Jansen, Spink, and Saracevic (2000) studied Web-based user queries and classified the queries as unique, modified, or identical. In their study, the first query by a user was always a unique query. On average, a query contained 2.21 terms. We found that the average number of terms used per session and query string varied between 1.56 and 8.58 terms, depending on the type of task. In our study, unique search terms were identified as the number of unique terms used once or more during a search session.

Related to our fifth sub-question, interesting results were found in the category of relevance judgement. Relevance judgement was actually performed with two, quite different assessment strategies: sequential relevance judgement strategy or/and an aggregated relevance judgement strategy (see Figure 10.2, below).

a) Sequential RJ strategy



b) Aggregated RJ strategy

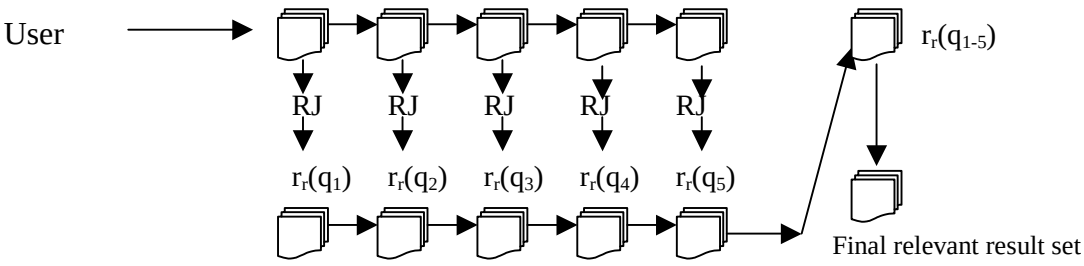


Figure 10.2: Illustration of relevance judgement strategies

This finding suggests that relevance judgement is utilised in a more complex and varied way and not viewed as one single action at the end of the search session. It was also established that the aggregated strategy involved three different approaches to how documents are saved and stored when found relevant.

In the present study, we distinguish between the relevance judgement strategy described above and the relevance judgement for a single document. Usually, an object is judged as accepted or rejected (Spink & Greisdorf, 2001) in a binary decision. In the patent domain, the relevance assessment is made on a 'graded' scale of predefined criteria for relevance. These criteria are closely connected to the domain, the task, and the content of the information. This confirms previous findings on degrees of relevance (Wilson, 1973), partial relevance (Saracevic et al., 1988), regions of relevance (Spink & Greisdorf, 2001), and graded relevance (Kekäläinen & Järvelin, 2002b, Kekäläinen, 2005). It is important to consider that not just a single document or object as a whole is judged for relevance. Specific elements/components and parts of the document are considered differently with regard to importance (that is, where the relevant information might be found). We established a list of such elements and components.

Finally, closely connected to the sixth research question is the information use to allow completion of a task. In the patent domain, the task completion stage or the end result (a report or statement/acknowledgement regarding the outcome of the review of a patent application) is closely related to the work material used during the task performance process. Examples of such types of material used in the final product are selected paragraphs, terms, classification codes and other relevant sections from prior art applications, and similar components from technical documents. This confirms the importance of considering the completion stage in a search and work task, as pointed out by Vakkari (2003) and Savolainen (2009).

10.2 A framework for patent IS&R

From the empirical findings described above, we constructed a conceptual framework for the patent handling process (see Figure 10.3, below). This framework divides the patent task process into several stages. Each stage has a set of sub-processes. Between the single actor and groups of actors are interconnected with different experiences and knowledge types that affect the actor (and group of actors). The experience and knowledge levels are not stable entities but constantly changing. In a similar way, the information need may change, depending on the task and problem at hand, as well as be affected by the experiences and knowledge of the actor or group of actors. The *task initiation stage* involves registration and the delegation and assignment of tasks within the workplace and group, and the task preparation stage encompasses activities such as reviewing a new or revised PA (i.e., a patent application sent back to the applicant by the PE after a first reading/revision with suggestions of enhancements and updates to the first version), planning and structuring the task, and formulating different requirements specific to that task. The IS&R stages include activities such as collecting appropriate information before searching, formulating the information need, and selecting one or more sources to be used. The information retrieval stage includes engaging with one or more electronic sources; query formulation and reformulation; and, finally, relevance judgement. Activities in the task completion

stage might include compiling and extracting relevant information and making a final decision or suggesting a solution. Finally, we have the stage of information outcome – the activities of using the information retrieved and judged to be relevant: external (e.g., in reports) and internal information creation. The arrows in the figure show that the individual stages within the framework are interconnected to each other.

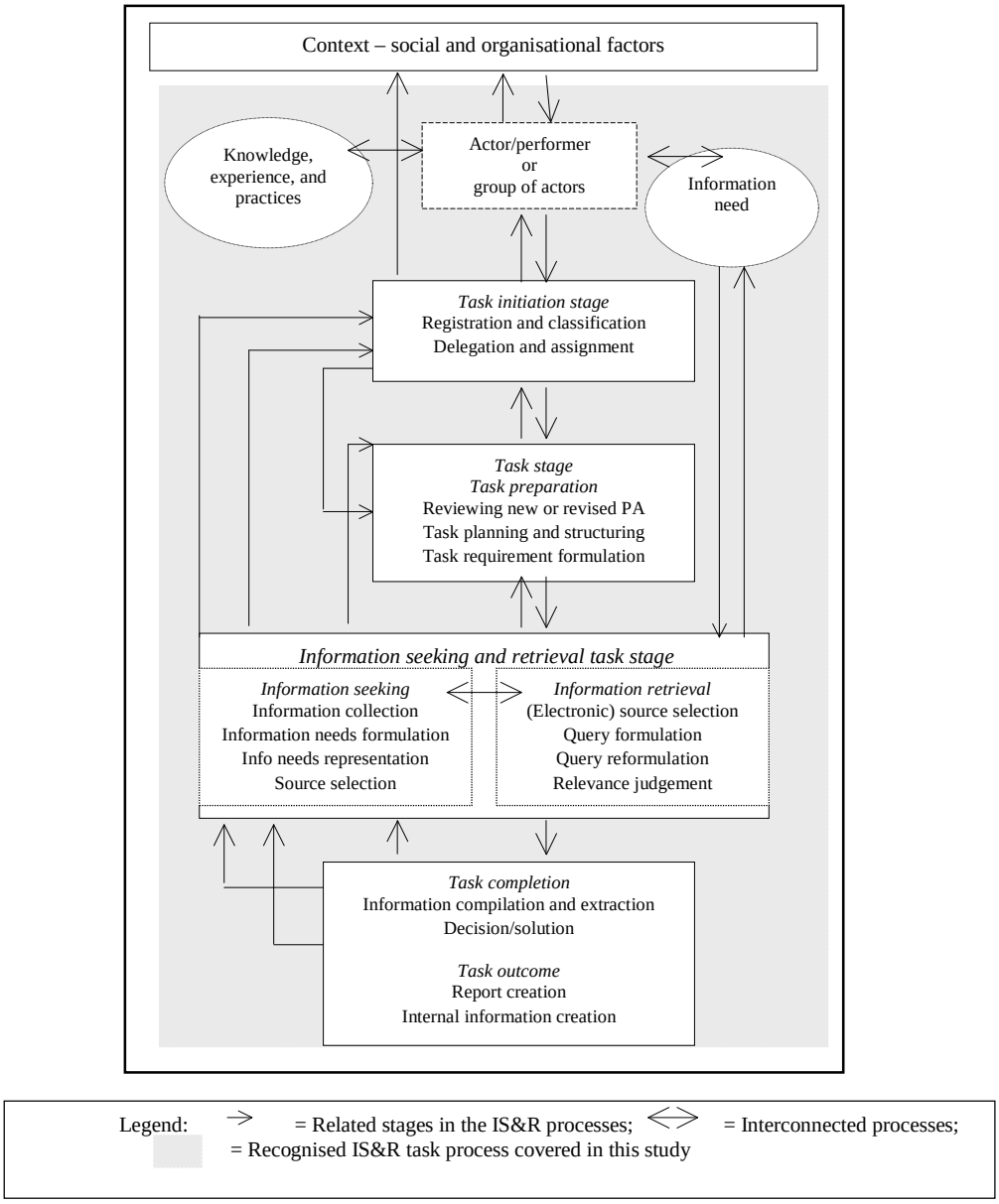


Figure 10.3: Framework for the patent handling process

This framework serves as the setting in which the further examination will take place. Relationships and processes and the effects will now be discussed.

10.3 Relationships in patent IS&R

This section examines the relationships between variables at the three main task levels described in the section above. Significant relationships and dependencies were demonstrated in a set of variables within the levels of work tasks and IS&R tasks.

The work task category (addressed in research question 1) involves *the effects of work task characteristics on work task approach*. We studied this problem by correlating variables such as task structuring, user knowledge of task topic and perceived task difficulty, task constraints, and completion time.

Task planning may be clear and/or structured. When only clear planning of the patent task was done, the patent application was often submitted by a company, while clear and structured planning of the patent task was done often when the PA was submitted by a patent bureau.

Patent applications submitted by companies usually correlated with formulation of multiple information needs by the patent engineer, use of a combination of source types, and the use of a high number of relevance judgements during the IS&R process. In turn, handling of patent applications that are subject to both clear and structured task planning (often submitted by bureaux) often involves the use of single sources and results in a low number of relevance judgements. These tasks often have a short duration. Cosijn (2006) found that the type of work task (in that study, academic school assignments) had a significant influence on the type of relevance judgement made. Task planning may involve knowledge of what information is needed if one is to structure and build an appropriate strategy for executing the task; therefore, the planning of the task itself is important and will affect the process.

Not very surprisingly, *time* (considered in hours) obviously has an effect on RJs. When many sources were used and when the aggregated relevance strategy was used, the task took longer to perform. One reason for this may be that the searcher is learning during the search process or that there is more information to interpret and handle.

High *domain knowledge* leads to an aggregated relevance assessment strategy, while a low level of domain knowledge usually was linked to a combined (aggregated and sequential) relevance assessment strategy. This confirms similar research on the relationship between prior knowledge and numbers of objects judged for relevance (e.g., Sping and Greisdorf, 2001; Vakkari & Hakala, 2000). Furthermore, when the *task knowledge* is outside the user's field of knowledge, use of a large number of sources is commonplace, as is the use of a low number of terms and query strings per session, on average. So uncertainty leads to careful use of search terms (that might be caused by difficulties in finding appropriate search terms) but also exploration of a large number of sources; therefore, availability of a larger number of sources is needed, to support term suggestions / query expansion such as synonyms (e.g., Efthimiadis, 2000).

The information need category (the subject of research question 2) was related to *the effects of different task characteristics on information need formulation*. This problem was studied via variables such as perceived information need, information need

structuring, expressed information need (representation thereof) and the number of expressed terms and search expressions, and information need change. Three instances of information need affect other stages in the patent search process and are discussed below, in the following order: perceived information need, expressed information need, change of information need.

First, a muddled perceived information need (in terms of specification of what is probably needed in terms of types of sources to visit, PA document components to inspect, or document components such as images to be inspected) leads to use of a high number of search strings.

Secondly, the expressed information need showed significant relationships. Prior to the search, the information need may be expressed either as a single or as multiple terms / keywords / classification codes. We found a significant relationship between information need expressed with a single term / classification code and the use of a small number of sources. Multiple source types are used when the information need is expressed with multiple terms / classification codes. Further to that, task structuring affects the information need expressed. A clear approach to task planning often leads to the use of multiple expressed information need.

Thirdly, during the search process, there might emerge a change in the information need. When such change occurs, it usually results in the use of a single source type, such as only electronic information sources, while a stable information need leads to use of a combination of source types. Change in the information need also leads to the use of a low number of search terms / keywords. This may be because the uncertainty of the information need limits the actor to being more elaborate when constructing the queries. If the information need changes, one uses few unique classification codes (1–3), while a stable information need encourages use of many unique classification codes.

The category ‘source selection’ (see research question 3) was related to *the effects of task characteristics on the types of sources and source content*.

The main aspects of this were the type of source and the number of sources. A combination of different source types leads to a combined RJ strategy, and usage of a large number of sources leads to a long duration, the use of combined RJ strategies, and a high number of relevance judgements within each patent task.

The category ‘query formulation’ (the subject of research question 4) had to do with *the effects of task characteristics on query formulation*. Three aspects of the query formulation showed relationships: terms used per string and session, the number of query terms, and the average number of combinations of query types.

Some correlations have already been reported; for example, a patent engineer who has insufficient knowledge of the topic often uses a low number of terms per string/session, on average. When information needs change, usually one finds a low number of query terms being used. Additional relationships will be reported on below, in the relevance judgement section.

The category 'relevance judgement' (see research question 5) was related to *the effects of task characteristics on relevance judgement performance*. Some dependencies have already been mentioned, and others will follow below.

A very important finding was that we were able to identify relevance judgement strategies and that the choice of RJ strategy correlates with the use of different source types. A combination of source types shows a correlation with a combination of RJ strategies, while aggregated RJ correlates with usage of a single source. The reason for this may be that the patent engineer visited the specific source in question several times (i.e., in several instances of opening and closing), with each time representing a specific aspect of the information need. This could result in several separate relevance judgement sessions. Combination of RJ strategies is related to the number of expressed information needs. Also, when a search process involves multiple information needs, they often occur in conjunction with a combination of RJ strategies. The type of relevance strategy – i.e., aggregated and/or sequential – also shows a weak correlation with the number of sources used. Use of many sources is connected with utilisation of a combined RJ strategy. This finding is a contribution in line with a problem Vakkari (2000b) discussed, which relates to how relevance evaluations change during task performance. Finally, a relationship was found between the RJ strategy applied and the content types judged for relevance. When a combined RJ strategy is utilised, terms and abstracts are important content types, while codes and abstracts are important when only the aggregated RJ strategy is used.

It is not only the documents that are judged for relevance but also their individual elements. We found dependencies between the number of elements judged for relevance and the number of sources, the number of sources used in the final product, and the combination of types of query items. Finally, and not surprisingly, RJ has an effect on *time* (hours spent on a task).

When the PE used a large number of sources and when the aggregated relevance strategy was used, the task took longer to perform. One reason for this may be that the searcher is learning during the search process or that there is more information to interpret and handle.

The information use (research question 6) category was created to address *the effects of task characteristics on use of information in order to complete the task*. The only significant dependency relationship with information use was identified between the average number of elements judged for relevance and the number of information sources used for the final product. When two information sources are used for finalising of the final product (an application report), 3–8 elements in the documents, on average, are judged for relevance. This finding is in line with the result from a study by Huuskonen and Vakkari (2006), in which the authors state that 'it was important to evaluate and combine information from several resources' (p. 19) and continue that the typical outcome evaluation in IR has concentrated 'on the immediate answering of questions based on material retrieved' (p. 19) and that future research should focus on the process of utilisation of the items retrieved. The present study sheds some light on this matter in that we offer description and categorisation dealing with what sources are used (see Subsection 6.4.1) and what document components are used (Subsection 6.4.2). Dependency was also found between use of a large number of combinations of query types (terms, classification codes, dates, document numbers,

etc.) and a high number of information components being used in the final product (see Subsection 7.2.3).

In Figure 10.4, below, we show a general schematic overview of the dependencies between the variables within the categories considered. The variables are condensed to categories and merged into one figure. We can see dependencies between almost all categories of variables. Note that the section numbers in the figure (e.g., 7.2.2) point to more detailed descriptions of these relationships, found in Chapter 7.

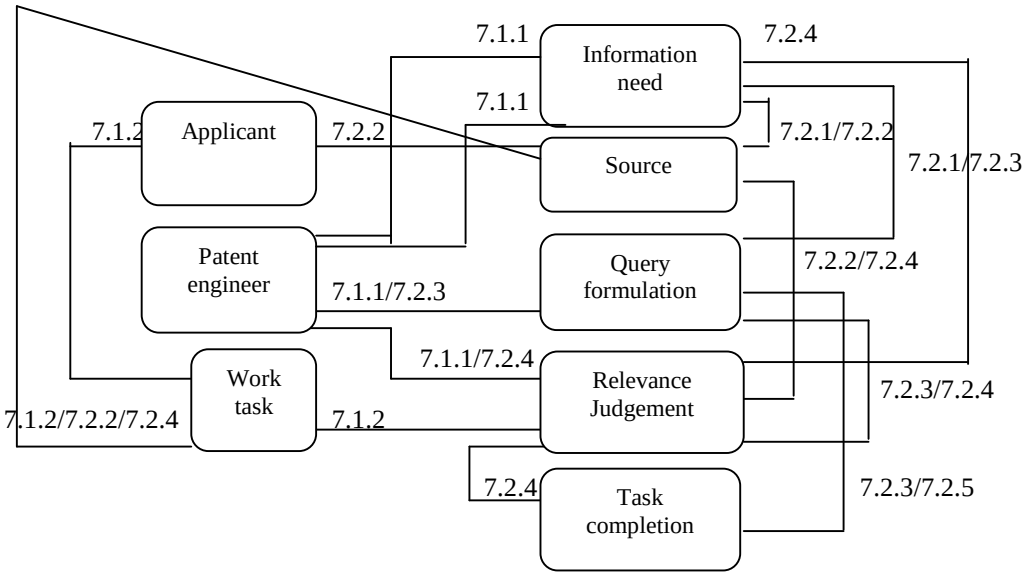


Figure 10.4: Schematic overview of dependencies between categories of variables

10.4 Collaborative information search

The seventh and last research question was concerned with *collaborative information retrieval activities* and how these activities were manifested in the course of the IS&R task performance process. Our empirical findings show that the patent task process involves collaborative activities throughout each stage in the IS&R process.

One prevailing assumption**,** especially within IR research (e.g., TREC[61]), is that an IS&R situation is an isolated activity. However, the empirical findings in our study show that patent information search is not merely an individual effort; it inherently involves various collaborative information search activities. This finding is important and contributes to reconfirming similar studies claiming that information retrieval is not performed in isolation (e.g., Karamuftuoglu, 1998; Romano et al., 1999; Fidel et al., 2000; Sonnenwald & Pierce, 2000; Hansen & Järvelin, 2000; Herzum & Pejtersen, 2000; Hansen & Järvelin, 2005; Foster, 2006; Reddy & Spence, 2008).

First, we identified that, within the patent domain, the task performance processes involve collaborative activities. Collaborative information seeking in the patent domain has not, to our knowledge, been reported upon before.

[61] http://trec.nist.gov/.

Secondly, we found that collaboration took place both asynchronously and synchronously. This was also the point of departure of Morris and Horvitz (2007) when they developed a system supporting asynchronous and synchronous functions.

Thirdly, observation showed that CIR activities were performed during *all* IS&R stages, including the initial stages of the IS&R processes in planning of the task as well as in formulation of queries. Most of the research on collaborative information retrieval has focused on specific parts of the seeking or searching process, such as query construction, or on collaborative search technologies that support, for example, collaborative querying and collaborative filtering as described by Foster (2006). Sonnenwald and Pierce (2000), who studied information behaviour in the context of command and control (C2), confirm this finding. The authors found that there was a continuing necessity of information exchange during work operations among the command-and-control staff. However, the authors did not explicitly talk about information searching, while our findings more specifically describe collaborative information retrieval actually occurring throughout the patent work task process.

Further to this, the following subtasks in the IS&R process especially involved collaborative activities:

Planning tasks. Even if there is a formal procedure of structuring the work task, a patent application may have aspects that lead to departure from that procedure. Furthermore, issues such as how to approach a specific process or procedure, or in what way to consult information sources never used before, require collaboration. This also involves *sharing* of *personal and subjective opinions*.

Problem definition. Collaboration also occurred in definition of the specific problem at hand. Novel areas of invention and both the complexity and the variation of sub-problems embodied in a patent application come with a need for support if one is to find the right focus and the core problem. In complex cases, the core problem may be hidden or divided into several parts. Sharing representations (such as classification codes or synonyms) of the information need may also be an important activity.

The characteristics of an information object. Three types of *history* were identified for an object: the document history, log history, and link history. These could be reused by colleagues. Another element found that could be shared and reused was the *contextual relationships between* information objects such as annotations, references, and citations.

Search paths and query construction sequences. Once the problem and the information need are defined, there may be a need for support for *query formulation and reformulation.* We found that for a given set of documents related to a topic, there was often a reason for sharing and reusing search paths. A search path could be reused up to a certain point, at which a more specific and unique search sequence started. The possibility of choosing the right query keys might be increased via reuse of earlier query formulations.

Task decision and relevance assessments. Sometimes a patent application was divided between two persons and this caused them to work together during the process; the relevance assessments had to be made jointly as well.

Final task outcome and work task completion. The *final outcome* of the patent process was often discussed with colleagues – for example, for checking of information previously handled by the patent office.

Taken individually, some of the findings described above confirm previous research – e.g., on sharing of personal opinions (Herzum, 2000) and collaboration in sharing of queries (e.g., Robertson & Hancock-Beaulieu, 1992; Talja, 2002), while others, such as those related to shared contextual relationships between documents, collaborative reuse of parts of a document's history, and collaboration during the closing stage of the work task, may further knowledge of collaborative IS&R.

We also assigned two additional collaborative activities to the four levels of information sharing in group situations described by O'Day (1993b): *case-building* and *history-building* activity for an information object. We found that the actual case, such as a patent work task, in part or as a whole, could be reused for similar upcoming work tasks, which was more efficient and effective for the patent examiner. In a similar way, the history of a specific document or set of documents, or its elements, could be reused, for more efficient task performance.

Finally, we identified two specific categories of collaborative activities: document-related and human-related CIR activities. The importance of human colleagues as information channels, for feedback and opinions, has been confirmed by Serola (2006). The first category has to do with collaborative activities manifested in document form (whether electronic or paper-based), and the second is composed of human-to-human collaborative activities. These might be co-located or distributed activities.

The pragmatic and holistic methodology used in our study indicates that the findings described above concerning CIR in patent work processes also contribute to the development of a CIR-enhanced conceptual framework for IS&R. The framework could be interpreted as describing two processes in parallel: The first and *primary process* describes the general IS&R process and the relationships among the various components. Additionally, on the right side in the depiction of the framework, we describe a *secondary process* of collaborative activities, involving both document-related and direct human-related CIR events. According to the framework, IS&R is dynamic and interactive and involves interrelated processes embedded in larger work tasks. The model explicates CIR activities at each stage of IS&R. Task performance may involve other processes as well, but here we will mainly discuss the collaborative activities in relation to the conceptual patent work task framework (the area with grey shading in Figure 10.3, below).

The framework points to elements related to *when*, *which kind of*, and *how* collaborative activities are manifested in work task performance. The *when* refers to when in the task performance process single subtasks (i.e., tasks of a particular kind) are performed. The *how* refers to document-related collaborative activities that are mediated through documents and textual information. These may be written notes that are used by others or search logs that are processed and (re)used for similar tasks. Document-based collaborative manifestations could be made explicitly (the activity is done with the goal of sharing with others) or implicitly (the activity may eventually lead to a collaborative activity but is not primarily intended to do so). The direct

human CIR activities are of a collegial character – e.g., discussing and adjusting to consensus within a department on how to proceed in specific situations.

In the figure below, we can see that all main stages have related CIR activities. At the level of the performer, collaboration may be done with internal and external domain expertise as human-to-human communication. The task initiation stage involves CIR activities related to delegation and assignment of work tasks (human- and document-based). The task preparation work showed collaborative events such as sharing and working together on a work task in which personal opinions are shared among patent engineers. During that process, case-building documents are shared to ensure proper preparation.

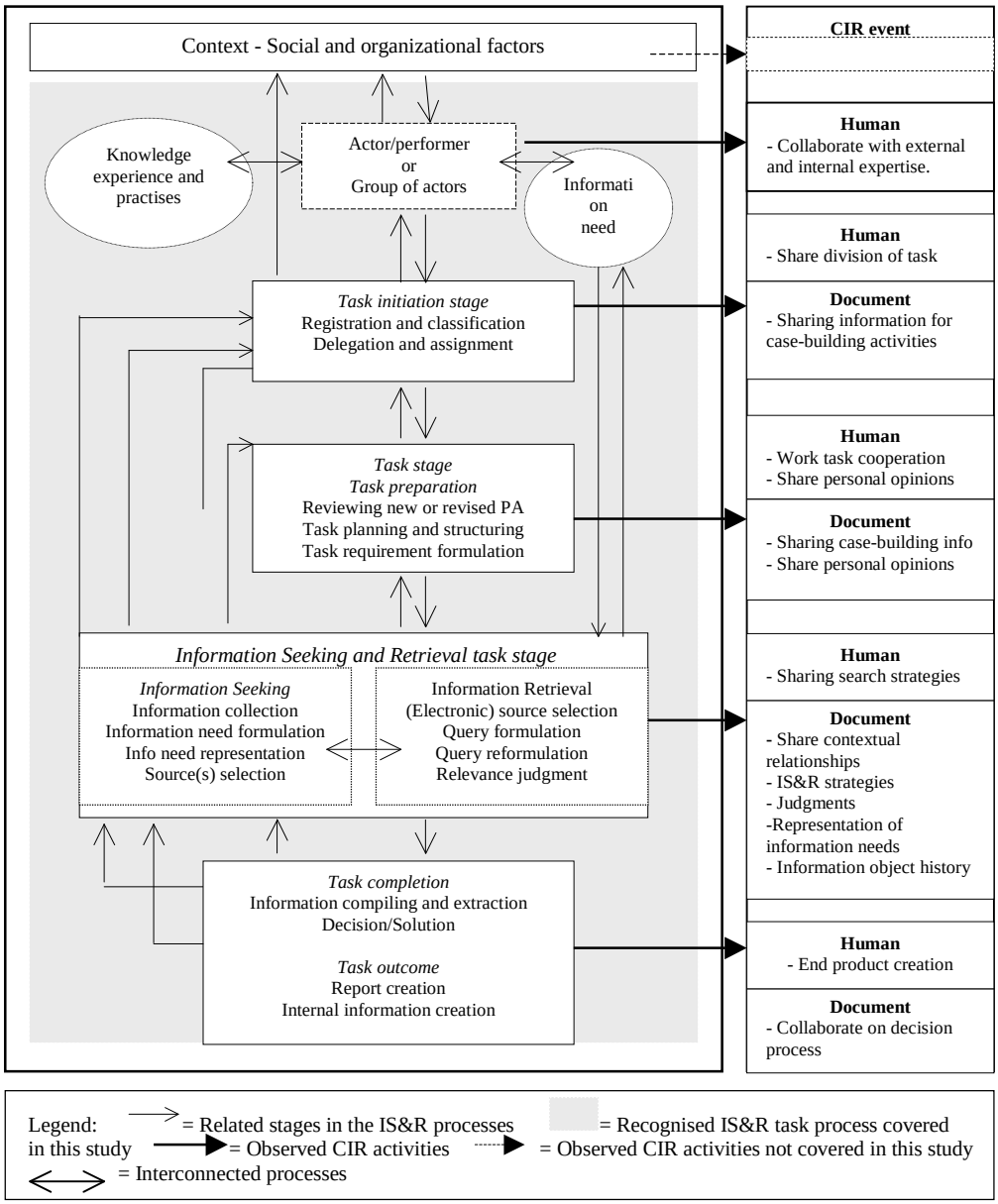


Figure 10.5: Framework for the patent handling process, including CIR

Most of the collaborative information retrieval activities were found in the IS&R task stage. Search strategies were a very important factor, and strategies and experiences were used collaboratively both verbally (human-to-human) and through documents written by colleagues and related to the current patent task. So, communication and debating concerning complex search strategies are necessary, especially if the search task involves several topics and/or information needs. This stage also featured other aspects of document-related collaboration, such as implicit collaboration through written representations of different types of information needs, contextual relationships between specific documents, and the history (such as links and logs) related to a specific document through annotations. In addition, assessments and judgements made by colleagues may be reused. Finally, the framework also shows that the final stage of the IS&R embedded work task involves collaboration. The end product may be finished through collaboration (human), and the final decision process may be based on earlier judgements and decisions for similar or parent written reports and applications.

Finally, this work set out to investigate a complex ISR situation and has proved not only that the task is complex (Byström & Järvelin, 1995) but also that the general assumption often made that the IR activity is performed by a single person may not be valid in a complex and information-intensive information context. In our case, all patent engineers were rather experienced. If our study had included participants with less experience, for example, there might have been more collaborative activities.

10.5 The methodological approach

In relation to our second research problem, a methodology for analysing the data of the task-based PIR studies was constructed and refined. In the study described here, we focused on two aspects: The first concern was that we utilise a combination of data and analysis collection methods. Secondly, we developed a method for analysing, describing, and systematically categorising patent IR sessions and modelling session-based information retrieval. In addition, schematic visualisations and schematic diagrams were developed to illustrate its application.

One big challenge in our study was how to collect data and what methods to use, given the setting of a patent work environment with real-life conditions and practices, involving professional patent engineers. In response to this, we collected both qualitative and quantitative data. Our methods allowed us to observe not only what real-life activities the PEs performed both online and physically but also how these activities were performed over time. This approach also made it possible to describe and characterise the patent domain examined, as well as the work processes and procedures, both at a general level and in more specific cases. Naturally, every method does have its limitations.

As our data collection methods we used and combined log data, diary data, and on-site observations of patent actors' performance. Our study complied with Brennan's (1992, pp. 59–61) statement that a) quantitative research can unfold more structural features of the research target, while qualitative research supports understanding of processes, and b) qualitative methods support the subject's perspective, while the quantitative is related more to the researcher's own focus. For example, the observations and diaries revealed collaborative IR activities, which were not visible from the log data. In line with what Hyldegård (2006) reports, using diaries may generate very useful data in

relation to individuals' and group members' behaviour, task performance, and activities and IS&R processes. We found that the diary should be designed in such a way that it allows and encourages the users to fill it in completely. Thus a delicate balance exists between assignment of pre-coded categories and the openness of user-formulated categories. This will affect the data quality. Before being used by the participants, a diary should be validated in its accuracy and usability. This was done in the pilot study. Unlike Hyldegård, we used an electronic diary, which proved easy to use, especially as one of the items asked the user to submit the search logs via the diary. It was easy for the participants just to copy and paste things into the diary, especially lengthy database logs. Furthermore, it proved essential that the researcher acquired some understanding before constructing the diaries, so that participants found the categories and content of the diaries relevant. However, some issues still require further discussion:

Type of study setting. We found it natural as well as challenging to perform this research in a real-life and natural setting (Goguen & Linde, 1993). The results of the study can both inform and complement studies using, for example, simulated work tasks (e.g., Borlund, 2000) based on small and more context-free laboratory experiments. The real-life work task setting made us confront the real conditions and requirements of work task performance, and it was certainly rewarding to gain this insight in an information-intensive domain. It will be possible to perform follow-up studies based on the findings reported here as well as construct enhanced and perhaps more focused research designs using a similar set-up or take smaller parts of the study set-up into a lab-based setting. Furthermore, very often, the participants in simulated and laboratory-based studies are recruited from academic settings, without domain and task skills matching professional participants. Our work was a departure from this.

Control over data. Given the real-life setting and other issues in observing professional patent engineers in their work tasks, we knew from the outset that we could not claim full control over the data collected and what actually was produced. Without control, one may not be able to foresee certain data output fully and design for it. On the other hand, such work may be rewarding in that unexpected datasets may emerge and new angles may be unfolded. So there are both problematic aspects and rewarding aspects. In our case, we were aware that our methods would create not entirely controllable data and that we would need to be prepared to adapt to the data as they were produced. For example, for different variables, we did not know which categories the data would result in. Another aspect of the data issue was the problem of normalisation of the data collected, in order to allow comparison of different types of data.

Time. Time was another uncertain aspect of the data collection process. Since we took a task-based approach, the basic unit for observation and data collection was the individual work task. Especially when observing the on-site work tasks, we did not know beforehand when a task should be deemed to end or what the process of the observed task would lead to. This made us constantly prepared to step in and step out of individual work task processes, depending on the patent engineers' daily routines and planned work. On the rewarding side, this gave us important insight into the workers' daily work. Finally, another important issue of the data collection process is that we observed only shorter units in the larger work task process. Usually, a patent work process takes between one and two years, and the sessions we observed gave us

only small windows from that perspective. However, we managed to observe enough IS&R units for us to perform systematic collection and analysis of the IS&R processes and to draw conclusions.

Combined methods. We set out to use a combination of quantitative and qualitative data collection and analysis methods, in order to harvest and capture as diverse and dynamic data as possible and still have reasonable control over the data collected. This proved successful, although combining methods for data collection and analysis was, to some extent, a cumbersome way to organise, merge, and tabulate all the data types in workable data sheets. We must note here that using a combination of methods has some implications, concerning both productive aspects and problematic aspects of combined methods. The utility of different but complementary data collection methods support an *exploratory* study approach in a new domain. Along with the exploratory approach, using combined methods may contribute to a dynamic way of observing, collecting, analysing, and understanding what one investigates and is therefore a *productive aspect* of our methodology. Furthermore, the use of several methods may contribute to a more *explanatory* perspective in that collected and enriched datasets in numeric form may be explained by qualitative data and qualitative data may then inform models and frameworks.

The different methods used may result in a very diverse set of data – for example, continuous and non-continuous data or binary data and data with multiple degrees – and may therefore create *problematic aspects* to our methodology. For example, to rectify the combination of different types of data, a normalisation process was utilised. A set of multiple methods also makes for time-intensive data collection. Some methods are demanding in terms of scheduling with the users and handling of incoming real-life work tasks whose analysis may be appropriate for study. In a few cases, we needed to wait some time before the right observable task came along. In other situations, we needed to abandon the work task just begun, because the task was cancelled, for example. Another problematic aspect was synchronising two quite different data collection methods during the same real-life work task.

In view of our second research problem, a methodology for analysing the data of task-based patent information retrieval studies was developed. In the present study, we have focused on two elements. As noted above, the first concern was to utilise a combination of data and analysis collection methods, and secondly, we developed a method for analysis, description, and systematic categorisation of patent IR sessions and for modelling of session-based information retrieval. In addition, schematic visualisations and schematic diagrams were devised that illustrate its application.

We developed schematic diagrams that visualised the IR processes and demonstrated that various types of data can be systematically collected and categorised, then mapped, through this protocol. The schematic diagram covers the main aspects of an IR process and depicts the process from two perspectives along a time scale. The vertical axis presents the relationships between one main ‘category’ (e.g., source (S)) and multiple search sessions, queries, terms used, and documents judged and saved that are related to the source in question. In the horizontal dimension, several parallel channels depict separate processes. For example, in one horizontal channel, all sources used are depicted. Similar channels exist for mapping other IR processes, such as search sessions, queries, terms used, and relevance judgements.

Additionally, each category may have subclasses and each of them have unique (numeric) values mapped to it. These values may be measured and compared with those for other IR tasks, performed within or outside the patent domain. Each action may also have a timestamp, by means of which each action (with its value) can be located, retrieved, and re-examined. Each category and subclass of a main category has a unique label (e.g., 'S1', in which S = source and 1 corresponds to the source INSPEC). Each time that source is used, it is labelled as S1. What makes the scheme usable and productive is that a) all actions can receive a timestamp, b) the order of all activities with regard to the overall process is retained, and c) each category and subclass is assigned a value. Using these categories and values enables comparison with other, similar IR processes. This is true for those within the patent domain, but it is also our belief that the scheme can be used in other domains too. Finally, the labelling of the different aspects of IR (as well as their values) was also integrated into a formal (set-theory) description of the patent IR processes.

In summary, these results suggest that the methodological approach utilised can provide insight-generating, rich, and valuable data.

10.6 Limitations

The methodology used in this study was feasible, although tedious, and resulted in unfolding of interesting and new aspects of IS&R processes, of which some confirm previous results from other researchers. It is worth noting, however, that the results of the study are based on the circumstances that existed at the specific time when the investigation was performed. Some of the processes observed may have changed since the study was performed, which could render it difficult to reproduce in full some of the specific situations reported on here. Changes in internal organisational procedures and enhancements to technology may affect ability to reproduce the study. Because of the methodology used, care must be taken in generalising the results from the present study. On the other side, using the limited number of subjects resulted in a surprisingly substantial quantity of data for analysis and from which it was possible to describe and explain patent IS&R processes.

It was our intention to try to bring some understanding of what is actually taking place in IS&R in a real-life setting. Therefore, the empirical investigation described in this thesis does not claim to examine complete work tasks solely; rather, it examines some aspects of the work tasks in relation to IS&R performance. For example, in our observations, we were not able to cover entire patent handling tasks from beginning to end (this could take up to 18–24 months). Instead, we chose to focus on important and manageable episodes of IS&R activities. Even though the IS&R episodes observed did not represent the complete work task involving the IS&R processes, the monitored episodes were definitely enough and contributed to a rich and complex body of data to be analysed. We also believe that gathering data from 54 separate work tasks can be considered enough for reaching the level that is critical for our qualitative study. Finally, this study focused on the IS&R embedded in professional work tasks, and, therefore, we have not investigated the attributes etc. of the performers of the tasks in depth.

The study described here is a user-oriented study. It was performed in a real-world setting and entailed access to observable situations. This is in line with the distinction between a system-based study and a user-oriented approach as described by Robertson and Hancock-Beaulieu (1992):

> The conflict between laboratory and operational experiments is essentially a conflict between, on the one hand, control over experimental variables, observability, and repeatability, and on the other hand, realism. (Robertson & Hancock-Beaulieu, 1992, p. 460)

Limited control over the variables was gained through the pilot study, in which an initial set of variables was designed for incorporation into the electronic diaries and observation protocols. Complete control was impossible, since the study took place in a real-world professional situation. Therefore, repeatability in the sense of conducting a completely identical study in the same real-life setting is not possible. However, the protocol and formalisation developed in the study may solve this problem. The goal may be not to reproduce the same exact study but to conduct a similar study, which may explore the relationships of the set of variables established in the present study.

10.7 Conclusions

In this study, we have analysed the relationships of task properties for information access in the patent domain. The goals for the thesis were reached through a longitudinal real-world study of the online information searching and use of information within the patent domain with professionals performing their work tasks.

a) We showed that there is a set of significant relationships among different aspects of the task-based IS&R process.
b) New insights were revealed, showing that the general IS&R process also involves collaborative information handling processes.
c) The study established a methodology for systematically studying empirical IS&R activities and processes, in the context of a study carried out over a longer span of time and in a real-life professional work setting.
d) Session-based patent search processes were exemplified through schematic visualisations and diagrams.
e) We confirmed the conclusion, previously found but in separate works, that multiple search sessions, information needs, sources, queries, and relevance judgements have been utilised in patent IR. These findings deviate from the presumptions of traditional IR research models.

These insights and results contribute to an enhanced understanding of the specific properties of individual aspects of, and relationships within, IS&R activities in the patent domain as well as to a modified and enhanced IS&R model.

The research questions and point of departure of the present study, and its findings, continue a chain of earlier research-based studies. This study confirms, in part or in full, previous findings and theories (e.g., Bates, 1989, 1990; Kuhlthau, 1993; Leckie et al., 1996; Byström & Järvelin, 1995; Ellis & Haugen, 1997; Vakkari, 1999, 2000b, 2001, 2003; Wilson, 1999; Hansen & Järvelin, 2000, 2005) and makes its own contribution to general task-based IS&R studies. This study is connected to an

established and growing base of knowledge and contributes to its development and enhancement. Besides the core of the IS&R activities, we acknowledge and consider the importance of the beginning and the completion of the IS&R performance, as pointed out by, for example, Kuhlthau (1993), Vakkari (2003), and Savolainen (2009). The model in this research visualises all stages (in the patent domain) as connected (see Figure 10.5).

Patent IS&R features and relationships: Given the insight acquired via the detailed exploration and examination of the real-life patent domain and patent engineers performing professional work tasks, the present study can contribute to an extended research agenda by providing classification of variables and features that are involved in the patent IS&R process. It has also uncovered relationships (see Figure 10.4) between some of the variables observed and has newly identified certain aspects of the IS&R process, such as collaborative searching as visualised in the extended IS&R framework.

We have discussed real-life task-based information access in the patent domain at the work task level and the IS&R level. We have proposed that, in the course of the IS&R process, *multiple information needs* may occur and that they may change (Bates, 1989, 1990), as well as that part of the information need may be muddled while at the same time another part is clear. We showed that in an information-intensive domain, such as the patent domain, *multiple sources or channels* – such as patent databases and Web search engines – are used in the completion of a work task. Furthermore, *multiple queries and multiple sessions* are employed during a work task process. Sometimes the same queries are used in different channels, and sometimes a session may be repeated in different channels. Sessions might feature both complex queries and short, simple ones. We also showed that *relevance judgements* are performed in a complex and dynamic way, and we identified two types of strategies here: sequential and aggregated relevance judgements. The evolving nature of the real-world relevance judgements was observed effectively through longitudinal-style process-oriented approaches to IS&R. For example, as illustrated in Chapter 9, we showed how different relevance judgement strategies were applied over time within an IS&R task and how the search process evolved as depicted in the visualisations of search processes.

Dependencies observed between different features within the work task process show that IS&R is a complex, interactive, and integrated process in which all stages are connected in the completion of a work task. These features of patent IS&R clearly prove that previous assumptions that the search process is a simple and linear search activity need to be reconsidered, both practically, in the design of experiments, and for system design in which systems, interfaces, and means of interacting with them need attention – theoretical and otherwise.

Collaborative information retrieval: In the theoretical field, our findings extend existing general IS&R models (e.g., Bates, 1989, 1990; Kuhlthau, 1993; Leckie et al., 1996; Byström & Järvelin, 1995; Ellis & Haugen, 1997; Wilson, 1999) to involve collaborative information handling events. Collaborative IS&R can be, partially or fully, integrated into the overall work task process, as reported by, for example, Hyldegård (2006b). We showed that each stage of the general IS&R process may be related to a corresponding collaborative information handling element. This has

practical and empirical implications. The design of experiments and field observations of IS&R processes needs to incorporate features that acknowledge features of collaborative information retrieval. Some of these features have been presented in this study.

Methodology of real-life application: We performed the study in a real-world context since we wanted to examine the work-related IS&R activities both as specific units (tasks) and as a process. The dynamic nature of real-life IS&R processes embedded in professional work tasks has been shown, and the study has demonstrated that dependencies exist between different levels of the IS&R process. We combined qualitative with quantitative data collection and analysis methods, and, in order to analyse the search process and its features, we developed and utilised a sequential model for mapping of search activities and values for these activities. With this mapping, each search task can be analysed and compared with other search tasks. Thus pattern analysis as well as detailed analysis of single tasks can be carried out.

Session-based patent IR: Through analysis, description, and systematic categorisation of patent information retrieval activities, we developed a way to model session-based retrieval sessions. Furthermore, we provided schematic visualisations and diagrams for these sessions.

Finally, the results of this study suggest that the constraints and opportunities of the domain play an important role in ability to determine and understand the IS&R processes; therefore, both the study and the design of the information flow must be tailored to the specific domain in question. The study described in this thesis shows us that patent IS&R operations are performed by individuals or as a collaborative effort. It also shows us that IS&R is a fairly dynamic and complex information handling activity. In examination of this complexity, significant relationships between the work task and IS&R tasks showed how work tasks and IS&R tasks are related.

10.8 Future work

Understanding and obtaining fine details of a real work task situation may serve the purpose of designing further interactive IS&R experiments; this knowledge could thus benefit a more realistic real-world contextual experimental design, based on user requirements elicitation and evaluation. We think that continuing investigations such as those presented in this study will contribute to wider and fuller understanding that introduction of realistic derived components will give us better tools but also more complex and dynamic components for a study. Also, subsequently, better experiments may result, perhaps leading, in turn, to improved design of interactive information systems.

Our study's finding of collaborative information handling activities points in the direction of it being very important to consider collaborative aspects in the design of studies of user behaviour and IS&R processes as well as in the design of information systems (such as IR systems). Future research could deal with task variation, task complexity, or type of task. Furthermore, we need to develop frameworks in which to examine information (seeking and retrieval) and theories that address methodologies

for evaluation of collaborative IR systems. Consequently, this knowledge will also have an impact on the evaluation of IR systems.

Future studies could investigate the behaviour of users more closely. On account of the goals of this study, personal differences and group characteristics were not investigated; these could yield additional knowledge and understanding of certain IS&R phenomena. However, for this, both more actors (patent engineers) and more patent work tasks related to these actors are needed. Finally, additional research is needed for generating understanding of the circumstances in which the collaborative information handling activities take place. In future research, the variables used in our study may be fine-tuned and connected with more workable metrics.

REFERENCES

Ackerman, M. & Malone, T. (1990). Answer Garden: a tool for growing organizational memory, In: *Proceedings of the ACM SIGOIS and IEEE CS TC-OA conference on Office information systems*, April 25-27, 1990, Cambridge, Massachusetts, USA, 31-39.

Akers, N. J. (1999). The European Patent System: an introduction for patent searchers. *World Patent Information* 21 (1999), 135-163.

Allen, T. J. (1977). *Managing the flow of technology*. Cambridge, Mass.: MIT Press.

Allen, B. (1997). Information needs: a person-in-situation approach. *Proceedings of the International Conference on Information Seeking in Context* (ISIC), Tampere, Finland, 111-122.

Anderson, B. & Alty, J. L. (1995). Everyday Theories, Cognitive Anthropology and User-Centred System Design. In Kirby, M., Dix, A. and Finlay, J. (Eds) *HCI '95 Proceedings of the HCI'95 conference on People and computers X*. August 1995. New York: Cambridge University Press, 121-135.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Wokingham, UK: Addison-Wesley.

Bates, M. J. (1979a). Information Search Tactics. *Journal of the American Society for Information Science* 30, 205-214.

Bates, M. J. (1979b). Idea Tactics. Jo*urnal of the American Society for Information Science 30*, 280-289.

Bates, M. (1989). The design of browsing and berrypicking techniques for online search interface. *Online Review* 13 (5), 407-424.

Bates, M. (1990). Where should the person stop and the information search interface start? *Information Processing & Management* 26 (5), 575-591.

Bashir, S. & Rauber, A. (2009). Identification of Low/High Retrievable Patents using Content-Based Features. In. *Proceedings of the 2nd international workshop on Patent information retrieval* (PaIR'09), Hong Kong, China, 9-16.

Belkin, N. J., Cool, C., Stein, A. & Thiel, U. (1995). Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems With Applications* 9 (3), 379-396.

Belkin, N. J., Marchetti, P. G. & Cool, C. (1993). BRAQUE: Design of an interface to support user interaction in Information Retrieval. *Information Processing & Management* 29 (3), 325-344.

Belkin, N. J., Oddy, R. N. & Brooks, H. M. (1982a). ASK for information retrieval: Pat 1. Background and Theory. *Journal of Documentation*, 38(2), 61-71.

Belkin, N. J., Oddy, R. N. & Brooks, H. M. (1982a). ASK for information retrieval: Pat 2. Results of a design study. *Journal of Documentation*, 38(3), 145-164.

Bennett, J. L. (1972). The user interface in interactive systems. *ARIST*, 7, 159-196.

Bonino, D., Corno, F. & Ciaramella, A. (2010). Review of the State-of-the-art in Patent Information and Forthcoming Evolutions in Intelligent Patent Informatics. *World Patent Information* 32 (1), 30-38.

Borgman, C. (1989). All users of information retrieval systems are not created equal: an exploration into individual differences. *Information Processing & Management* 25 (3), 237-252.

Borlund, P. (2000). *Evaluation of interactive information retrieval systems*. Åbo, Finland: Åbo Akademi University Press, Ph.D. Thesis.

Borlund, P. & Ingwersen, P. (1997). The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. *Journal of Documentation* 53 (3), 225-250.

Brannen, J. (Ed.). (1992) *Mixing Methods: Qualitative and Quantitative Research*. Burlington: Ashgate Publishing.

Byström, K. (1999). *Task complexity, information types and information sources*. *Acta Universitatis Tamperensis* 688. Tampere: Tampere University Press. [Doctoral dissertation].

Byström, K. (2000). The effects of task complexity on the relationship between information types acquired and information sources used. *New Review of Information Behaviour Research* 1**,** 85-101.

Byström, K. (2002). Information and information sources in tasks of varying complexity. *Journal of the American Society for Information Science and Technology* 53(7), 581-591.

Byström, K. & Hansen, P. (2002). Work tasks as unit for analysis in information seeking and retrieval studies. *The Fourth International Conference on Conceptions of Library and Information Science: Emerging Frameworks and Methods. CoLIS4*, Seattle, WA, USA, July 21-25, 2002, 239-252.

Byström, K. & Hansen, P. (2005). Conceptual Framework for Task in Information Studies. JASIST - *Journal of the American Society for Information Science and Technology* 56 (10), 1050-1061.

Byström, K. & K. Järvelin (1995). Task complexity affects information seeking and use. *Information Processing & Managemen*t 31 (2), 191-213.

Cleverdon, C. W., Mills, J. & Keen, E. M. (1966). *Factors determining the performance of indexing systems.* Cranfield, UK: Aslib Cranfield Research Project, College of Aeronautics. (Volume 1: Design; Volume 2: Results).

Cosijn, E. (2006). Relevance judgements within the context of work tasks. In: Ruthven, I. & al. (Eds.) *Information Seeking in Context: Proceedings of an International Conference on Research in Information needs, Seeking and Use in Different Contexts* Copenhagen: ACM Press, 20-29.

Cosijn, E. & Ingwersen, P (2000). Dimensions of relevance. *International Journal of Information Processing and Management* 36(4), 533 – 550.

Creswell, John W. (1998). *Qualitative inquiry and Research Design. Choosing Among Five Traditions*. London: Sage.

Croft, W.B., Metzler, D. & Strohman, T. (2009). Sea*rch Engines: Information Retrieval in Practice*. Boston: Addison Wesley.

De Mey, M. (1977). The cognitive viewpoint: Its development and its scope. *Communication & Cognition* 10(2), 7-23.

Denzin, N. K. & Lincoln, Y. S. (1994). Introduction: Entering the field of qualitative research. In: N. K. Denzin and Y. S. Lincoln (Eds.) *Handbook of qualitative research*. Thousand Oaks, CA: Sage, 1-17.

Dervin, B. (1997). Given a context by any other name: Methodological tools for taming the unruly beast. In: Vakkari, P., Savolainen, R. & Dervin, B. (Eds.) *Information Seeking in Context: Proceedings of an International Conference on Research in Information needs, Seeking and Use in Different Contexts*. London: Taylor Graham, 13-38.

Dervin, B. & Nilan, M. (1986). Information needs and use. In: Williams, M. E. (Ed.) *Annual Review of Information science and Technology (ARIST)*, Volume 21, 3-33.

Dey, I. (1999). *Grounding Grounded Theory. Guidelines for Qualitative Inquiry*. Academic Press: San Diego.

Efthimiadis, E. (2000). Interactive query expansion: A user-based evaluation in a relevance feedback environment. *Journal of American Society of Information Science and Technology* 51 (11), 989-1003

Ellis, D. (1989). A Behavioural approach to information retrieval design. *Journal of Documentation*, 46(3), 318-338.

Ellis, D. & Haugen, M. (1997). Modelling the Information Seeking Patterns of engineers and Research scientists in an Industrial environment. *Journal Of Documentation*, 53(4), 356-369.

Elsweiler, D. & Ruthven. I. (2007). Towards Task-based Personal Information Management evaluations. In: Charles L.A. Clarke, C., Fuhr, N., Kando, N., Kraaij, W and de Vries, A (Eds.) *Proceedings of the 30th Annual ACM Conference on Research and Development in Information Retrieval*. 2007, Amsterdam, The Netherlands, July 23 - 27, 2007. ACM Press: New York, NY, USA, 23-30

EPO - Guidelines for Examination in the European Patent Office (status April 2009). Part B. Available at http://documents.epo.org/projects/babylon/eponet.nsf/0/1AFC30805E91D074C125758A0051718A/$File/guidelines_2009_part_B_en.pdf (accessed 2011-04-04)

Ehrlich, K. & Cash, D. (1994). Turning information into knowledge: Information finding as a collaborative activity. In Pro*ceedings of the first annual conference on the theory and practice of digital libraries*. College Station, TX, USA, 119-125.

Feinman, S., Mick, C., Saalberg, J. & Thompson, C. (1976). A conceptual framework for information flow studies. In: Martin (ed.) *Information Politics: Proceedings of the 38th Annual Meeting of the American Society for Information Science* 13(1), 106-116.

Fidel, R. (1984). Online Searching Styles: Case-Study-Based Model of Searching Behavior. *Journal of American Society of Information Science*, 35 (4), 11-221.

Fidel, R. (1985). Moves in online searching. *Online Review*, *9(1)*, 61-74.

Fidel, R., Bruce, H., Pejtersen, A. M., Dumais, S. Grudin, J. & Poltrock, S. (2000). Collaborative Information Retrieval (CIR). *The New Review of Information Behaviour Research* 235-247.

Freund, L. *(*2008). *Exploiting task-document relations in support of information retrieval in the workplace*, doctoral dissertation, Faculty of Information Studies, University of Toronto.

Foster, J. (2006). Collaborative information seeking and retrieval. In: *Annual Review of Information Science and Technology*, 40, 329–356.

Fujii, A., Utiyama, M., Yamamoto, M. & Utsuro, T. (2008). Overview of the Patent Translation Task at the NTCIR-7 Workshop. In: *Proceedings of the NTCIR-7 Workshop Meeting*, 2008, Tokyo, Japan, 389-400.

Gerstberger, P. & Allen, T. (1968). Criteria used by research and development engineers in selection of an information source. *Journal of Applied Psychology 52*, 272-279.

Giddens, A. (1979). *Central problems in social theory: action, structure and contradiction in social analysis*. Berkeley, CA: University of California Press. 50.

Gougen, J. & Linde, C. (1993). Techniques for Requirement Elicitation. In: Fickas, S. and Finkelstein, A. (Eds.), *Proceedings of IEEE International Symposium on Requirements Engineering*, 1993, 152 – 164.

Gravetter, F. & Wallnau, L (2000). *Statistics for the behavioural Science*. (5th edition). Wadsnorth: Belmont.

Hackman, J. R. (1969). Towards understanding the role of tasks in behavioral research. *Acta Psychologica, 31*, 97-128.

Haake, J. M., Wiil, U. & Nürnberg, P. (1999). Openness in shared hypermedia workspaces: the case for collaborative open hypermedia systems, *ACM SIGWEB Newsletter*, 8(3), 33-45.

Hansen, P. (1999). User Interface design for IR Interaction. A Task-oriented approach. Aparac-Gazivoda, T. & Saracevic, T. (Eds.). *Digital Libraries: proceedings of the Third International Conference on the Conceptions of the Library and Information Science, Dubrovnik, Croatia, May 23-26, 1999*. Zagreb: Zavod za informacijske studije Odsjeka za informacijske znanosti, Filozofski fakultet, Dubrovnik, Croatia, 1999.

Hansen, P. & Järvelin, K. (2005). Collaborative information retrieval in an information-intensive domain. *Information Processing and Management,* 41 (5), 1101-1119.

Hansen, P. & Järvelin, K. (2004). Collaborative Information Searching in an Information-Intensive Work Domain: Preliminary Results. *Journal of Digital Information Management*, 2(1), 26-30.

Hansen, P. & Järvelin, K. (2000). The Information Seeking and Retrieval process at the Swedish Patent- and Registration Office. Moving from Lab-based to real life work-task environment. *Proceedings of the ACM-SIGIR 2000 Workshop on Patent Retrieval*, Athens, Greece, July 28, 43-53.

Hansen, P. & Karlgren, J. (2005). Effects of foreign language and task scenario on relevance assessment. *Journal of Documentation*, 61(5), 623-638.

Harper, R. H. R. & Sellen, A.J. (1995). Collaborative tools and the practicalities of professional work at the International Monetary Fund. In: Mack, R., Mark, L., Collins, D. & Instone, K. (Eds). *Proceedings of SIGCHI Conference on Human factors in computing systems*, Denver, CO., USA, 122-129.

Hearst, M.A. (2009). *Search User Interfaces*, Cambridge University Press.

Hertzum, M. (2000). People as carriers of experience and sources of commitment: information seeking in a software design project. T*he New Review of Information Behaviour Research*, Vol. 1, 135-149.

Herzum, M. & Pejtersen, A. M. (2000). The information-seeking practices of engineers: searching for documents as well as for people. *Information Processing & Management* 36(5), 761-778.

Hill, B., Long, J., Smith, W. & Whitefield, A. (1993). Planning for Multiple Task Work. An Analysis of a Medical Reception Worksystem. In: Ahslund, S., et al. (Eds.) *INTERCHI'93 Conference on Human Factors in Computing Systems*. Bridges Between Worlds. Amsterdam, The Netherlands, 24-29 April 1993, 314-320.

Hsieh-Yee, I. (1993). Effects of Search Experience and Subject Knowledge on Online Search Behavior: Measuring the Search Tactics of Novice and Experienced Searchers. *Journal of the American Society for Information Science* 44, 161-174.

Huuskonen, S. & Vakkari, P. (2006). Situational relevance and task outcome. In: Ruthven, I., et al. (Eds.) *Information Interaction in Context*, IiiX. Copenhagen: ACM Press, 24-32.

Hyldegård, J. (2006). Using diaries in group based information behvior research- a methodological study. In: Ruthven, I., Et & al. (Eds.) *Information Interaction in Context*, IiiX. Copenhagen: ACM Press, 153-160.

Hyldegård, J. (2006b). Collaborative information behaviour - exploring Kuhlthau's Information Search Process model in a group-based educational setting. *Information Processing & Management* 42(1), 276-298.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation* 52 (1), 3-50.

Ingwersen, P. (1992). *Information Retrieval Interaction*. London, UK: Taylor Graham.

Ingwersen, P. & Järvelin, K. (2005). *The Turn. Integration of Information seeking and retrieval in Context*. Dordrecht: Springer.

Jansen, B., Spink, A. & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & management* 36, (2000), 207-227.

Jensen, E. (2009). Sensemaking in military planning: a methodological study of command teams. *Cognition, Technology & Work*. 11 (2), 103-118,

Jochim, C., Lioma, C., Schütze, H., Koch, S. & Ertl, T. (2010). Preliminary study into Query Translation for Patent Retrieval. In. *Proceedings of the 3rd international workshop on Patent information retrieval* (PaIR'10), Toronto, Canada, 57-66.

Johnson, D.J. (2003). On context of information seeking. *Information Processing & Management*, 39(5), 735-760.

Joho, H., Azzopardi, L. & Vanderbeuwhede, W. (2010). A Survey of Patent Users: An Analysis of Tasks, Behaviour, Search Functionality and System Requirements. In Belkin N. and Kelly, D. (Eds.): *Proceedings of the Third Information Interaction in Context Symposium (IiiX 2010)*. New Brunswick, NJ, US, 13-24.

Järvelin, K. (1986). On information, information technology and the development of society: an information science perspective. In: Ingwersen, P., Kajberg, L. & Pejtersen, A.M. (Eds.), *Information technology and information use: Toward a unified view of information and information technology*. London: Graham Taylor, 35-55.

Järvelin, K. (2007). An Analysis of Two Approaches in Information Retrieval: From Frameworks to Study Designs. *Journal of the American Society for Information Science and Technology* 58(7), 971-986.

Järvelin, K. & Ingwersen, P. (2004). Information seeking research needs extension towards tasks and technology. *Information Research*, 10(1) paper 212 [Available at http://InformationR.net/ir/10-1/paper212.html. Last visited 2011-03-17]

Järvelin, K. & Repo, A. (1983). On the impacts of modern information technology on information needs and seeking: A framework. In: Dietschmann, H. J. (Ed.), *Representation and exchange of knowledge as a basis of information processes*. Amsterdam, NL: North-Holland, 207-230.

Järvelin, K. & Wilson, T.D. (2003). On conceptual models for information seeking and retrieval research. *Information Research*, 9(1) paper 163. Available at http://informationr.net/ir/9-1/paper163.html. Accessed 14 May 2009.

Karamuftuoglu, M. (1998). Collaborative Information Retrieval: Towards a Social Informatics View of IR Interaction. *Journal of the American Society for Information Science* 49(12), 1070-1080.

Kari, J. (2007). Conceptualizing the personal outcomes of information. *Information Research*, Vol. 12 No 2, January 2007. Available at http://informationr.net/ir/12-2/paper292.html. Accessed 2009-05-29.

Kari, J. & Savolainen, R. (2007). Relationships between information seeking and context: A qualitative study of Internet searching and the goals of personal development. *Library & Information Science Research* 29, 47-69.

Kekäläinen, J. (2005). Binary and graded relevance in IR evaluations - Comparison of the effects on ranking of IR systems. *Information Processing and Management* 41, 1019–1033.

Kekäläinen, J. & Järvelin, K. (2002a). Evaluating Information Retrieval Systems under the Challenges of Interaction and Multi-Dimensional Dynamic Relevance. In: Bruce, H., Fidel, R., Ingwersen, P. & Vakkari, P. (Eds.). *Emerging frameworks and*

methods. CoLIS4. Proceedings of the Fourth International Conference on Conceptions of Library and Information Science. Seattle, WA, USA, July 21-25, 2002. Greenwood Village, Colorado: Libraries Unlimited, 253-270.

Kekäläinen, J. & Järvelin, K. (2002b). Using Graded Relevance assessments in IR Evaluation. *Journal of the American Society for Information Science* 53(13), 1120 – 1129.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2), 1-224. DOI: 10.1561/1500000012.

Korfhage, R. (1997). *Information Storage and Retrieval*. New York: Wiley.

Kuhlthau, C. (1997). The Influence of Uncertainty on the Information Seeking Behavior of a Securities Analyst. *Proceedings of an International Conference on Information seeking in context (ISIC)*, Tampere, Finland: August 14-16 1996. London, UK: Taylor Graham Publishing, 268-274.

Kuhlthau, C. (1993a). *Seeking meaning. A process approach to library and information services*. New York: Ablex publications.

Kuhlthau, C. (1993b). A principle of uncertainty for information seeking. *Journal of Documentation* 49 (4), 339-355.

Kuhlthau, C. (1991). Inside the search process: information seeking from the user's perspective. *Journal of the American Society for Information Science* 42(5), 361-371.

Kuhlthau, C. & Tama, S. L. (2001). Information search process of lawyers: a call for "just for me" information services. *Journal of Documentation* 57(1), 25-43.

Leckie, G., Pettigrew, K. & Sylvain, C. (1996). Modelling the information seeking of professionals: A general model derived from research on engineers, health care professionals and lawyers. *Library quarterly* 66(2): 39-52.

Leong, M-K. & Kando, N. (2000). (Eds.) *ACM-SIGIR Workshop on Patent Retrieval*, Athens, Greece, July 2000.

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge Series on Human Computer Interaction 9. Cambridge: Cambridge University Press.

Mick, C., Lindsey, G. & Callahan, D. (1980). Toward usable user studies. *Journal of the American Society for Information Science 31(5)*, 347-356.

Mizzaro, S. (1998). How many relevance in information retrieval? *Interacting with Computers* 10(3), 305-322.

Morris, M.R. & Horvitz, E. (2007). SearchTogether: An Interface for Collaborative Web Search. In: Shen, C., Jacob, R. and Balakrishnan, R. (Eds.), *Proceedings of the*

20th annual ACM symposium on User interface software and technology, Newport, Rhode Island, USA, October 7 - 10, 2007, 3 – 12

Newton, D. (2000). A survey of users of the British library patent information centre. *World Patent Information*, 22(4), 317-323..

Norman, D. (1988). *Psychology of everyday things*. BasicBooks.

O'Day, V. & Jeffries, R. (1993a). Orienteering in an Information Landscape. How Information Seekers Get From Here to There. In: Ahslund, S., et al. (Eds.) *Conference on Human Factors in Computing Systems (INTERCHI'93).* Bridges Between Worlds. Amsterdam, The Netherlands, 24-29 April 1993, 438-445.

O'Day, V. & Jeffries, R. (1993b). *Information Artisans: Patterns of Result Sharing by Information Searchers*. Report HPL-93-19, Systems Technology Laboratory February 1993, 13 pages. Hawlett-Packard. Available at http://www.hpl.hp.com/techreports/93/HPL-93-19.pdf. Accessed 2010-02-10.

Parapatics, P. & Dittenbach, M. (2009). Patent claim Decomposition for Improved Information Extraction. In. *Proceedings of the Second International Workshop on Patent information retrieval* (PaIR'09), Hong Kong, China, 33-36.

Pharo, N. (2002). *The SST Method Schema: a Tool for Analysing Work Task-Based Web Information Search Processes*. Acta Universitatis Tamperensis; 871. Tampere: Tampere University Press. [Doctoral dissertation].

Pinelli, T. E., Bishop, A. P., Barclay, R. O. & Kennedy, J. M. (1993): 'The information-seeking behaviour of engineers', in A. Kent and C. M. Hall (Eds.): *Encyclopedia of Library and Information Science*, vol. 52, supplement 15, Marcel Dekker, New York, 167-201.

Piroi, F & Tait, J. (2010) CLEF-IP 2010: *Retrieval Experiments in the Intellectual Property Domain*. *IRF Report 2010-00005*. Information Retrieval Facility, Vienna, 2010.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. & Carey, T. (1994). *Human-Computer Interaction*. Wokingham, England: Addison Wesley.

Rasmussen, J., Pejtersen, A. M. & Goodstein, L. P. (1994). *Cognitive systems engineering*. New York: Wiley.

Reddy, M. C. & Jansen, B. J. (2008). A model for understanding collaborative information behavior in context: A study of two health care teams. *Information Processing and Management* 44(1), 256–273.

Reid, J. (1999). A new, task-oriented paradigm for information retrieval: implications for evaluation of information retrieval systems. In: Aparac, T., Saracevic, T., Ingwersen, P. & Vakkari, P. (Eds.). *Digital Libraries: Interdisciplinary concepts, challenges and opportunities* (CoLIS3). Zagreb: Zavod za informacijske studije

Odsjeka za informacijske znanosti: Filozofski fakultet; Lovke: Naklada Benja, 97-108.

Robertson, S. E. & Hancock-Beaulieu, M. M. (1992). On the evaluation of IR systems. *Information Processing and Management* (28) 4, 457-466.

Romano, N. C. Roussinov, D., Nunamaker, J. F. & Chen, H. (1999). Collaborative Information Retrieval Environment: Integration of Information Retrieval with Group Support Systems. *Proceedings of the 32nd Hawaii International Conference on System Science* (HICSS-32), 1999.

SPRO Annual Overview for fiscal year 2004. Protect your ideas (2004). Available at http://www.prv.se/english/pdf/PRV_year_04_eng.pdf. Patent- och Registrerings-verket: Stockholm. Accessed 2006-08-07.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In: Fox, E., Ingwersen, P. and Fidel, R. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 138-146.

Saracevic, T. (1996). Relevance reconsidered ´96. In: Ingwersen, P. and Pors, N. O. (Eds.) *Second International Conference on Conceptions of Library and information Science: Integration in Perspective*, Copenhagen, Denmark. Oct. 13-16, 1996. Copenhagen, Denmark: The Royal School of Librarianship, 201-218.

Saracevic, T., Kanto, P., Chamis, A. Y. & Trivison, D. (1988). A study of information seeking and retrieval. 1. Background and methodology. *Journal of the American Society for Information Science* 39, 161-176.

Savolainen, R. (2009). Information use and information processing: Comparison of conceptualizations. *Journal of Documentation* 65(2), 187-207.

Savolainen, R. (2007). Filtering and withdrawing: strategies for coping with information overload in everyday contexts. *Journal of Information Science* 33(5), 611-621.

Schamber, L., Eisenberg, M.B. & Nilan, M.S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing & Management* 26(6), 755–776.

Seale, Clive (1999). *The quality of qualitative research*. Sage publications: London.

Serola, S. (2006). City planners' Information Seeking Behavior: Information Channels Used and Information Types Needed in Varying Types of Perceived work Tasks. In: Ruthven, I. & et al. (Eds.) *Information Interaction in Context*, IiiX, Copenhagen: ACM Press, 42-45.

Siegel, S. & Castellan Jr, N. J. (1988). *Nonparametric statistics for the behavioural sciences*. 2nd edition. London: McGraw-Hill.

Smith, W., Hill, B., Long, J. & Whitefield, A. (1997). A design-oriented framework for modelling the planning and control of multiple tasks work in secretarial office administration. *Behaviour & information technology,* 1997, Vol. 16, no. 3, 161-183.

Solomon, P. (1997). Discovering information behavior in sense making. II. The social. *Journal of the American Society for Information Science 48(12),* 1109-1126.

Sonnenwald, D. & Lievrouw, L. (1997). Collaboration during the design process: a case study of communication, information behaviour, and project performance. In: Vakkari, P. & Savolainen, R. & Dervin, B. (Eds.) *Information Seeking in Context*, London: Taylor Graham, 279-204.

Sonnenwald, D. H. & Pierce, L.G. (2000). Information behaviour in dynamic group work contexts: interwoven situational awareness, dense social networks and contested collaboration in command and control. *Information Processing & Management* 36(3): 461-479.

Spink, A. (2004). Multitasking information behaviour and information task switching: An exploratory study. *Journal of Documentation* 60(4), 336-345.

Spink, A. & Greisdorf, H. (2001). Regions and Levels: Measuring and Mapping Users' Relevance Judgments. *Journal of the American Society for Information Science*, 52, 161-173.

Strauss, A. & Corbin, J. (1998). Bas*ics of Qualitative research. Techniques and Procedures for Developing Grounded Theory*, (Second edition). Sage Publ.: Thousand Oaks.

Talja, S. (2002): Information sharing in academic communities: Types and levels of collaboration in information seeking and use. *New Review of Information Behavior Research 3*, 143-159.

Talja, S. & Hansen, P. (2006) Information Sharing. In: A. Spink and C. Cole (Eds.) *New Directions in Human Information Behavior*. Information Science and Knowledge Management Series no. 8. Dordrecht: Springer Verlag. 113-134.

Tang, A., Tory, M., Po, B., Neumann, P. & Carpendale, S. (2006). Collaborative Coupling over Tabletop Displays. *Proceedings of the SIGCHI 2006 Conference on Human Factors in computing systems*, April 22–27, 2006, Montréal, Québec, Canada, 1181 – 1190.

Taylor, R. S. (1968). Question-negotiation and information seeking in libraries. *College and Research Libraries* 29, 178-194.

Tseng, Y-H., Lin, C-J. & Lin, Y-I (2007). Text mining techniques for patent analysis. *Information Processing and Management* 43 (2007), 1216-1247.

Tseng, Y-H. & Wu, Y-J. (2008). A Study of Search Tactics for Patentability Search A Case Study on Patent Engineers. In: Tait, J. (Ed.). *Proceeding of the 1st ACM workshop on Patent information retrieval, PaIR'08*. Napa Valley, USA. 33-36.

WIPO, IFIA & BUD, (1998). *Patent Information in Support of Inventive and Innovative Activities: General Introduction*. WIPO/IFIABUD/98/2, March 1998. Available at http://www.wipo.int/edocs/mdocs/innovation/en/wipo_ifia_bud_98/wipo_ifia_bud_98_2.doc Accessed 2009-12-14.

WIPO 2009: International Patent Classification (2009). Guide. 8th version. Available at http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc_2009.pdf. Accessed 2010-09-28.

Vakkari, P. (2003). Task-based information searching. *Annual Review of Information Science and Technology* Vol. 37 413-464.

Vakkari, P. (2001a). A theory of the task-based information retrieval process: a summary and generalization of a longitudinal study. *Journal of Documentation* 57(1), 44-60.

Vakkari, P. (2001b). Changes in search tactics and relevance judgments in preparing a research proposal: A summary of findings of a longitudinal study. *Information retrieval* 4(3/4), 295-310.

Vakkari, P. (2000a). Cognition and Changes of Search Terms and Tactics during Task Performance: A Longitudinal Study. In *Proceedings of the RIAO'2000 Conference*. Paris: C.I.D., 894-907.

Vakkari, P. (2000b). Relevance and contributing information types of searched documents in task performance. In: Belkin, N. et al. (Eds.). *Proceedings of the 23rd annual international ACM SIGIR* conference on *Research and development in Information Retrieval*, 2-9. New York: ACM Press.

Vakkari, P. (1999). Task Complexity, problem structure and information actions: Integrating studies on information seeking and retrieval. *Information Processing & Management* 35 (6), 819-837.

Vakkari, P. (1998). Growth of Theories on Information Seeking. *Information Processing & Management* 34(3/4), 361-382.

Vakkari, P. and Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. *Journal of Documentation* 56(5), 540-562.

Wang. P. (1997). User's information needs at different stage of a research project: a cognitive view. In Vakkari, P., Savolainen, R., & Dervin, B. (Eds.) *Information Seeking in Context*, 307-318. London: Taylor Graham.

Veinot, T. C. (2009). Interactive Acquisition and Sharing: Understanding the Dynamics of HIV/AIDS Information networks. *Journal of the American Society for Information Science and Technology* 60(11), 2313-2332.

Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation* 55 (3), 249-270.

Wilson, T. D. (1997). Information behaviour: An interdisciplinary perspective. *Information Processing & Management*, Vol. 33, No. 4, 551-572.

Wilson, T. D. (1981). On user studies and information needs. *Journal of Documentation* 37(1), 3-15.

Wilson, P. (1973). Situational relevance. *Information Storage and Retrieval* 9, 457-471.

Wixon, D., Holzblatt, K. & Knox, S. (1990). Contextual design: An emergent view of system design. In: Carrasco, J and Whiteside, J (Eds.). *Proceedings of Human factors in computing systems (CHI'90)*, 329-336. New York: ACM Press.

Zhu, J. & Tait, J. (2008). A proposal for Chemical Information Retrieval Evaluation. In. *Proceedings of the 1st international workshop on Patent information retrieval* (PaIR'08), Napa Walley, California, USA, 15-18.

APPENDICES

Appendix A: Classification of variables by task level
(see Chapter 3)
Variables, descriptions and values of the task level empirically identified in study.

Work task		
Variable	**Description**	**Range of attributes**
Category 1a: Patent application		
Type of patent application	Type of patent application	International National
Patent applicant	The patent work task externally initiated	Company private person
Patent work task initiation - Externally	Author of patent application	Company Bureau Private person
Patent work task initiation - Internally	Assignment of patent WT within a group	Group-leader Colleague
Category 1b: Task performer / Actor		
Goal of work task	Individually perceived goal for a patent WT	Organisational Group Individual
Perceived task difficulty (Task knowledge)	The WT is perceived by the performer as different levels of difficulty	Low Medium High
Domain knowledge	A WT is judged by the performer as being within, partly within or outside own domain knowledge	Within Within/Outside Outside
Work task constraints	Organizational constraints when performing a patent WT	Cost IT support legislation IPR
	Personal constraints when performing a patent WT	Interruptions by colleague Shortage of time team support
Category 1c: Patent work task		
Work task structuring	The WT is planned and structured by the performer	Structured Unstructured
Task completion time	Amount of time units used for WT completion	Hours Days

IS&R task		
Variable	**Description**	**Range of attributes**
Category 2: Information need		
Problem need formulation	The overall problem to be solved are perceived as clear in various levels	Clear Muddled
Problem need formulation	The overall problem to be solved are perceived as clear in various levels	Clear Muddled
Perceived Information need clarity	The search task has a clear or an unclear structure regarding the perceived information need	Clear Unclear
Planned Information	The search task is planned in a structured/unstructured way	Structured

need structure	by the performer	Unstructured
Information need decomposition	The information need is decomposed as regard to the overall WT	WT decomposition IS&R decomposition
Information need change	The information need is changed during the course of the task performance process	Yes No
Expressed information needs	The amount of information need expressed for search task	Single Multiple
Expressed information need	The information needed for task completion has been expressed in a narrative way	Narrative Terms
Document component needed for formulation of information need	The requirements of the search task are retrieved text/image objects such as requirements for formulation	Text objects (Sections, References, Terms, classification codes) Image objects (Images, structures)
Type of information needed	Different requirements for retrieval for the search task	Text Image
Category 3: Source		
Source	The number of sources used within the search task	Number of
Source combination	Type of combination of sources in order to complete the search task	Paper Human Electronic
Source	Numbers of search session made for each source	Numbers
Source content type	Different types of contents of the sources used	Classifications Code ; Full text; Abstract; Bibliographic; Images; Dictionary; Lexicon
Category 4: Query formulation		
Query	The numbers of unique terms used in a query string	A-priori expressed terms
Query	The number of different query elements used in a query string	Terms; classification codes; doc-id; synonyms; dates; country; structures (biomed); Images
Query synonyms	Number of synonyms used in a query string	Numbers
Query	Average number of query terms per string	Numbers
Query	The number of combinations of query elements used in a query	Types of combinations
Query	The number of unique of classification codes used per task	Numbers
Category 5: Relevance judgment		
Relevance Judgment	RJ used in the work task initiation, information seeking task and information retrieval task of the task performance process	TI IST IRT
Relevance Judgment tactic	Different types of RJ strategies within the same search task performance resulting in several sets of documents judged	Sequential Aggregated
Relevance Judgment	Type of document elements judged for relevance	Summary; Figure; Claims; Description; Term; classification codes; bibliographic
Category 6: Information use		
Information use	Type of components of information objects actually used to complete the task	Paragraphs Images Classification codes Abstracts
Category 7: Collaborative Information handling		
User effort	User efforts used to complete WT	Single Collaborative

Appendix B: Interview form (see Chapter 4)

Interview "Guide"

Procedure of the interview:
The Interview will follow an open and "theme-based procedure. The questions will be posed as naturally as possible and we will rather use an informal discussion mode talking around these questions. Since I have previously done two visits, I have acquired some knowledge about the general work and situations. The interview will contain questions about the participant (background and experience etc.), followed by questions concerning different knowledge levels and finally questions about the domain/context (goals, roles, constraints, applications etc).
The interview will be tape-recorded and notes will be written down. These will later on be transcribed and analysed.

Pilot: 10-11 May 2000
Revised Interview Scheme: 13 May 2000

INTERVIEW FORM

Name:
Date:

Background: User/Actor-related

DEMOGRAPHIC
Education
Question (1): What kind of education do you have?
Vilken utbildning har du

- External: Graduate/Post-graduate
- Internal: Courses
- Domain: Engineering, Physics

Domain experience
Question (2): How many years have you been working
- at SPRO and
- in your present position?

ACTORS
Reason: Position in organisation and work-domain area.
Question (3): What is your position?
Question (4): In what area is your expertise and in what area do you work?

TYPES OF TASK EXPERIENCE
Questions:
- (5) How many years have you worked with this type of work (PA)?
- (6) How long have you worked in this specific subject-area?
- (7) Have you worked in other subject areas as well?
- (8) What are the characteristics of your work-tasks?

ISR KNOWLEDGE

Questions:
- (9) Which databases and systems do you use regularly?
- (10) How frequently do you use them?
- (11) What IR-related courses have you attended and when?
- (12) How many years have you worked with IR-related activities?
- (13) Could you describe the information seeking and retrieval activities? Do you use any specific method/strategy (individually developed or recommended by the organisation)?

TYPES OF INFORMATION KNOWLEDGE
Question (14): What types of information are important for a PA task and what are the requirements regarding different types of information in order to perform your tasks?

TYPES OF INFORMATION SOURCE KNOWLEDGE
Questions:
- (15) What types of information sources do you use in a problem-solving task?
- (16) What are the requirements regarding the access to different information sources (paper/humans/ electronic) in order to perform your task?
- (17) What are the specific requirements regarding the access to electronic information sources specifically (databases and IR-systems) within SPRO?

<u>Organisation/Context/ Domain-related</u>

GOAL (S)
Reason: Organisational goal(s)/Group-related goal(s)/Individual goal(s) perceived /understood by the actor.
Questions(18): What are the specific goal(s) and tasks related to your work?

ROLES
Question (19): Describe how your work is related to your group/department and organisation as well as to external patent organisations

CONSTRAINTS
Reason: External factors that are affecting the specific task performance process, such as the official requirements, time, place, client requirements, situations, actors….
Question(20): Describe what difficulties and problematic factors might affect the accomplishment of your typical tasks

COOPERATION
Reason: Description of types of collaborations in the task performing process.
Question(21): Describe types of collaborations in the task performing process.

PATENT APPLICATIONS
Reason: To acquire the set of official and labelled patent applications that the respondents work with as well as other possible tasks that occur.
Question(22): Describe the patent applications that you generally are working with at SPRO and those that are most common in your own work.

PATENT APPLICANTS
Reason: Description of different types of applicants such as individuals and company and their characteristics.
Question(23): Please, could you describe the different types of applicants that send in applications and their characteristics

OTHER QUESTIONS

Appendix C: Note form for observation
(see Chapter 4)

D: NOTE FORM FOR OBSERVATION

Date:
Time started:
Participant:
Patent Application (PA) ID:
Type of Patent Application (A1):
Stage of process in the Patent Application (A2)

Notes on Tasks performance

Task Planning/Construction (PA)

Information Need

Knowledge Sources

Type of Information

Information Seeking Strategies

IR techniques and Process

Inferences
(questions around focus items and others situational-based events)

General physical work environment.

A3. Describe the task regarding the construction of the task performance process and information seeking process?

A4. Describe the task according to difficulty/complexity/duration etc

A5. What are the Requirements? regarding Task structuring and planning?

A6. Describe how you formulate or construct your information need?

A7. How would you describe your information need in order to solve the patent application task?

A8. What is the current information need?

A9. Why are you using this information source at this stage in the process?

A10. What type of information or information units are you looking for?

A11. Describe why you are taking this path in the seeking process?

OB1. Types of general search Techniques

OB2. Types of request Formulation

OB3. Types of queries Formulation

OB4. Types of reformulation

Relevance and Success

A12. Describe what is the relevant information in this stage of the task performance process?

Usage

A13. In what way will the information retrieved be used in different stages in the task performance process (During and after)?

Other

Appendix D: Task performance electronic diary activity log (see Section 4.1)

DIARY OF TASK PERFORMANCE PROCESS

BACKGROUND OF TASK

Date:	
Participant:	
Time started:	
Patent Application (PA) ID (to keep track of them during time):	
Type of Patent Application:	
Current stage of process in the Patent Application:	
Time ended (with this specific task)	

DIARY FORM

TASK STAGES	Descriptions of Actions/Activities
INITIATION	
History of the current Patent Application	
Preparation of current task	
CONSTRUCTION	
Formulation of the YOUR Problem-Solving task and Information Need Formation	
PLANNING	
Planning of overall Task Performance Strategy	
Planning Information Seeking and Retrieval Strategy	
TASK PERFORMANCE	
Information Seeking Task Name sources (human/paper) and information types of interest	
Relevance Judgment	
Information Need	
Information Retrieval task Name sources (electronic) and Mark information types of interest	
Relevance Judgment (mark docs relevant to claims/information need)	
COMPLETION	
Stopping	
Relevance Judgment (mark docs relevant to claims/information need)	
Information Usage	

Appendix E: Task-based protocol for data analysis, Section A: Internal task information
(see Section 4.2)

Section A: Internal Task Information.

Task No.				
Participant (Name)				
Task category				
Collected data for analysis	Interviews (written notes + recordings	Diary (Log-statistics + written statements)	Observation (written notes)	Post-interview (e-mail correspondence + complementary documents)
Official PRV documents for analysis	G-note (granskningslapp)	Reply-document	Written opinion (WO)	

Section B: Main variables, indicators and categories with values

1. Organizational/Contextual Setting

a. Type of applicants (%)	Company/Organization			Private					
b. Application preparation	Company			Private			Bureau		
c. Types of incoming applications (formal)	A	C	A+ITS	PCT1	PCT2	Class	Assign.	B	other
d. Collaboration									
i. Type of setting knowledge/awareness	Internal				External				
ii. Type of information knowledge/awareness (requirement and need)	Information related to search process				Information related to task performance procedure/structure				
iii. Type of channel knowledge/awareness	Individual/group: within own subject area	Individual/group: other, task-topic related subject area			Media: electronically-based information		Media: paper-based information		

2. Task and task performance process

a. Task type	A (ISR)	B (only Retrieval)		C (Classification)	D (Other)	
c. What are the task performance stages?	Task initiation phase	Task preparation and structuring phase	Information seeking phase	Information retrieval phase	Relevance judgment phase	Task completion phase
d. Topic of Task	Within examiners subject area			Outside examiners subject area		
e. Perceived Task difficulty/complexity	Easy			Difficult		
	Structured			Unstructured		
ii.						
g. Information (incoming application) requirement characteristics/state	Clear			Muddled		
i.						
h. Stage of current task in the overall task performance process?	Initial phase (Beginning)		Interactive and answering phase (Middle)		Concluding phase (End)	

3. Magnitude/Scope

a. Constraints during task performance	Limited time to accomplish task	Other non IR related tasks (interruptions)	Support/expert help with problem solving	Database connection problems	Cost	Other
b. Effective hours spent on task						
c. Completion time						

4. User preferences

a. Education	Graduate			Post-Graduate	
b. Domain (patent) experience (years)	0-2	2-5	5-10	10-20	20 -
c. Subject area experience	0-2	2-5	5-10	10-20	20 -
d. IR knowledge	Pre-PRV IR Knowledge	PRV 18 month intro course	Update DB-courses	Internal group discussion	
e. Role and within PRV	Examiner	Guru-group	Supervisor	Other	

5. Information need

a. Type of information need (perceived)	Stable/clear	Stable/Unclear	Unstable/clear	Unstable/ unclear
i.				
	Structured		Unstructured	
ii.				
b. How is the information need formulated (output)?	Descriptive (Sentence, phrase and their relations)		Non-descriptive (Terms, concepts and class. codes)	
i.				
	Short (1-2 sentences)		Long	
ii.				
	Single aspect		Multiple aspects	
iii.				
c. Type of information need representation	Text	Image	Text/Image	

d. Requirements for information need formulation (input) (as stated in "preparation of current task"	Document	Section	Reference	Term	Image	Codes On application
e. How many terms are expressed in the information need description?						
f. Information need change?	Yes		No			

6. Sources

a. Numbers of sources used							
b. Type of sources	Paper-based		Human-based		Electronic-based		
c. Type of source content	Classification codes (Thesaurus)	Full text	Abstract	Bibliographic	Image	Dictionaries	Lexicon

7. Retrieval Strategies

a. Information need decomposition	Related to source				Related to information need structure					
b. Type of query-terms	Keyword		Synonyms		Classification codes		Document number			
c. Numbers of unique query-terms used in task										
d. Numbers of synonyms used in task										
e. Numbers of items per query string	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10

f. Term and info need-term relationship matrix	INTerm- Qterm match	INTerm - Qterm no match	INTerm-synonyms match	INTerm-synonyms no match	other

g. Query structure (Use of operators)	Structured	Unstructured

h. Number of combinations of term-concept-codes	Single item e.g. term or code etc.	Items of same type, e.g. term/tem or code/code, etc.	DocID	Term-classification code
Numbers of unique classification codes entries				

8. Relevance Judgment

a. When is RJ performed within the task performance process?	Task initiation stage Beginning	Information seeking stage	IR stage Middle	Task completion End
b. Number of RJ made during the task performance process				
c. How is relevance judged?	Sequenced		Concatenated	

d. What parts in the retrieved information is judged relevant?	Summary/ Abstract	Full Image/ Figures	Part of Image/ Figure	Claims	References	Bib-info	Description	Terms	Class-ification codes
SS1-n									

e. What degree of relevance is assigned to the relevant documents? (Tague 48)	A doc that matches whole info need (X) = (a=a)	2 docs that matches whole info need (Y) = (a=b)	Covers a similar but not the same info need (A) = (b > a)
SS1-RJ1-n RJ6-A-report RJ6-ITS-report			
f. Combinations of different relevance judgments (x,y,a)			
SS1-n RJ6			

9. Use of information retrieved

a. What type of information is used?	PA documents	Articles	Web documents	Images	Classification codes

b. What type of information elements is used?	Images	Paragraphs	Abstract	Sentences	References	Terms	Classification codes

c. When in the task performance process is the retrieved information used?	Task Initiation stage	Information Seeking stage	Information retrieval stage	Task completion stage

d. How is the retrieved information used?	Produce First Prel. Report	Produce Final Report	Produce Second Prel. Report	Produce Written Opinion	Produce Internal Search Note
i.					

Retrieved information used in relation to…	Describing overview	Topic (classification)	Describing details	Describing a method	Describing an object	Describing usage
ii.						

Appendix F: Excerpt from the matrix (see Chapter 4)

The Matrix – an excerpt of variable number 8 (Retrieval Strategy) with indicators 8c, 8d, and 8e with its values. To the left is the number of the task.

	8c No. of unique query-terms	8d No. of synonyms	8e No. of terms per session and query string
1	*	*	*
6	13	4	3.71
15	10	19	7.71
22	*	*	*
25	NA	NA	NA
28	18	2	3.31
33	19	12	2.27
34	7	3	2.77
35	NA	NA	NA
36	7	0	1.93
37	23 + STRUCT	8	5.00
43	5	4	2.97
46	*	*	*
49	13	22	8.58
50	*	*	*
52	9	4	1.8S
54	7	2	2.09
61	12	1	2.07

Legend: NA=not applicable *= no data collected.

2cnew Knowledge of task topic
2f Clear/Muddled requirement need
2e* Clear/Unclear Perceived information need
2e** Structured/Unstructured Perceived information need
2d Perceived task difficulty
2e Perceived Information need (combined 2* and 2**)
5b Expressed information need (long-short)
5b* Expressed Information need (narrative – non narrative)
5b** Expressed information need (multiple - single)
5d Document components needed for information need formulation
5c Type of information needed (text - image)
5e Number of expressed terms
5f Information need change
6a Number of sources
6b Source type consulted
6c Number of content types used
6dii Number of search strings used
6do Number of search "items" requested
6d# Average number of search items/strings
6e Number of search sessions per task
8a Task or /and information need decomposition
8b Type of query term
8c Number of query term
8d Number of synonyms
8f Perceived information need term and query match
8hi Number of combinations of query types
8hii Average number of query types combination
8i Number of unique codes
9ai RJ application in TPP stages
9aii RJ applications in TPP stages
9bi Average number of RJ in TPP
9bii Number of RJ in TPP
9c Application of RJ (sequential - aggregated)
9d Average number of elements judges for relevance
9dii Number of different elements judged in single task

10a Information sources used for final product
10bi Type of information component used in final product
10dii Part of information used from PA in new product

Appendix G: Results: Descriptive analysis of work and IS&R task processes (see Chapter 6)

Table 6.1: Categorization of domain goals and their frequency

Domain level	Domain goals
Organisational level	g) supporting the development and growth of Swedish industry and further development of the patent domains; (7)[62] h) providing applicants with high quality searches and services (5) i) protecting ideas (4); j) helping applicants (2); k) disseminating information and knowledge (1) l) quick answers (1)
Group/Team level	d) creating and developing praxis and consensus within the work team as regard to judgments, education, etc. (3) e) if necessary, providing information, knowledge and protection of applied invention (2); and f) processing as many applications as possible (2)
Individual level	f) give each application a good qualitative judgment; (6) g) find what is already known and therefore not to accept redundant applications or parts thereof thus identifying patents that really are unique and therefore possible as applications; and (5) h) give the applicant a strong protection for his ideas. (5) i) provide patent searching as a service; (4) j) support the development and growth of Swedish industry (1)

Table 6.2: Distribution of types of application preparation by task types in terms of patent applications made (N=49)

Application preparation	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*	*Total*	*Group 1 A+PCT2 +AS +C*	*Group 2 PCT1+A +ITS*
Company	0	8	0	5	0	0	13	0	13
Private	2	0	0	0	0	0	2	2	0
Bureau	9	2	7	3	6	7	34	29	5
N=	11	10	7	8	6	7	49	31	18

Table 6.3: Distribution of type of task constraints (descendent order) across task type.

Type of Constrain	*A %*	*PCT1 %*	*PCT2 %*	*A+ITS %*	*AS %*	*C %*
Interruptions (visitors, internal meetings, courses etc)	14	41	100	23	8	0
Time	27	14	0	23	33	0
IT connections	27	18	0	18	17	0
Problem Solving support	18	14	0	6	17	0
Cost	9	4	0	12	17	0
Other	5	9	0	18	8	0
	100	100	100	100	100	0
N=74	22	22	1	17	12	0

Table 6.4: Distribution of completion time (days) by number of tasks.

Completion time (days)	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*	*Σ*
1	1	3	4	0	3	0	11
2	10	2	4	2	2	7	27
3-4	2	5	0	3	0	0	33
5-6	0	1	0	2	0	0	17
7-8	0	0	0	2	0	0	14
9-10	0	1	0	0	0	0	9
Total days	29	37	12	40	7	14	139
Average # of days/task	2,23	3,08	1,50	4,45	1,40	2,00	2,57

[62] The numbers in brackets, for example (7), mean that 7 out of 10 professional patent examiners made comments on this particular aspect.

% of total days	21	27	8	29	5	10	100
n=tasks	13	12	8	9	5	7	54

Table 6.5: Distribution of completion time (hours) by number of tasks.

Completion time (hours)	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*	*total*
1-4 (2.5)*	4 (1**)		8 (2)	3 (1)	8 (3)	12 (6)	3 (1)
5-8 (6.5)	12 (2)	22 (3)	33 (6)	6 (1)	6 (1)	5 (1)	14 (1)
9-12 (10.5)	54 (5)	12 (1)	0	9 (1)	0	0	7 (1)
13-16 (14.5)	46 (3)	14 (1)	0	28 (2)	0	0	6 (1)
17-26 (21.5)	44 (2)	186 (7)	0	90 (4)	0	0	13 (1)
Total hours	160	234	41	136	14	17	602
Hours/task	12,31	19,5	5,12	15,11	4,67	2,43	11,14
n tasks	13	12	8	9	4	7	53

* = class midpoint; **= number of tasks

Table 6.6: Distribution of perceived overall work task difficulty (task knowledge) by task type (N=52)

Task knowledge	*A*		*PCT1*		*PCT2*		*A+ITS*		*AS*		*C*		Σ	
Easy (E)	7	58	6	54	7	78	6	67	1	25	7	100	34	65
Difficult (D)	3	25	5	46	2	22	1	11	3	75	0	0	14	27
E/D	2	17	0	0	0	0	2	22	0	0	0	0	4	8
%		100		100		100		100		100		100		100
total	12		11		9		9		4		7		52	

Table 6.7: Distribution (percentage) of task structuring by task type (N=54).

Type of task structuring	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
Structured	62	73	100	78	80	100
Unstructured	23	9	0	22	20	0
Unclassifiable	15	18	0	0	0	0
%	100	100	100	100	100	100
N	13	11	9	9	5	7

Table 6.8: Distribution of problem formulation clarity of information needed across task types. (N=53)

Problem formulation clarity	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
Clear problem formulation	8 (67)	9 (81)	7 (78)	3 (33)	4 (80)	7 (100)
Muddled problem formulation	3 (25)	2 (19)	2 (22)	5 (56)	1 (20)	0 (0)
Clear/Muddled problem formulation	1 (8)	0 (0)	0 (0)	1 (1)	0 (0)	0 (0)
Total	12	11	9	9	5	7

Table 6.9: Distribution of patent engineer domain knowledge across task type. N=54

Engineer domain knowledge	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
A-within own topic area (a)	39	36	100	89	40	100
Outside own topic area (b)	46	28	0	11	40	0
Combination of within/Outside own topic area (c)	15	36	0	0	20	0
%	100	100	100	100	100	100
Wholly or partially outside (b+c)	62	64	0	11	60	0
N=54	13	11	9	9	5	7

Table 6.10. Distribution of unique CIR activities by type across individual PA tasks

	Type of CIR activity		
Task	Document-related	Human-related	Total
1	7	2	9
2	9	7	16
3	2	3	5
4	11	2	13
5	15	3	18
6	15	4	19
7	6	5	11

8	12	5	17
9	7	8	15
10	10	10	20
11	4	6	10
12	2	0	2
Total	100	55	155
Mean	8,33	4,58	12,92
%	65	35	100
N=54			

Table 6.11: Distribution of perceived clarity of information needed by task type

Perceived clarity of information needed	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
SC	4 (31)	9 (81)	7 (78)	1 (11)	1 (20)	7 (100)
SU-UC-UU	9 (69)	2 (19)	2 (22)	8 (89)	4 (80)	0 (0)
N=54	13 (100)	11 (100)	9 (100)	9 (100)	5 (100)	7 (100)

Legend: sc=structured/clear; su=structured/unclear;
uc=unstructured/clear; unstructured/unclear

Table 6.12: Distribution of planning information need structuring across task type. (N=47)

Structuring of information need planning	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
Structured	4 (36)	9 (82)	7 (78)	1 (20)	1 (25)	7 (100)
Unstructured	7 (64)	2 (18)	2 (22)	4 (80)	3 (75)	0 (0)
N=tasks (%)	11 (100)	11 (100)	9 (100)	5 (100)	4 (100)	7 (100)

Table 6.13: Distribution of change of information need by task type. (N=47)

Information need change	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
Change	38	0	0	75	0	0
No change	62	100	100	25	100	100
N=tasks	13	11	9	8	5	1
%	100	100	100	100	100	100

Table 6.14: Distribution of information need decomposition by task type. (N=36)

Information need decomposition	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
% of information need decomposition related to overall search task performance	100*	27	0	50	50	0
N=tasks	12	11	2	8	2	1
% of information need decompositions related to the information need	100	100	100	100	100	100
N=tasks	12	11	2	8	2	1

Legend: *=percentage

Table 6.15: Distribution of information need expression across task type in terms of single or multiple information need stated (N=44).

Expressed information need as single or multiple need	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
Single	0	30	0	11	0	100
Multiple	100	70	100	89	100	0
N=tasks	13	10	2	9	3	7

Table 6.16: Distribution of PA components by task type in terms of components needed for information need formulation. (N=54)

PA Document components needed for information need formulation	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*	
Sections	13,0	19,4	24,0	12,5	12,5	47,0	128,4
References	21,0	5,0	28,0	12,5	25,0	0	91,5
Terms	26,0	25,0	0	25,0	12,5	0	88,5
Classification codes	29,0	31,0	16,0	33,3	37,5	47,0	193,8
Images	11,0	19,4	32,0	16,6	12,5	6,0	97,5
N	38	36	25	24	8	15	146
%	100%	100%	100%	100%	100%	100%	

Table 6.17: Distribution of number of sources selections by task type. (N=54)

Source selection	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*
Number of sources[63] selected	86	62	10	56	20	7
Average of sources selected	6,61	5,6	1,1	6,2	4,0	1,0
N tasks	13	11	9	9	5	7

Table 6.18: Distribution of source type selection by task type (N=54)

Source type	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*	
P			3			7	10
H							0
E	3	3	5	1	1		13
PH			1				1
HE	2						2
PE	2	2		3	2		9
PHE	6	6		5	2		19
Total / Av. Source types per task	29/**2,23**	25/**2.27**	10/**1.11**	22/**2.44**	11/**2.20**	7/**1.00**	
# tasks	13	11	9	9	5	7	54

Legend: P = paper-based source; H = Human source; E = Electronic source

Table 6.19: Distribution of types of source content by task type in terms of number of source elements used. (N=179)

Types of source content	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*	*%*
Classifications Code	22	18	0	23	29	0	19
Full text	21	23	62	18	29	78	27
Abstract	21	16	15	18	23	11	18
Bibliographic	7	18	0	16	0	0	10
Images	15	20	23	16	12	11	17
Dictionary	12	2	0	3	6	0	6
Lexicon	2	2	0	5	0	0	2
%	100	100	100	100	100	100	100
N=	32	25	7	21	9	5	100

Table 6.20: Distribution of number of unique terms expressed prior to search by task type (N=34).

Number of unique terms expressed prior to the search	*A*	*PCT1*	*PCT2*	*A+ITS*	*AS*	*C*	*Tot*
# Tasks	*13*	*10*	n.a.	*8*	*3*	n.a.	*34*
Terms	80	52	n.a.	65	13	n.a.	210
Average terms/task	6.15	4.72	n.a.	8.12	4.33	n.a.	6.17

Table 6.21: Distribution of all terms and synonyms per session used by task type. (N=36)

Terms/synonyms per session	**A**	**PCT1**	**PCT2**	**A+ITS**	**AS**	**C**
Average term/session[64]	6.66	4,27	n.a.	11.77	3.33	n.a.
Average synonyms/session	1,58	1.45	n.a.	8.11	0.66	n.a.
Total average term +synonyms/session	8.25	5.72	n.a.	19.88	4.00	n.a.

Table 6.22: Distribution of terms used per session and query string by task type. (N=30)

	Average String per session	*Average term per string*	*N*	*Range*
A	7.66	2.01	11	1,56 <-> 2,77
PCT1	11.44	1.95	10	1,13 <-> 8,58

[63] With selected sources we mean that we did not count repeated use of the same type of source within a task solving performance.

[64] In this context, a session is characterized by a starting point and an ending of an electronic search query.

A+ITS	9.15	4.11	9	1,56 <->7,71
AS	6.75	1.61	3	1,00 <-> 1,93
C	0	0	0	0
PCT2	0	0	0	0
			36	

Table 6.23: Distribution of the combination of query and term usage by task type (N=34[65])

Average of combinations		*A*	*%*	*PCT1*	*%*	*A+ITS*	*%*	*AS*	*%*
Double term or code (D)		12,8 (10)	49	8,45 (11)	42	14,71 (7)	37	0	0
	Range	7-20		1-22		2-29		-	
Double Doc (Doc)		3,00 (8)	11	2,00 (5)	10	3,00 (5)	8	1 (1)	14
	Range	1-7		1-3		1-6			
Term/Code (Tcc)		3,16 (6)	12	2,66 (6)	13	2,00 (4)	5	2,00 (3)	29
	Range	1-6		1-8		1-3		1-3	
Doc-Code (Dcc)		0	0	4 (2)	20	3 (2)	8	0	0
Author (A)		0	0	2	10	0	0	0	0
Term/doc (TD)		1	4	1	5	0	0	1	14
Term-Country (Tco)		2	8	0	0	0	0	0	0
Cite/Cited (CC)		2 (2)	8	0	0	4,5 (2)	11	2	29
Combi Com)		0	0	0	0	1	3	0	0
Author/code (ACC)		0	0	0	0	2	5	0	0
Term/Author (TA)		1	4	0	0	2	5	0	0
Term /Structure (Tstru)		1	4	0	0	8 (2)	20	1	14
		25,96	99,5	20,11	100	40,21	100	7.00	100
Total # tasks		12		11		8		3	

Legend (most used combinations): double= 2 terms or 2 classification codes in a query; doc= document ID in a query; Tcc= term and classification code combination in a query; Country=country name; cite/cited= document that are cited within another document or have citations to other documents; combi=a combination of ; Structure=structures of, for example, chemical compounds; *= number of tasks

Table 6.24: Distribution of unique classification codes used per task by task type. N=35

Classification codes used per task	*A*	*PCT1*	*A+ITS*	*AS*
total	51	43	45	14
average	3,93	4,3	5	4,67
N=tasks	13	10	9	3

Table 6.25: Distribution of the stage of RJ by task type in terms of numbers of relevance judgments made during task stages.

Relevance: Relevance judgments in TPP stages	*A-task %*	*PCT1 %*	*PCT2 %*	*A+ITS %*	*AS %*	*C-task %*	*Total %*
TI stage (e.g. task initiation and preparation)	6	13	9	7	13	0	9
IST stage (e.g. information gathering (e.g. human-human, Information need formulation, source selection)	15	9	9	7	20	0	11
IRT stage (e.g. query formulation, relevance judgment)	79	78	82	87	67	100	80
TI+IST total	21%	22%	18%	14%	33%	0	20%
N (judgments)	33	23	11	15	15	7	104
N (tasks)	13	11	9	9	5	7	54

Legend: TI=task initiation stage; IST=information seeking task stage; IRT=information retrieval task stage

Table 6.26: Distribution of relevance judgment strategy application across task type (N=52)

	A	*%*	*PCT1*	*%*	*PCT2*	*%*	*A+ITS*	*%*	*AS*	*%*	*C*	*%*
Sequenced relevance judgment	13	100	4	28	0	0	7	100	4	80	0	0
Aggregated relevance judgments	12	92	11	100	9	100	7	100	3	60	7	100

[65] The low number of tasks in this table is due to no data for task types PCT2 and C

N=tasks	13	11	9	7	5	7

Table 6.27: Distribution (%) of type of document elements judged for relevance across task type. (N=52)

Types of document elements used	*A %*	*PCT1 %*	*PCT2 %*	*A+ITS %*	*AS %*	*C %*
Summary	16	24	17	18	17	5,5
Full Figure or part of Figure	21	17	26	18	17	11
Claim	16	13	20	14	17	5,5
Description	16	24	20	21	25	5,5
Term	10	11	0	11	0	0
Classification code	14	6	2	14	17	39
Reference	7	2	15	4	8	34
Bibliographic	0	2	0	0	0	0
	100	100	100	100	100	100
n=221	71	46	46	28	12	18
Average number of judged components per task	5.46	4.18	5.11	4.00	2.40	2.57
N (tasks)	13	11	9	7	5	7

Table 6.28: Distribution of number of information sources used across task type.

Information sources used	*A %*	*PCT1 %*	*PCT2 %*	*A+ITS %*	*AS %*	*C %*
Patent Application (PA) online	100	100	100	100	100	100
Article (online)	15	25	11	62	0	0
Web-page	15	8	0	25	0	0
Image	62	75	67	87	3	1
Classification code	100	6	11	62	3	1
Bibliographic	0	0	0	37	0	0
N tasks=54	13	12	9	8	5	7
N information types=137	38	32	17	30	11	9
Average information type/task	2.9	2.6	1.9	3.7	2.2	1.3

Table 6.29: Distribution of information components used by task type in terms of average percentage of components used. (N=54).

Patent document Components	*A %*	*PCT1 %*	*PCT2 %*	*A+ITS %*	*AS %*	*C %*	*Total %*
Image	12	19	17	21	27	25	76
Paragraph[66]	18	19	21	18	18	25	79
Abstract	16	17	19				57
Section[67]	7		21	15		21	52
Reference	11		17		18	21	46
Terms	16	15					41
Classification code	20	17		15	27		59
	100	100	100	100	100	100	
	431	473	478	367	220	400	
N tasks=	13	11	9	9	5	7	
Av. # components and n tasks	4,3	4,6	4,4	4,10	2,2	4,0	

[66] A paragraph is a piece of text within the PA. It can be a sentence or a couple of sentences.
[67] In contrast to a paragraph, a section is defined as a whole separate piece of text with a heading.

Appendix H: Cross-tabulation analysis (see Chapter 7)

The tables contain χ^2 and Spearman rho calculations.

Table 7.1: Domain knowledge (dk) and task knowledge (tk)

Perceived task difficulty	HDK		LDK
High tk	7		6
Low tk	24		4
N=41	31		10
Significance	χ^2=3.31; p<=0.10; df=1		

Table 7.2: Relevance judgment strategy and Source type (6b)

Source type (P=paper, E=electronic, H=human)	A	%	SA	%
Single Source type	18	37	4	8
Combinations of sources (PHE, PH, PE, HE)	7	16	19	39
N=48	25	53	23	47
Significance	χ^2=12.27; p< = 0.10: df=1			

Table 7.3: Relevance judgment strategy vs. expressed information need

5b expressed information need	A	%	SA	%
Single (s)	10	24	0	0
Multiple (m)	9	22	22	54
N=41	19	46	22	54
Significance	χ^2=12.59; p< = 0.01; df =1			

Table 7.4: Relevance judgment strategy and Perceived information need

Perceived information need ((un)structured/(un)clear) 5a	A	%	SA	%
Structured/clear	21	47	6	14
Structured/unclear	3	7	14	32
N=44	24	54	20	46
Significance	χ^2=12.88; p< = 0.01;df=1			

Table 7.5: Relevance judgment strategy and hours

# of hours	A	%	SA	%
1-4	12	30	4	10
5-60	9	22,5	15	37,5
N=40	21	52,5	19	47,5
Significance	χ^2=4.01; p< = 0.05; df=1			

Table 7.6: Relevance judgment strategy and requirements for information need formulation

requirements for information need formulation	A	%	SA	%
1-4	20	36	11	20
5-6	4	8	20	36
N=55	24	43	33	57
Significance	χ^2=10.72; p> = 0.01; df=1			

Table 7.7: Relevance judgment strategy and Number of sources (6a)

# of sources 6a	A	%	SA	%
2-5	7	21	7	21
6-10	3	9	16	49
N=33	10	30	23	70
Significance	χ^2=23.85; p< = 0.10; df=2			

Table 7.8: Relevance judgment strategy and Domain knowledge (topic)

Domain Knowledge	A	%	SA	%
HDK	58	36	37	23
LDK	9	5	41	25
LDK/ HDK	8	5	10	6
N=163	75	46	88	54
Significance	χ^2=24.46; p< = 0.10; df=2			

Table 7.9: Relevance judgment strategy and Content type

Type of content 6c	A	%	SA	%
Classification codes	11	26	1	2
Abstract	6	14	6	14
Bibliographic	3	7	8	19
Term	1	2	6	14
N=42 (tasks)	21	50	21	50
Significance	χ^2=14.18; p< = 0.10; df=3			

Table 7.10: Change of information need and source type consulted

6b (Source types consulted)	Information need change	%	No information need change	%
Single source types (paper, electronic, human)	6	17	13	36
Combined source types (E/H/P)	0	0	17	47
N=36 (tasks)				
Significance	χ^2=4.37; p< = 0.05 df=1			

Table 7.11: Change of information need and average number of query terms

8c Number of query terms	Information need change	%	No information need change	%
1-3	11	33	0	0
4-18	11	33	11	33
N=33(tasks)				
Significance	χ^2=6.153; p< = 0.025; df=1			

Table 7.12: Change of information need and number of unique codes

8i Number of unique codes	Information need change	%	No information need change	%
1-3	18	51	4	12
4-9	6	17	7	20
N=35 (tasks)				
Significance	χ^2=3.31; p< = 0.10, df=1			

Table 7.13: Multiple/single terms as expressed information need and number of sources.

6a Number of sources	S	%	M	%
1-2	8	18	4	9
3-10	3	7	29	66
N=44				
Significance	χ^2=12.37; p< = 0.001; df=1			

Table 7.14: Multiple/single terms as expressed information need and source type consulted

6b (source type consulted)	S	%	M	%
PHE	1	3	15	49
P/H/E	9	29	6	19
N=31				
Significance	χ^2=7.92; p< = 0.005; df=1			

Table 7.15: Task topic knowledge and number of sources

6a (# of sources)	In	%	Out	%
1-3	20	43	1	2
4-8	15	32	11	23
N=47				
Significance	χ^2=6.75; p< = 0.01; df=1			

Table 7.16: Task topic knowledge and terms per session and query string

8e (average terms per session and query string)	In	%	Out	%
0-200	4	15	8	30
200-	13	48	2	7
N=27				
Significance	χ^2=6.01; p< = 0.025975; df=1			

Table 7.17: The perceived task difficulty and clear or muddled information need.

2f (clear/muddled information need)	E	%	D	%
C	29	60	7	15

M	5	10	7	15
N=48				
Significance	χ^2=4.840; p< = 0.05; df=1			

Table 7.18: Clear/muddled information need and source type consulted

6b (source type consulted)	C	%	M	%
E/H/P	20	50	2	5
HEP	9	22,5	9	22,5
N=40				
Significance	χ^2=6.38; p< = 0.025; df=1			

Table 7.19: The number of sources and type of patent applicant

1b (patent applicant)	1-3	%	4-7	%
Company	1	2	12	26
Bureau	17	37	16	35
N=46				
Significance	χ^2=5.79; p< = 0.025; df=1			

Table 7.20: The number of sources and hours per task

3b (Hours per task)	1-3	%	4-7	%
1-5	15	28	3	6
6-25	12	23	23	43
N=53				
Significance	χ^2=9.56; p< = 0.025; df=1			

Table 7.21: The task structuring and expressed information need

5b (expressed information need)	C	%	CS	%
Single	0	0	8	30
Multiple	9	35	9	35
N=26				
Significance	χ^2=4.11; p< = 0.05; df=1			

Table 7.22: The task structuring and source type consulted

6b (source type consulted)	C	%	CS	%
E/P/H	2	8	17	65
PHE	4	15	3	12
N=26				
Significance	χ^2=3.91; p< = 0.05; df=1			

Table 7.23: The task structuring and numbers of relevance judgments made in TPP

9b # of RJ made in TPP	C	%	CS	%
1-3	2	7	18	62
4 10	6	21	3	10
N=29				
Significance	χ^2=7.34; p< = 0.01; df=1			

Table 7.24: The task structuring and patent applicant

1b (patent applicant)	Clear	%	CS	%
Company	7	22	2	6
Bureau	1	3	22	69
N=32				
Significance	χ^2=14.89; p< = 0.001; df=1			

Table 7.25: The number of sources and hours per task

3b (hours per task)	C	%	CS	%
1-9	2	6	22	63
10-25	7	20	4	11
N=35				
Significance	χ^2=9.36; p> = 0.995; df=1; (99.5%)			

Table 7.26: The average number of elements judged for relevance and information source used for final product

10a (source used for final product)	1-2	%	3-8	%
1	7	27	2	8
2	2	8	15	58
N=26				
Significance	χ^2=8.60; p< = 0.005; df=1;(99.5%)			

Table 7.27: The average number of elements judged for relevance and # RJ made in TPP

9b (# RJ made in TPP)	PHE	%	P/H/E	%

1	3	8	15	42
2-10	13	36	5	14
N=36				
Significance	χ^2=9.11; p<=0.005; df=1; (99.5%).			

Spearman

Table 7.28: Information need change correlations

Average number of search terms/strings	.950
Number of query terms	.950
Number of combinations of query types	.950

Table 7.29: Number of sources (6a) correlations

Number of search sessions per task	1.00
Average number of query types combined	1.00
Average number of RJ in TPP	.943
Application of RJ (sequential/aggregated)	.943

Table 7.30: Average number of search items/strings

Number of expressed terms	1.00
Clear/Muddled information need	

Table 7.31: Number of search sessions per task

Average number of query types combination	1.00

Table 7.32: Number of different elements judged in single task

Average number of query types combination	1.00

Table 7.33: Domain Knowledge (within/outside) of task topic *(2cnew)*

Type of query terms	-1.00

Table 7.34: Perceived task difficulty (Task knowledge)

Application of RJ (sequential-aggregated)	.929

Table 7.35: Number of expressed search terms to use (5e)

Information need change during TPP	.950

Table 7.36: Information sources used for finalizing the product

Number of expressed terms	1.00
Average number of search items/strings	1.00
Number of query term	1.00

Table 7.37: Type of PA information component used in final product

Number of search strings used	.975