Usability and usefulness evaluation: an experiment on the DTU Digital Library

By

Bararunyeretse Roger

Supervisor: Haakon Lund

Number of words: 19,087

The Royal School of Library and Information Science

Copenhagen
ABSTRACT

The role of IIR systems, i.e., DLs. is to provide services that support users to perform work tasks that involve interaction behavior. Components of such services include users, the presented content as well as the system on which they are performed. The purpose of this study was to evaluate the usability and usefulness of an IIR system by investigating the role played by domain knowledge in users’ interactions with the system and the system content. Two classes of subjects were recruited from two different areas of expertise in order to participate in an experimental session on the system. Subjects’ information searching behaviors were observed and both objective and subjective measures were utilized. Objective measures were concerned with the number of search steps/tactics, the number of search terms issued by subjects while completing nine predefined tasks, the amount of tasks completion time, as well as the task completeness; whereas subjective measures were based on questionnaires that included some rating statements which elicited the perceived subjects’ level of satisfaction. Similarities/differences between the two classes were also observed and analyzed through documents relevance judgments, criteria for relevancy as well as system documents presentation order. Results of the study revealed that domain knowledge and IR knowledge complement each other in order to perform search tasks. In addition, variables such as satisfaction consisting of system navigation, system ease of use as well as the accuracy of system content are important criteria for usability and usefulness evaluations of an IIR system.

Keywords: Digital libraries; evaluation; human - computer interaction; information behavior; information retrieval; interactive information retrieval; information seeking; information systems; usability; usefulness.

Paper type: Thesis
ACKNOWLEDGEMENTS

First I wish to thank the supervisor of this study, Haakon Lund who tirelessly helped me in the study design, data collection and data analysis as well as all the study processes. Second, I thank my family for their great support and encouragement they have shown me, especially my wife for her comments and corrections.

Also my great gratitude goes to the 12 subjects who volunteered their precious time and participated in the experiment. Without their participation, the experiment would not have been a reality.

Many thanks to DTU DL staffs, i.e., Kasper, Bøgh; Søren, Hovmand; Thomas, Skov Jensen and Tove, Grethe Nielsen who helped me in the process of the experiment and provided me with the accessibility to the DTU DL and the system HCI facilities.
4. DATA ANALYSIS

3. 2. d. Post-session interview

3. 6. Measures

4. DATA ANALYSIS

4. 1. Subjects’ profile

4. 2. Search tasks and subjects’ interactions with the system

4. 3. Interaction measures

4. 3. a. The use of system search functionalities

4. 3. b. Subjects’ number of search tactics

4. 3. c. Subjects’ number of search terms

4. 3. d. Summary

4. 4. Subjects’ evaluation data

4. 4. 1. Evaluation of search results

4. 4. 2. Relevance judgments criteria

4. 4. 3. Effects of documents presentation order on system evaluation

4. 4. 4. Summary

4. 5. Usability and usefulness of evaluation

4. 5. 1. Performance data: time

4. 5. 2. Subjects’ level of satisfaction

4. 5. 2. a. System navigation

4. 5. 2. b. System ease of use

4. 6. Errors observed under the experiment on the system

4. 7. Subjects’ perceptions on search tasks

4. 8. The usefulness of system content

5. DISCUSSIONS AND CONCLUSION

5. 1. Evaluation criteria

5. 2. The design of system usability and usefulness

5. 3. Conclusion

5. 4. Limitations of the study

6. REFERENCES

7. APPENDIX
1. INTRODUCTION

Information retrieval (IR) is concerned with the process involved in the representation, storage, searching and finding of information which is relevant to a requirement for information desired by a human user (Ingwersen 1992, p. 49). IR is regarded as the core research field in information science, hence its objective to study as well as understand IR process in order to be able to design, build and test information retrieval systems by facilitating the effective communication of desired information that occurs between human generator and human user. Generally, human users approach IR systems to find useful information to support their activities at hand. Research in IR has been focusing on the system – based approach. Although systems in this traditional IR
approach were designed for human users, these users were not directly involved in system
evaluations. Studies on traditional IR emphasized on developing and evaluating retrieval algorithms
and indexing techniques. These studies were mostly conducted in laboratories and can be traced
back to the Cranfield studies, i.e. “TREC”. Therefore, claims that real end – users were not involved
in these studies were real.

The IR interaction is regarded as the interactive communication processes occurring during the
retrieval of information by involving the following three components: the user, the intermediary,
and the IR system.

In IIR, human users are typically studied along with their interactions with systems and information.
Research in IIR focuses most exclusively on human, their information needs and information
behaviours. In this context, the focus is on how users conduct searches while they interact with IR
systems as well as how they assess the relevance of retrieved documents. An increasing numbers of
IR systems are being developed for end-users. Callan et al. (2007, p. 29) note that the involvement
of human users into IR system evaluation as well as the study of users’ information search
behaviours and interactions have been identified as central concerns for IR researchers. Research in
IIR studies typical users along with the interactions with the systems and information. While classic
IR studies omitted humans in their evaluation model, IIR emphasizes on users’ behaviours and
experiences, including physical, cognitive and affective as well as the interactions that occur
between users and systems, and users and information. Since IRR studies include both system
evaluations as well as focussed studies of users’ information search behaviours and their
interactions with the systems and information, IIR is connected to different fields, such as IR; HCI
among others.
Evaluations of IIR systems, hereof digital libraries have drawn considerable attention during the last years. DLs are designed to be used by human users. The concepts of usefulness and usability have gained significant attention in user – centred evaluations; hereof the evaluations of DLs. Since these two concepts have their origins in the areas of information behaviour and HCI, they are important in pursuit of user satisfaction and system usage, as they study users’ interactions with representations of information systems and information objects. Järvelin and Ingwersen (2004) formation and analysis of users’ cognitive process as well as information seeking behaviour has been grounded on the interactivity occurring between the human user and both the information system and the information items, located behind the interface. In other words, components in interaction process, such as system, user and content (Tsakonas et al. 2004, p. 46) are typical in the evaluation of DLs. Fuhr et al (2002) have acknowledged these three components as key factors in the DL lifecycle for predicting the usage of a DL. Each component is in interaction with the other two; and the interactions between the three components are defined as usability, performance and usefulness of the system and each of them provides a number of attributes, which are considered as evaluation variables.

In the present study, the usability and usefulness was conducted through an experiment on the DTU DL. This IIR system covers 30,000 titles of academic journals and contains 125 millions full – text scientific papers without duplicates; 70,000 eBooks as well as 600,000 printed books.

This study investigated the role played by domain knowledge on the system usability and usefulness evaluations through its interaction cycle as well as the appropriate methodology and criteria for evaluation. These methods collected data on subjects’ interaction behaviours with the system, i.e., how they completed assigned tasks, in what time and whether or not they were satisfied with the system’s performance as well as the system’s content.
1. 1. Delimitations of the study

The study is organized into five sections as follows: Section 1: Introduces the study and provides a brief introduction to research questions, as well as the objective of the study. It is followed by Section 2 which is concerned with the theory utilized in this study, i.e., system acceptability. This is followed by section 3 which is an introduction to the methodology utilized in this study. Section 4 reports results for the study and discusses some usability problems with the DTU DL as well as a discussion on issues related to IIR usability and usefulness evaluation. Section 5 concludes by discussing the findings of the study as well as some limitations observed. The following subsection introduces the research questions for this study.

1. 2. Research questions

IIR evaluation is an essential component for the design of effective digital libraries which are designed to be used by users. Therefore they must be judged and accepted by the same users. This study conducted an experiment on a selected IIR system in order to assess system acceptance through two classes of subjects with different domain knowledge. It aimed to investigate the role that domain knowledge plays on usefulness and usability evaluation of an IIR system. To this end, the study explored the following questions:

1. To what extent does domain knowledge of either of the two groups influence their performance on an IIR system?
2. Are there any similarities or differences on the performance of these two groups in terms of their interactions with and the evaluation of an IIR system?
3. Does the system under evaluation support users with different domain knowledge?
1.3. Objective of the study

Basically, the objective of the study was to evaluate the usefulness and usability of an IIR system and try to answer the research questions by uncover how different domain knowledge can affect the interactions with and evaluation of the system. The hypothesis was that there might be some similarities/differences between experts in a technical domain and experts in information searching in their search behaviors which could affect the performance of an IIR system, hence the effects on the judgment of the whole system. Research question number one was concerned with subjects’ information behaviors by exploring the effects of domain knowledge on the evaluation of an IIR system. Research question number two compared two classes of subjects with different domain knowledge in terms of their interactions with and the evaluation of the IIR system. And finally, research question number three explored the system capabilities for supporting subjects with different domain knowledge. This was done based on the usefulness of the system content.

By answering these research questions, we are able to understand more about relationships between users’ use of the system as well as the relationships between users and the system content. To conduct the experiment, the study was built upon the theory of system acceptability in the fields of HCI and IIR as discussed in the following section.

2. THEORY

2.1. HCI and IIR in user – centred evaluations

There are many issues involved in the evaluation of IIR systems. A number of research studies
have been conducted in HCI and IIR which aimed to develop and evaluate new retrieval systems i.e., Yee et al (2003). However, most of these ongoing projects have only been connected to the laboratory studies of individual system components with a small number of users and tasks. A recent development has seen numbers of information systems being designed for end-users in order to involve them in the evaluation process. This is the case of IIR systems, where users are studied along with their interactions with system and information (Kelly, 2009, p. 2).

There have been ongoing discussions and the question that arises is on whether IIR is a distinct research area or is just a subfield of HCI (Beaulieu, 2000, p. 436). However, recent study by Ruthven (2008, p. 47) has shown that IIR is a distinct research field. IIR is more than simply developing interfaces for searching (Shneiderman, Byrd, & Croft, 1998) and the force of a good search in IIR comes from a technical knowledge of interactive systems development as well as from knowledge of people’s search behaviour and search context, including the environmental factors that may influence behaviour (Fidel & Pejtersen, 2004). Although HCI and IIR are different research areas, both are important in the studies of user – centered evaluations. IIR systems, i.e., digital libraries are designed to be usable for users. Therefore, if IIR system has low usability, then it will be less effective for users. And research has shown that the two fields can learn from each other and the best research in IIR often reflects best practice in HCI as well as IR (Ruthven, 2008, p. 47).

2. 2. Human – computer interaction (HCI)

The research area of HCI arose in the early 1980s as a distinct discipline and has since been recognised as being inherently interdisciplinary and as a requiring consideration of both psychological and computing issues as well as issues in many other areas (Diaper, 2002). It
emerged as multi-tasking minicomputers and personal computer systems became widespread (Grudin, 1992, p. 215). These systems addressed the tasks of individual subjects. There was an interest in personal information processing which resulted in an academic focus divided between computer science and experimental (perceptual and cognitive) psychology. According to Diaper (2004, p. 20), HCI is a specialised sub discipline of ergonomics for two reasons: for the first, it restricts itself to the study of systems that have a computer component, and for the second, it has emphasised on human psychology than in traditional ergonomics, which started with emphasise on matching human and the device physical capabilities.

This shift of emphasis was related to the view that computers are tools that enhance human mental abilities in contrast to earlier tools that augmented human physical capabilities. Indeed, HCI is all about the design of computer systems with the aim to support people so that they can carry out their activities as productive as possible and in a safe way. Diaper (cited in Diaper et al. 2004, p. 21) emphasizes that HCI has a broad and narrow definition: the broad view being that HCI is concerned with everything to do with people and computer, whereas the narrow view is concerned with usability, learnability, intuitiveness, and so on, with a focus on the user – computer interface. Designing a user – computer interface implies the study, planning and design of interaction between users and computers and the whole body of knowledge. Users in this context are those who directly interact with the computer. Therefore this field is often regarded as the intersection of computer science, behavioural sciences design and several others.

2.3. The concept of usability

This section discussed different aspects of the concept of usability and how it is related to the field of HCI and IIR. Hornbæk (2006, p. 79) has described usability as a core term in HCI. It is a
science that focuses its basic research in recognition and perception and the design of usable technology for IR systems. Gillan and Bias (2001, p. 352) have stressed that the relationship between basic science in cognition and perception and the design of usable interfaces has been a topic of some interest in HCI. The concept of usability is part of the field of HCI as it deals with the design of technology applied to develop a usable interface in order to permit interaction between the users and the computer as well as between the users and information. According to Dillon (2001) the classic HCI approach to evaluation has focussed on usability. Hence the term is often applied to describe a quality of information technology taken in designing IIR systems.

2.4. Usability and system acceptability

One of the challenges that face the designers of HCI systems is to produce a final system that responds to the expectations of its end – users. In other words, it has to respond to its users’ needs in order to be acceptable. The production of a suitable system for the end-users requires the knowledge of the users’ needs and the environments they work in. When this requirement is in place, then the best system can be produced for a particular task and a particular user working in a particular environment. There has been lack of consensus on usability definition as claimed by Shackel (1990, p. 31). However, a general framework for usability embraces the four principal components of any work situation: user, task, system and environment. Furthermore, designing a good usability product will depend on successful harmony in the dynamic interplay of the four components. Therefore, “usability” can be defined in terms of interaction between user, task and system in the environment.

Although the aim of those involved in the development of computer systems has been that of developing useful and effective systems, there have always been problems on formulating that
intention. Usability engineering is devoted to solve usability problems by producing systems that respond to users’ expectations. Therefore, usability engineering is not exclusive to HCI. It is used in interactive information retrieval systems, i.e. digital libraries in order to improve the way in which their information technology (IT) services operate.

Some of the techniques of HCI practitioners have been around in the field of ergonomics for some time. And the history of usability can be traced to Brian Shackel who raised the concept of ergonomics for the computer in 1959 (Faulkner, 2000, p. 6). The first attempt of a usability definition in terms of measures for “ease of use” was made by RB Miller in his paper of 1971 as cited in Shackel (1990, p. 31). Shackel’s (1986) formal definition of usability was based on effectiveness, learnerbility, flexibility and attitude. Shackel’s operational definition of usability was that it allows a system to be evaluated throughout its development lifecycle.

However, Goodwin (1987, p. 230) acknowledges that the concept of usability is less easily defined and adds that it was affected by the types of tasks to be performed.

Usability has been defined by The International Organization of Standards (ISO, 1998, p. 2) “as the extent “to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction. In the ISO definition, effectiveness, efficiency and satisfaction are identified as key dimensions of usability. In addition, the definition provides a clear indication of the importance of carefully defining user and task models since it emphasizes on the articulation of specified users and specified tasks.

ISO’s “effectiveness” refers to the idea that a user is able to carry out intended tasks. However, this definition lacks the concept of time and ease of use. What is needed in order to consider the effectiveness of a system is that a task has to be fulfilled. This is in contrast to Shackel’s concept of effectiveness which is the measurement of time as well as performance.
ISO’s definition of “efficiency” emphasises on the time of performance. This concept of efficiency was not provided as a separate one in Shackel’s definition. However, it was part of his concept of effectiveness.

The third ISO’s definition refers to user’s satisfaction with the system. This implies the degree to which a system is acceptable by the user, how comfortable the user feels when interacting with the system. This could be referred to Shackel’s concept of attitude in his human terms because of him being associated with the field of ergonomics.

Nielsen (1993, p. 24) focussed mainly on the attributes which constitute usability. He argues that usability is a narrow concern compared to the larger issue of system acceptability. The emphasis is that a system is acceptable if and only if it is good enough to satisfy all the needs and requirements of the users. And system performance is a determinant factor for user acceptance of a system (Tsakonas & Papatheodorou, 2005, p. 403). The overall acceptability of a computer system is a combination of its social acceptability and its practical acceptability. The system acceptability model after Nielsen (1993, p. 25) is depicted bellow in figure 1.
The analysis of practical acceptability of a system is possible through categories such as cost or price, support, reliability, compatibility as well as the category of usefulness. The usefulness category is about how good a system is to achieve some desired goal and is divided into two subcategories, such as utility and usability (Grudin, 1992), where utility is the issue of whether the functionality of the system is proper, and usability is the question of how well users can apply that functionality.

Nielsen (1993, p. 26) defines usability as a quality attribute assessing how easy user interfaces are to use and outlines five usability attributes, i.e., learnability, efficiency, memorability, errors, and satisfaction.

While Nielsen’s definition (1993) focuses on the attributes of usability, other definitions, i.e. Bevan et al. (1991, p. 1) focussed on the views to how usability should be measured: (1) The product – oriented view, that is, how usability can be measured in terms of the ergonomic attributes of the object, (2) The user – oriented view, that usability can be measured in terms of the effort and
attitude of the user, and (3) The user performance view, that usability can be measured by examining how the user interacts with the product, with a particular emphasis on either ease of use or acceptability.

Furthermore, these views are complemented by the contextually-oriented view in that usability of a product is a function of the particular user or class of users under the study, the task they perform, and the environment in which they work.

Dumas and Redish (1993, p. 4) argue that usability means that people who use the product can do so quickly and easily to accomplish their own tasks and outlined four points, i.e., users, productivity, tasks, and the ease of use.

Dillon (2001, p. 58) supports the definition of the ISO Standards with an emphasis on specified users achieving specified goals in particular environments. In other words, the definition emphasises effectiveness, efficiency and satisfaction as measurable criteria of performance that are context–bound by the type of user, the type of task and the situation of use.

Therefore, usability evaluation is conducted by having representative end-users who interact with a task-oriented design, while the experimenter records data, such as time for performing tasks, errors, and user attitude (Dillon, 2001, p. 58). However, some problems remain unresolved, such as tasks that lack correct answer where the scoring of effectiveness can be problematic. Though speed measures, such as time taken to complete tasks, number of steps taken, and number of deviations from ideal path are considered as a primary quality of interaction, the question remains on who is to determine the efficiency and whether the time is equally weighted by the user, the designer or the owner. Dillon (2001, p. 60) extended the classic ISO approach by proposing alternatives such as process, outcome and affect as three levels of user experience.

Although there has not been consensus on the definition of usability, Rubin (1994, p. 18) states that any operational definition including one or more of the four factors: usefulness, effectiveness
(or ease of use), and attitude (or likability) as outlined by Booth [(1989) cited in Rubin, 1994, p. 18] is generally accepted by professionals in the usability community.

Lecerof et al. (1998, p. 864) have proposed a definition of usability as one that focuses on how the system is relevant to users' needs, that is, how well the system serves the users' needs; efficiency, that is how efficiently users can carry out their tasks using the system, the users’ attitude or their subjective feelings towards the system, learnability, and the system’s safety, i.e. such as the right to users to undo some unclear actions. Table 1 depicts an overview of different usability definitions and attributes as well as their sources.

<table>
<thead>
<tr>
<th>Source</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shackel (1986)</td>
<td>Ease of use, effectiveness, learnability, flexibility, user attitude</td>
</tr>
<tr>
<td>Reed (1986)</td>
<td>Ease of learn, ease of use</td>
</tr>
<tr>
<td>Bevan et al. (1991)</td>
<td>Usability measurement based on product, user, ease of use, acceptability of product</td>
</tr>
<tr>
<td>Dumans et al. (1993)</td>
<td>Users, productivity, tasks, ease of use</td>
</tr>
<tr>
<td>Nielsen (1993)</td>
<td>Learnability, efficiency, memorability, few errors, satisfaction</td>
</tr>
<tr>
<td>Lecerof et al. (1998)</td>
<td>Users' needs, efficiency, users' subjective feelings, learnability, system's safety</td>
</tr>
<tr>
<td>Dilon (2001)</td>
<td>Effectiveness, efficiency, satisfaction, process (as actions and processes involved in users' interaction with the system), outcome (variables measuring what users attain from the system), and affect (attitudinal, emotional, and mood - related elements of experience).</td>
</tr>
</tbody>
</table>

Table 1: Usability attributes in different standards, models and definitions
2.5. The concept of usefulness

Besides usability, usefulness was considered in this study as a bias for IIR systems evaluation, hereof DLs. Generally, the concept of usefulness is about whether DLs constitute valuable tools for the completion of users’ tasks by answering the questions if DLs support users’ information needs and work tasks completion. According to Tsakonas et al. (2007, p. 1237-38), users assess the applicability of the documents to their work tasks by the relevance of the source, the reliability, the level of the information, the format of the document as well as the coverage of the offered documents. A study by Hong et al. (2002) demonstrates that features of an IIR system have impact on both users’ perceptions of ease of use and usefulness. However relevance, which is considered as a system feature, was related only to perceived usefulness. This relation was interpreted as the association of relevance to the content of the DLs. In contrast to relevance, a judgment of usefulness can be made of a result or a process, rather than only to the content of an information object. Subsequently, it can be utilised to a specific result, to interaction sub sequences as well as to the session as a whole. For this study, relevance was employed as a usefulness criterion for interaction with the system content as the immediate goal was to gather topical documents. Usefulness has been studied differently for system acceptability. Liu (2004) justifies the important role played by the reliability and credibility of information for the selection of the appropriate resources, whereas in their relevance criteria, Vakkari and Hakala (2005) include the level of the offered information. Moreover, users’ information searching behaviours have demonstrated that despite retrieval of full text resources is significant, other levels of information, such as abstracts, are also preferred (Wolfram & Xie, 2002), possibly due to the content overview they offer to users (Krottmaier, 2002). Today’s evaluation practices include recording and analysis of usage statistics based on users’ format preferences, such as .PDF as well as .HTML file formats (Mercer, 2000).
2. 6. Usability evaluation methods (UEMs)

Usability evaluation is part of the HCI as a cognitive activity which requires that users (also called evaluators or participants) exercise their subjective assessments. Usability evaluation dates back to the beginning of HCI and usability evaluation methods (UEMs) go back more than 4 decades and published accounts of UEMs reaches back for more than 2 decades (Nielsen & Molich, 1990). However, since different methods and measures have been applied to approach usability evaluation, researchers have conducted studies that compare UEMs (Nielsen & Molich, 1990) for some time. They have substantive differences and include pluralistic and usability walkthrough, heuristic evaluations, cognitive walkthroughs, think aloud evaluations, and scenario – based and guidelines – based reviews (Bias, 1991; Duservire et al. 1991; Jeffries et al. 1991; Karat et al. 1992; Nielsen & Molich, 1990). These methods differ from each other depending on the source used for the evaluation – a source that can be users, usability experts, or usability models. But which usability evaluation technique is most appropriate in evaluating users’ interaction with a DL system? To answer this question, Blandford et al. (2004, p. 35) analyzed four usability evaluation techniques, but did not find any one technique to be better than another. In relation to this study, the following subsection presents an overview of two UEMs, such as expert – based and user – centred evaluations.

2. 6. a. Expert – based evaluations

Expert – based evaluations are inspection methods evolved in HCI with the aim for assessing a system without the involvement of end - users as evaluators, in order to avoid problems related to time and cost. Nielsen is probably the best known proponent for this type of method. Nielsen &
Mack (1994, p. 5-6) have outlined a series of eight inspection methods through which usability inspection is conducted. Once the software product has been produced, then the usability inspection is conducted with the aim to fixing the problems and improving the usability of the design. Nielsen admits that experts are the best evaluators if they know the domain of the interface. However, he adds that in the absence of such evaluators, representative end - users can be included in the evaluation.


Evaluation of interactive information retrieval systems (i.e., DLs) has been performed so far either in a system – based or a user – centred fashion. In the former, the evaluation emphasizes on measuring a set of properties of the system and very little focus is on the person who is the actual driver of the system. In the latter, the user is a central actor and his/her behaviour determines the outcome of the evaluation.

The user – centred approach was initiated in the 1970s as an alternative to the traditional system oriented approach. At this time it appeared necessary to extend the focus of the research concentrating on the individual actors of the information search and use processes, in the social, practical and cultural contexts. This approach focuses mainly on the user’s problems and his/her production of meaning, stressing that the recovery of information is efficient depending on the integration of the results with the user’s life, and more specially on how the user evaluate the utility of the information provided to solve his/her problems. In contrast to the user – centred approach, the system- oriented approach is concerned with information environment external to the individual user that said, the user being not involved in the process. However, the user - centred approach
examines the individual psychological and cognitive necessities and preferences and the way these aspects influence the standards of search and use of information.

Users have been the main focus in the field of HCI research. The aim of user – centred evaluations is to identify representative users, representative tasks, and develop a procedure for capturing the problems that users encounter when trying to apply a software product in accomplishing these tasks. Generally, two types of evaluations are carried out, i.e., as formative and summative evaluations (Hartson et. al. 2001, 374-375).

Formative evaluations on one hand are those used to collect information for the design as they focus on problems that need to be fixed during the prototype design stage before a final design can be accepted for release. In the context of usability, they are used to find usability problems to fix so that an interaction design can be improved. This type of evaluation collects qualitative information data about the system, i.e. opinions concerning users’ feelings about the system; the types of problems they experience with the system; changes they suggest can be made, and so forth.

Summative on the other hand are evaluations conducted in order to assess the efficacy of the final design by documenting the effectiveness, efficiency and user satisfaction of a product at the end of the development cycle. These evaluation methods require quantitative data which are useful at the end of the design. Hewett (1986, p. 198) describes the purpose of summative evaluations as being about:

“...assessing the impact, usability and effectiveness of the system – the overall performance of user and system”
These two types of evaluation have different goals as they are conducted differently. This study conducted a summative evaluation since the system under evaluation has already been implemented. Hartson et al (2001) stress that this method is conducted to assess or compare the level of usability achieved in an interaction design where a formal experimental design which includes a test for collecting quantitative and statistical data (e.g. user performance numbers) is required by involving a number of representative end - users.

With respect to acceptability of the evaluated system, this study employed the user – centred evaluation method. This method which seemed to be valid was expected to yield results that better represent the way the system was intended to be used in the real world. It was also seen as more complete, since the picture that one could get about the system would be richer, potentially taking into account not only system related factors such as effectiveness and efficiency, but also those related to user satisfaction and learnability.

2. 7. Usability evaluation in the context of IIR systems

The conception and implementation of any IIR system, i.e., a digital library requires some prior studies in order to be sure that the tool will be adequate and respond positively to the aspirations of its intended end - users. These studies are based on methods, concepts and theories drawn from some areas of information, such as HCI, e.g. usability studies in particular and information science (IS) especially studies about users’ needs and behaviors during the information search and use. As said before, the field of HCI is concerned with all contacts between computer systems and human use, and more precisely, the interaction occurring in the process.

A system is evaluated to ascertain the level of its performance or its value. Digital libraries can be judged by their effectiveness in retrieving relevant results (how well does a system or any of its
parts perform the roles or tasks for which it was designed?) and efficiency (at what cost?)
(Chowdhury & Chowdhury 2003, p. 272).

According to Tsakonas and Papatheodorou (2006, p. 403) system usability and system usefulness are two related properties of system interaction, which in combination, determine system satisfaction and usage. Usability testing has been a research area in the field of digital libraries and HCI. It focuses on the effects obtained when individuals use an information system in order to accomplish tasks. However, Saracevic (2000, p. 358) claims that there still are some fundamental concepts that need to be clarified, such as What is meant for a digital library? Why to evaluate? What needs to be evaluated? What are the criteria for the evaluation? How to apply these criteria in the evaluation? What is the purpose for the evaluation? These seem to be general questions in any and all evaluations.

Different methods and instruments have been utilized in order to evaluate the effectiveness and usability of digital libraries. An evaluation is a judgment of worth. Saracevic (2000, p. 359) defines evaluation as an appraisal of the performance or functioning of a system, or part thereof, in relation to some objectives; and the performance can be evaluated in relation to the effectiveness, efficiency or the combination of these two instruments.

In comparison to other areas of research in digital libraries, little work is being done in order to understand the purpose and usability of digital libraries (Theng et al. 2000, p. 238). Borgman et al (2000, p. 229) are of the same view and state that “Relatively little has been done on evaluating usability of digital libraries in any context. Marchionini et al (2000, p. 2) have expressed their view in relation to the evaluation of digital libraries:
“Evaluation of a digital library may serve many purposes ranging from understanding basic phenomena (e.g., human information seeking behaviour) to assessing the effectiveness of a specific design to measuring sufficient return on investment. Human-centred evaluation serves many stakeholders ranging from specific users and librarians to various groups to society in general. Additionally, evaluation may target different goals ranging from increased learning and improved research to improved dissemination to bottom the profits. Each of the evaluation goals may also have a set of measures and data collection methods. Finally, the evaluation must have a temporal component that can range from very short terms to generations”.

Accepted definitions of usability focus more on users themselves and include multiple usability attributes (Nielsen, 1993). However, the areas of usability testing in digital libraries have covered efficiency, effectiveness, navigation, functionality, metadata appropriateness and so on. A research review by Jeng (2005a, 2005b) concluded that usability is a multidimensional construct. She further suggested a model for the evaluation of usability in digital libraries where she examined the effectiveness, efficiency, satisfaction, and learnability of digital libraries. The model was tested, and the results revealed that there is an interrelation between effectiveness, efficiency, and satisfaction. At the same time, the results also identified users’ perceptions of ease of use, organization of information, terminology, attractiveness, and mistake recovery.

Attributes of usability, particularly user needs and user satisfactions have been under investigation in many of the digital library studies. Fox et al (1993, p. 503) who were interested in understanding users’ information needs and their perception of existing information systems conducted an interview with potential end – users and experts in the fields of library and computer sciences. A study by Buchanan et al. (2009, p. 648) concludes that effectiveness, efficiency, aesthetic appearance, terminology, navigation, and learnability are key attributes of system usability, while
relevance, reliability, and currency are key attributes of system usefulness. The access to full resources is of big importance for students using a digital library for their assignments. They do not want to spend a lot of time to access the information they need.

The usefulness and usability of a digital library depend on the effectiveness and efficiency of the user’s experience with the system. Therefore it is important for users to be able to accomplish goals for their tasks (Buchanan et al., 2009, p. 638-640).
Usability has been tested on the California digital library (CDL) and results revealed that end - users want a digital resource supporting a search, evaluate, and access cycle. Intended end users want to quickly locate a resource, evaluate the sources’ applicability in accordance to their needs, and then access the actual resource or move on by repeating the cycle (Lee 2009, p. 3-4). Consequently, if users are not able to efficiently work within the cycle, they will then move into an entirely new digital resource.
Another key aspect in usability testing is the user’s ability to navigate the system and learn new tool functions efficiently. The user’s lack of awareness of their location on the system’s website can be a barrier and users’ disorientation can lead to cognitive overload. In addition, the learnability, or instructional design of the web site’s functions can be part of a fundamental aspect of usability because the users’ ability to learn the system is their first experience (Buchanan, 2009, p. 640).
A study by Oulanov and Edmund (2002, p. 484) used a questionnaire as the primary method of usability evaluation with affect, efficiency, control, helpfulness, and adaptability as evaluation criteria. At the University of the Pacific, Krueger et al., (2004, p. 287) used a formal usability evaluation technique in order to assess students’ awareness of information resources offered through the library’s Web pages. 134 students were recruited as participants and were asked to perform eight tasks. In order to understand the purpose of digital libraries, Theng et al (2000, p. 238) applied both questionnaires and usability heuristics to evaluate the usability of the ACM
Digital Library, where 45 undergraduates were asked to complete questionnaires on how satisfied they were with the design and structure of DLs. Usability inspection method was used to evaluate the design and functionality of Networked Computer Science Technical Reference Library (NCSTRL) (Hartson et al, 2004, p. 108). The results revealed that the system’s design was functionally oriented rather than an approach based on user task threads, and the terminology used on the system was a jargon and terms were used as designer – centered rather than user – centered.

The University of Illinois at Chicago (Augustine et al. 2002, p. 355) applied a formal usability test where 12 students were recruited as subjects in order to assess the clarity and ease of navigation and the users’ search for information. Table 2 bellow depicts an overview of a review on the test on usability in Academic Digital Libraries.

<table>
<thead>
<tr>
<th>DL Website</th>
<th>Subjects</th>
<th>Applied methods</th>
<th>Areas of application</th>
<th>Researchers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rutgers University Libraries</td>
<td>3 Students</td>
<td>Pretest &amp; Post-test questionnaires</td>
<td>Performance &amp; Satisfaction</td>
<td>Jeng (2005a, 2005b)</td>
</tr>
<tr>
<td>Websites &amp; The Queens College</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lib. Websites</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CUNY + Web</td>
<td>10 Students</td>
<td>Questionnaire</td>
<td>Interface</td>
<td>Oulanov et al. (2002)</td>
</tr>
<tr>
<td>University of the Pacific</td>
<td>134 Students</td>
<td>Formal usability test</td>
<td>Awareness of Info. resources</td>
<td>Krieger et al. (2004)</td>
</tr>
<tr>
<td>ACM DL</td>
<td>45 Undergraduates</td>
<td>Questionnaire</td>
<td>Design &amp; structure</td>
<td>Theng et al. (2000)</td>
</tr>
<tr>
<td>UIC (Univ. of Illinois at</td>
<td>12 Students</td>
<td>Formal usability test</td>
<td>Users' search for Information</td>
<td>Augustine et al. (2002)</td>
</tr>
<tr>
<td>Chicago)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: An overview of a review on usability tests in Academic Digital libraries
2. 8. Work task

The concept of work task has been discussed by Borlund (2000) as an approach to the evaluation of IIR systems. This approach aims to merge the system – driven and the cognitively user – oriented approaches, the former being based on traditional laboratory experiments and the later referring to operational tests. The traditional system driven approach on one hand has been dealing with the improvement of the match between the query and document as well as the specifications of computationally effective and efficient retrieval algorithms. This approach is related to the experimental settings and investigations, i.e. the TREC experiments (Harman 1997) which are grounded on the Cranfield studies (Cleverdon & Keen, 1966). The traditional model (the laboratory model) is depicted in figure 2. In the figure, IR system is represented by a database, retrieval algorithms, requests and stored relevance assessments. The system components are represented in the middle and the evaluation components are on the top, left and bottom in the light shaded area. The main focus of these studies is to assess how effective a retrieval performance of a given system is. This evaluation which most of the time takes place in a laboratory involves a test collection, a set of queries as well as a set of representations of information needs. Therefore the relevance judgments for this evaluation are considered objective and static between a request and the document as seen by an assessor. Users have been involved in the interactive TRACK of TREC (Over, 1997). The user – oriented approach has three main goals: the first focuses on the identification of relevance (Saracevic, 1996), the identification of relevance criteria and their use, and considers relevance as being a dynamic and context specific phenomenon; the second goal is the cognitive view point where the user is involved in the process and judges the relevance as shown in the model (the standing man in figure 2); the third goal for this approach is considered as
a process – oriented approach which considers information retrieval as a process like problem solving, browsing, trial and errors (Kulthau 1993).

Figure 2: The Laboratory model schematized (Ingwersen & Järvelin, 2005, p. 5).

The conflict among the above approaches has been a conflict between, on one hand, control over experimental variables, observability, and repeatability, and on the other hand, realism (Robert & Hancock-Beauliue, 1992, p. 460). Borlund (2000, p. 76) has proposed an experimental setting with a focus on the issues of experimental control and realism consisting of three components: (1) The involvement of potential end – users as test persons; (2) The application of dynamic and individual
information needs (real and simulated information needs); and (3) The use of multidimensional and dynamic relevance judgments.

This study has focused on student - subjects from DTU as representatives of real end - users of the evaluated system compared to student – subjects in the area of library and information science. To this end, the hypothesis was that in the case real end – users are absent in the experiment, there would then be an absence of human interaction in the process of system evaluation. Furthermore, in the absence of human interaction, then the interpretations of individual and potentially dynamic information need would be absent. And finally, in the absence of the two first components, then the multidimensional and dynamic relevance will not be assessed.

The approach of work task is under the second component – where the concept of simulated work task situation falls (Borlund & Ingwersen 1997). A simulated work task situation is further defined as fulfilling two goals: firstly, it serves as a description of the universe of the information need situation in which the user is supposed to see himself, and based upon which the user formulates the search statement to the system. It consists of an indicative request, a definition, and a simulated work task situation; secondly, the concept serves as a foundation for the user when assessing situational relevance of the retrieved documents (Borlund & Ingwersen, 1997, pp. 227-228). Furthermore, the simulated work task situation helps the user with a description of the source of need, the environment of the situation, the problem to be solved; and helps the user to understand the objective of the search (Borlund & Ingwersen, 1997, p. 229).

Indeed, a simulated work task is a short “cover story” describing an IR requiring situation and providing at the same time an experimental control by being the same for all the participants. This control is useful for one to compare the search results for the group of participants. The simulated work task situation works as a catharsis of the so - called “breakdown situation”, which is a
cognitive state creating an information need that has to be satisfied in order to enable the user to
deal with the situation and moves on.

Simulated work task is applied in the case of DTU DL as a tool that triggers the simulated dynamic
information needs in order to help facilitating the evaluation of the system in a way very close to the
process presented in figure 2 where the assessment of effectiveness and relevance are to be
conducted by and compared between real end – users of the system (technical domain experts) and
IR experts.

In HCI research, there is an issue of realism that deals with the involvement of potential end – users
as subjects, who develop subjective and individual information needs upon provided work tasks.
Therefore the concept of task is used in the field of HCI (Borlund 2003, p. 9). According to Preece
et al (1994, p. 411), a HCI task analysis is the analysis of:

“A task is a structured set of activities in which actions are undertaken in some sequence. Tasks are
what the human has to do (or thinks he or she has to do) in order to accomplish a goal. At some
point, the human physically interacts with a device by performing an action. Thus, we can define a
task (or internal task) as the activities required, used or believed to be necessary to achieve a goal
using a particular device”.

2. 9. The concept of relevance

More and more end – user public access information systems (i.e., DLs.) have been developed
in the past decades. With this development, changes have occurred in the field of information
retrieval (IR). Today, IR process is taken from the perspective of the end – user as well as that of
the system in use. As it is shown in figure 2: (The laboratory model), the traditional model of the
IR task which was based on a matching function between a user’s query and a set of documents stored in the system has led to the concept of extending the system’s boundary to include the end – user as an essential component of the system. Although it has been assumed that the query is a representation of an information need and that information originates from a user’s cognitive state, users have not been directly involved in the experiment for the judgement of the outcome of the system. And despite the fact that binary relevance judgments on document surrogates in a bibliographic database may have been an acceptable compromise for the purposes of performance testing in a laboratory, Beaulieu (2000, p. 432) acknowledges that the complexity of relevance judgments has become even more evident in the context of full-text retrieval. Moreover, relative judgments based on inter-document assessments may also lead to the reassessment of the initial relevance criteria themselves. Hence the challenge in questioning the validity of relevance and its associated performance measures of recall and precision stems from the fact that information retrieval is inherently an interactive process. This issue of defining relevance in IR evaluation has led to the involvement of end – users in the process (Saracevic, 1975).

The approach of realism has been centered on requests and relevance judgments and as testers became more aware of the relevance revolution, it became clear that an approach of real people with real information needs was in need for the assessment of relevance. The relevance movement emphasized on relevance being judged based on the information need – and in addition, in relation to the underlying situation of the need, by the application of situational relevance. Therefore, the concept of relevance is treated in relation to its dynamic and multidimensional nature, by being assessed interactively in a non–binary way.
Relevance has then been considered as one of the most fundamental aspects of IR (Tombros et al. 2004). It is a multi-dimensional concept, as it relates to content, which can be considered subjective (Thornley & Gibb, 2007). Relevance considers both the users’ (cognitive) knowledge and (subjective) perceptions; is a situational (influenced by the information problem), complex and multidimensional, and although dynamic and constantly changing, is also systematic, observable and measurable (Barry & Schamber, 1998). In the context of system usability and usefulness, relevance is associated with the issue on how well the system enables users to accomplish tasks and how well information retrieved contributed to the user needs.

2.10. Subjects’ domain knowledge and individual differences in evaluations

The HCI literature generally discusses the importance of using “appropriate test subjects” when carrying out a usability and usefulness evaluation. Users’ task and their individual characteristics and differences are two important issues for the evaluation (Nielsen 1993, p. 43). Egan (1988) studied individual differences between subjects’ performance. Nielsen (1989, p. 243) analyzed the usability of hypertext systems and found that 4 out of 10 usability problems were due to individual differences between subjects, while two out of them were caused by task differences. In relation to the level of user expertise, this study considered Nielsen’s three dimensions (1993, p. 43): (1) the user’s knowledge about the domain (ignorant versus knowledgeable); (2) the user’s experience with computers in general (minimal vs. extensive); and (3) the user’s experience with the system being evaluated (novices vs. experts). Indeed, these three main dimensions on which user’s experience differs are shown in figure 3:
Figure 3: The three main dimensions on which users’ experiences differ: knowledge about computers in general, expertise in using the specific system, and understanding of the task domain (After Nielsen, 1993, p. 44).

It is pointed out, that it is important to choose subjects who are representative of the intended target user community with respect to parameters such as their demographic profile, i.e., gender, age, education, etc. (Preece et al. 2002; Rubin, 1994). This is an additional set of factors that differentiate subjects (Cazja et al. 1988, Fowler et al, 1987).

On the evaluation of IIR systems, there are two main factors that may affect searcher’s choice of tactics and terms (Ingwersen 1996) which in turn may affect the relevance assessments of the system output. Domain or subject knowledge is a key issue in regulating searchers’ ability to articulate information needs conceptually and in the way they express search terms. This can also be reflected in subjects’ query reformulation due to their relevance feedback.

Studies have examined the effect of domain knowledge/familiarity on search tactics and results have shown that there are behavioral differences between searchers with different levels of domain knowledge. Some of these studies have demonstrated the effects of domain knowledge on search
tactics. A study by Hsieh-Yee (1993) has found that for users who have an experience in information searching (IR), subject knowledge affected their searching tactics. When they worked with a topic typically unfamiliar to them, they used the thesaurus more for term suggestion, monitored their search more closely, included more synonyms, and tried out more term combinations than when they engaged in a search on a familiar subject area. Moreover, domain knowledge was said to be in association with searchers’ ability to select appropriate search terms (Vakkari et al. 2003), or efficient selection of concepts to include in the search (Wildmuth, B. 2004), or the ability to better utilize the assistance from a thesaurus (Sihvonen & Vakkari, 2004). Search tactics are very important in IIR. Subjects conducting information searches must have some search formulations that are combinations of the choice of search terms, operators, and tactics (Vakkari 2003, p. 440-442). Bates (1979) has defined search formulation tactics as tactics to assist in the formulation and reformulation of searches, while term tactics are tactics that help select and revise terms in search formulation. According to Bates, (1979, p. 207), “A search strategy is a plan for the whole search, while a tactic is a move made to further a search”. A move can be defined as any change in formulating a query. A move in a search forms the basic unit of analysis. It is an identifiable thought or action that is a part of information searching (Bates 1990), for improving search results (Fidel 1991). A query represents the user’s information need, which consists of search terms and of possible operators connecting them.

3. METHODOLOGY

In light to the research questions, this section presented the IIR system under evaluation as well as different methods and instruments employed for data collection for this study.
3. 1. The evaluated system: DTU DL.

For this study, the DTU DL was selected for evaluation. The user interface for the system is designed with a similar appearance to Web search engine such as Google. As an integrated search system, the Google like syntax can cross search in different subscribed literature of DTU library, ranging from journals, scientific articles, and printed books, eBooks – including references from PubMed, the Web of Science and so forth. Two formats are used in the display of system content, namely HTML format for bibliographic information as well as abstracts and PDF format for full – text documents.

Regarding IR, DTU DL offers a more complete access to content through simple and advanced search interfaces. Furthermore, AND Boolean is utilized as a default between search terms. It is quite possible for users to key in the query terms in a search field, which can also contain a set of parameters useful to refine the query. There is a similarity between the syntax and that of a typical Web search engines, i.e., it includes a number of search terms and support a basic set ops operations. On its usability and usefulness evaluation, the system is tested once in a while in order to fixe some areas of problems. DTU DL was an already implemented IIR system and consequently has been tested by Userminds. Therefore, this study conducted a summative evaluation in order to assess the efficacy of the final design by documenting subjects’ level of satisfaction, i.e., by system navigation, system ease of use as well as the usefulness of the system content based on relevance assessments of retrieved documents.

3. 2. Subjects and their identification codes

The process of determining what type of subjects to include in an experiment depends on the
methods intended to be used. In order to participate in the main experiment, subjects were told in advance that the test was concerned with the system under evaluation. In addition, they were provided with a schedule for the experiment in order to make the arrangement in accordance to their availabilities. They were informed on where to show up for the experiment where a short introduction to the system was provided. Ethical issues were also taken into consideration and subjects were informed that their identities would not appear anywhere, thus they were assigned an ID code, i.e., ID 01 as their numerical identity. After a personal contact with them, invitations as an official solicitation were sent electronically (appendix A). However, since the DTU DL had recently conducted a usability test, the experimenter made sure that subjects from DTU should not have participated in the previous test. This aimed to avoid the problem of having subjects who likely knew something about the aim of the experiment and the desired outcome (Kelly, 2009, p. 69).

3.3. Sampling

Two classes of subjects with different domain knowledge were part of the experiment. They were recruited from the Technical University of Denmark (DTU) and The Royal School of Library and Information Science (RSLIS), currently IVA. Two reasons were considered for the recruitment: (1) Subjects from DTU are the targeted audience for the DTU DL and there was a premise of having declarative knowledge, in this case domain knowledge under which the system operates; (2) Subjects from IVA understand IIR systems, i.e., digital libraries and have expertise in the use and evaluation of such systems. Twelve subjects were recruited for this experiment. However, two of them conducted the pilot test while ten (five from each group) were taken into the evaluation process. Each class was represented by three males (60 %) and two females (40 %). They consisted of 5 masters of library and information science graduates (IVA student subjects); 2 DTU graduates,
in Civil Engineering and Planning, Innovation & Management respectively; and 3 DTU undergraduates, one in the faculty of Software and two in that of Technical Biomedicine. IVA students were aged between 25 and 38, while those from DTU were between 21 and 27 year old. Surprisingly, only two of DTU students had knowledge about the DTU DL, which means that the three others had never used the system. One of those with knowledge about the DTU DL has been using the system once or twice a semester for searching articles for his assignments, while the other has used it just two times before the experiment. In contrast to DTU students, IVA students were without any knowledge about the technical domain of knowledge under which the system operates. However, the hypothesis was that they might differ with their counterparts in their information search behaviors because of their area of study. The debriefing interview elicited their perceptions about how easy the system was to learn and how satisfied they were with the system performance and its content after an intensive interaction with it by performing assigned tasks. A time slot of 60 minutes was reserved to perform tasks. However, this was not a limit, since it was expected that some subjects would need more time to assess the performance of the system for judging the relevance of its content. The whole experimental process then lasted between 42 minutes and 95 minutes for each subject. For every search task, subjects ended the search session only when either their information needs were satisfied or the subject just gave up his/her task. The role of the experimenter was to observe subjects while performing their searches. Moreover, as a subject might modify his/her search tactics during the search session, the retrieved documents were defined as those displayed on the screen in the response to the last tactic in a search session.

3. 4. Pilot test

A preliminary test was conducted with the aim to observe the applicability of the proposed
methods, including the time the experiment was to last, tasks, and instruments for data collection. This test was done by two subjects among those recruited for the main test (Campbell, 2001, p. 13) utilizing the same actual test instruments. Some usability problems were revealed, and the test helped to redesign test instruments and search tasks as well as redefine the test procedures. A part from the time estimation and tasks/instruments redefinition, this test allowed more exercise with the system under investigation in the same manner as in the actual experiment. Furthermore, it provided the experimenter with detailed feedback from test subjects concerning the methods, as well as a comfort feeling with the administering of the experiment.

3.5. Test instruments and data collection techniques.

This subsection discussed multiple instruments and techniques employed for collecting data. They consisted of the apparatus and techniques used in evaluating the usefulness and usability of IIR. One IIR system, i.e., DTU DL was selected for subjects to evaluate.

3.5.1. Apparatus

The experiment was conducted on a 17” monitor resolution with 1152 x 864 pixels. Software “Adobe Connect”: [www.connect.forskningsnettet.dk](http://www.connect.forskningsnettet.dk) was employed to register the time lasted for all tasks as well as to record subjects during the performance of tasks. This software has an integrated video and voice recorder. However, the HCI facilities provided by the DTU DL have no integrated webcam. This was only possible using the experimenter’s own laptop with which he managed to take some clips outside the DTU DL facilities (see figure 4).
Subjects were instructed to talk aloud as they completed their tasks and data were recorded. This method was useful for collecting both qualitative and quantitative data in relation to subject’s information search behaviors and their perceptions on system usefulness and usability. The system usefulness was assessed based on the relevance of the system’s content, whereas the system usability was assessed based on the system performance. Subjects were recorded on every sequence of a search strategy and their feedback after every task as well as the time lapsed for performing search tasks. At the end of every test session, the experimenter spent a time for debriefing. At this stage, it was possible to replay the recorded data in order to compare them with notes and discuss
general reactions and opinions from the session. This stage helped ensure that everything was noted and discussed before any new data was collected.

3.5.2. Data collection techniques

Multiple techniques were utilized to collect data, including recordings of talking/thinking - out – aloud, direct observations, post – search interview (Ingwersen & Järvelin, 2005, p. 247) as well as search tasks and evaluation questionnaires.

3.5.2. a. Think/Talk – out - loud protocol

Talk – aloud or think – aloud (TA) is a discount usability evaluation method which was first introduced in systems development in the 1980s by Lewis (C. Lewis, 1982) and is probably the single and most important method in use for practical evaluations (Nielsen, 1993, p. 195). TA has traditionally been used by academic researchers as a quantitative data collection method. However, this method has been accepted in practical evaluation of human computer interfaces (Denning et al. 1990, p. 1285).

This study employed the “talk – out – loud protocol” as described by Rubin (1994) and Nielsen (2000) where subjects were asked to articulate their thoughts and decision – making verbally as they engaged in solving a series of tasks, describing their actions, how they perceived the system and so forth. These verbal thoughts from subjects allowed the observer (experimenter) to determine both what they were doing when interacting with the system as well as why they were doing it. With the Adobe Connect Software, it was feasible to record these data mostly on different steps in their searches as well as their perceptions about the system and system content.
Though this method seemed to be relatively inexpensive as commented by Nielsen (1993), some of the subjects had difficult time to articulate their thoughts while they were engaged in task completion. The experimenter encouraged them to continue thinking – out - loud in order to generate data to help understand how subjects responded to the system. Studies on whether thinking aloud changes the way subjects perform search tasks have been conducted. A study by Rhenius and Deffner (1990) showed that subjects talking out loud take longer to perform tasks. In addition, subjects who talk out loud are less flexible at changing problem solving strategies when they first start talking. Bainbridge (1979) has also outlined two problems that a subject can face when talking - out - loud while performing tasks at the same time. Firstly, verbalization can slow subjects down, thus making any performance measurements less representative of the subject’s regular working speed. Secondly, subject’s problem solving behavior can be influenced by the very fact that he/she is verbalizing his/her thoughts.

3. 5. 2. b. Evaluation questionnaires

The questionnaire is said to be one of the most popular methods employed to collect data from subjects in IIR evaluations (Kelly, 2009, p. 91). This method which consisted of both subjective and objective evaluations was employed in combination with the talking – out – loud method in order to gather thorough feedback from subjects because predefined questions on a questionnaire balance nicely with the feedback from other methods. Questionnaires were administered for two reasons (Dumas & Redish, 1999, p. 208): (1) to ask every subject the same question; and (2) to avoid forgetting to ask the question. Questionnaires consisted of closed questions as well as opened ones. Closed questions were provided with a specific response set with a five point Likert scale, while open questions provided a room for subjects to respond in any way they saw fit, (i.e., is there
something you liked/disliked about the system?). On one hand, closed questions were used to gather quantitative data, thus they were useful for providing numeric representations of subjects’ attitudes and feelings and allowed the experimenter to make statistical comparisons. On the other hand, open questions were used to produce qualitative data. Thus they were useful for gaining more unique and varied insight into subjects’ experiences and for understanding the reasons behind particular attitudes and behaviors. Furthermore, open questions were useful as they allowed the experimenter to better interpret and contextualize subjects’ responses to closed questions. In this study, two types of questionnaires were completed as pen – and – paper and one was followed up as a post – session interview. They are discussed in the following subsections.

3. 5. 2. b. 1. Pretest evaluation questionnaire

A pretest evaluation questionnaire for collecting demographic data (appendix A) was sent electronically to subjects prior to the main experiment. This type of questionnaire aimed to gather information about their background and was useful for the interpretation of the data from the experiment as well as to verify the qualifications of subjects. Data on demographic variables, such as age, gender (Collins & Aguínaga, 2001, p. 26), together with the subjects’ skills in computer were employed to measure their effects on subjects while interacting with IIR.

3. 5. 2. b. 2. Post – search task evaluation questionnaire

This questionnaire (appendix B) was administered to gather data on judgments and ratings following each task. It aimed to simulate subjects for interacting alone with the system while talking –out - loud. It collected both qualitative and quantitative data in form of feedback about the
subjects’ experience employing the system to complete assigned tasks. Thus its particular objective was to assess the system - task interaction as well as the user – system interaction. Subjects were presented with a list of 9 predefined tasks. Each task represented the goal or purpose of the search – what a user wanted to accomplish by searching. Some tasks aimed to obtain an immediate reaction to the subjects’ experience, the system navigation, finding documents for thesis writing, finding known items, etc. Other tasks were related to the subjects’ information needs and information retrieval in order to get the subjects’ activities and relevance judgments. The combination of user and task models helped to define the particular behaviors and activities the system was intended to support. Furthermore, it helped determine the appropriateness of particular evaluation measures and study subjects.

3. 5. 2. c. Observations

Observation data were used to describe the activities, subjects, and the meaning of the observations from the observer’s perspective. Such data were needed as a supplement to verify the information provided in the questionnaires and data recorded by the talk – out - loud method as it allowed the observer to fully understand the situation described. Mullings (1984, p. 1) has recorded that:

“Observation is a way of collecting data in a purposeful and systematic manner about the behaviour of an individual or a group of people at a specific time and place - observation studies events as they actually occur and also what people do rather than what they say they do - observation can be used to study both users and usage.”
With this method, the experimenter was seated near subjects and observed them as they interacted with the system while completing the assigned search tasks. The focus was on particular events and behaviours and notes describing their observations were taken. As Nielsen puts it (1998):

“Watch what people actually do. Do not believe in what people say they do…users self – reported data are typically three steps removed from the truth”.

This real – time observation collected and measured data on the number of search tactics as well as the number of search terms employed by subjects. Furthermore, unusual interactions or instances where subjects seemed to be confused or frustrated were noted. It was also noted that usability could be assessed during the experimental task by observing critical incidents and breakdowns, i.e. difficulties and problems encountered by users which either interrupted normal interaction or caused the user to abandon the current task (Sutcliffe, et al. 2000, p. 749).

3. 5. 2. d. Post - session interview

A post – session interview (appendix C) with open – ended questions was employed as a follow up after tasks completion. This method collected qualitative as well as quantitative data to serve a variety of goals, such as measuring subjects’ level of satisfaction, getting user feedback on their perceptions about the system’s usefulness and usability, as well as seeking user input. This type of data collection provided subjects with more opportunities to speak in their own voice instead of merely responding to the categories of questions that the experimenter had defined for them, as they might with a real questionnaire and was an occasion for the experimenter to interact with subjects. According to Kelly (2009, p. 95), open – ended questions allow one to get more individualized responses as well as some flexibility with respect to probing and follow up.
3. 6. Measures

Measurement is a key issue in the evaluation of IIR systems. And any usability test requires collecting both performance (objective) measures and subjective measures (Dumas & Redish 1999, p. 164). Performance measures were those based on counts of subjects’ actions and behaviors that the experimenter could see, i.e., interaction measures, while subjective measures were those related to subjects’ perceptions, opinions and judgments of the system, i.e., the overall satisfaction with the system. Contextual, interaction, performance and usability have emerged as the standards measures of IIR systems (Kelly 2009, p. 100). Contextual and interaction measures sought to characterize subjects, while performance and usability measures were more related to the information seeking situation, such as task – type and subjects’ familiarities with the domain. Contextual measures for the evaluation of the DTU DL were collected through questionnaires and were characterized as socio – cognitive measures or individual difference measures.

The second set of measures in the thesis was used to characterize the interaction between subjects and the IIR system (DTU DL) and the subjects’ search behaviors. These measures included the number of queries (search tactics) issued by each subject as well as the number of search terms (query length). Typically, these kinds of data for measurement are collected from log data; however, they were collected by observations and notes. The third measure for the study was performance in relation to the relevance judgments of the system content. Subjects were required to state their criteria for judging a reference as being relevant. The fourth measure used in the study aimed to collect data on subjects’ feedback in order to evaluate their attitudes and feelings towards the system and their interactions with it.
4. DATA ANALYSIS

Both quantitative and qualitative methods were employed to analyze data collected for this study. Quantitative methods were employed to conduct descriptive data analysis, such as average/mean and mode/frequency. The following are different steps for data analysis procedures: (1) for each subject and all tasks, identify time lasted for tasks completion, number of search tactic(s) per subject, number of query terms per subject, subjects’ criteria for relevance judgments, (2) then calculate the average/mean time employed, number of strategies, number of search terms; (3) calculate the mode/frequency of subjects’ level of satisfaction as well as the usefulness of the system content. Mean is the sum of the scores in a distribution divided by the total number of scores, whereas mode is the score that occurs the most frequently (Kelly, 2009, p. 133).

4.1. Subjects’ profile

Table 3 displays data on subjects’ demographics. They consist of information about subjects’ current status, their average level of knowledge in interacting with computer as well as their average knowledge about the DTU DL.

<table>
<thead>
<tr>
<th></th>
<th>Current status: Undergraduates?</th>
<th>Current status: Master student?</th>
<th>Level of computer knowledge? (average)</th>
<th>Knowledge about the DTU DL?</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTU</td>
<td>3</td>
<td>2</td>
<td>3.8</td>
<td>2</td>
</tr>
<tr>
<td>IVA</td>
<td>0</td>
<td>5</td>
<td>4.2</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3: Subjects’ current status, self – confidence with computer usage and knowledge about the DTU DL
Results showed that subjects from IVA were slightly confident in their capabilities for interacting with computer. They were generally considered as computer literate with an average of 4.2, while the average was 3.8 for DTU subjects. All IVA subjects were graduates compared to 2 graduates from DTU. 3 DTU subjects were not aware of the system under evaluation, while the other 2 have used it for articles searching. With regard to IVA students, they did not have any experience with the DTU DL since it is only allowed for access to those affiliated to DTU and some guests. However, their difference in interaction with computer and the search knowledge might be some of the reasons explaining the difference between IVA subjects and DTU subjects. This was seen later in completing the assigned tasks.

4. 2. Search tasks and subjects’ interactions with the system

System’s performance could be a significant determinant of subjects’ satisfaction with the system. And this was first assessed by looking at subjects’ search performance. To this end, task 1 was an introduction where subjects were instructed to search for a topic of their interest and use “MyLibrary” function to save their results. All IVA subjects found the system very useful because it saved them time by saving the search history for a later use compared to only two DTU subjects who completed the task. The other three didn’t know what MyLibray meant exactly. One IVA student even commented: “MyLibrary contains results of previous searches – so they may be easily retraced, thus it is useful as often one remembers something regularly and then it is useful to be able to retrace the steps.”
As Kelly (2009, p. 4) commented, with such observable behaviors we must infer cognitive activity. Users who save a document may do so because the document changes or adds to their understanding of their information needs.

Task 2 also aimed to introduce subjects to the “help function” which could be a useful tool in their later search tactics. After all, this task showed how IVA students found it very useful. At every search formulation, they looked at this function to find how best they could formulate and reformulate their search queries. Furthermore, after a failed search tactic, they were quick to go back to the help function. It was not the same case for DTU students. All DTU subjects have completed the task but only one found it useful by clicking on its links. Others confused this task with the previous one. This again showed the difference between experts with knowledge in the domain of IR and those without knowledge in that domain. Vakkari (2001, p. 50-51) has noted that searchers’ IR knowledge steers their choice of options provided by the system.

### 4.3. Interaction measures

The following subsection discussed data on interaction measures – the interaction between subjects and the system and subjects’ search behaviors. Information search behavior has been defined by Wilson (2000, p. 49) as the “micro – level” of behavior employed by the user in interacting with a specific IIR system in order to search for relevant information. It consists of all the interactions with the system, whether at the level of human computer interaction or at the intellectual level. In this study, subjects’ search behaviors were measured by (1) the use of search functionalities provided by the system (i.e., adoption of advanced search strategy with the combination of AND Boolean); (2) the number of search tactics (queries) issued by subjects; and
(3) subjects’ number of search terms. These factors involved mental acts, i.e. judging the relevance of information retrieval. Data for these measures were collected through carefully observations with a focus on subjects’ query formulations. These observations might be more regarded as subjects’ behaviors when interacting with IIR. However, they might also reflect on how well the design of the system usability and usefulness supports subjects and yield data on what they like and dislike with the system.

4. 3. a. The use of system search functionalities

Data displayed in table 4 discussed the level of subjects’ use of the help system and the advanced functionality. As it is the case, the mean number of times IVA students visited the help system and used the advanced search functionality was higher than that of DTU students.

<table>
<thead>
<tr>
<th>Subjects’ ID</th>
<th>IVA stud.</th>
<th>DTU stud.</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>02</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>03</td>
<td>1</td>
<td>0,5</td>
</tr>
<tr>
<td>04</td>
<td>0,5</td>
<td>0,5</td>
</tr>
<tr>
<td>05</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>4,5</td>
<td>3</td>
</tr>
<tr>
<td>Mean</td>
<td>0,9</td>
<td>0,6</td>
</tr>
</tbody>
</table>

*Table 4: Subjects’ mean uses of the DTU DL help tools and advanced functionality.*

This data were worth to measure subjects’ information searching knowledge as well as compare what IVA and DTU students know about a particular human computer interaction domain. Given
that DTU DL allows to search with query operators, i.e., phrase matching and Boolean operators as well as prefixes for search expansion, it was expected that subjects’ searching knowledge should include at least experience using the Boolean system as well as the advanced functionality. However, DTU students made little use of interactive search possibilities provided by the system. Only 2 of them have fully used them compared to 4 IVA students. The mean of use of the DTU DL search functionalities differed between the two classes of subjects with IVA students scoring 0.9 per student compared to 0.6 per DTU student. One possible reason could be that IVA students were experienced in using IIR system search help features. This reflected the idea that subjects’ IR knowledge steered their choice of options provided by the system (Hsieh-Yee, 1993; Vakkari, 2001). One IVA subject argued: “These futures are system access points”. He added: “they allow users to solve different problems, or answer general questions about the use of the DTU DL and access of its content”.

4.3. b. Subjects’ number of search tactics

The study also looked at subjects’ interaction with the system since it may also be an important factor influencing subjects’ satisfaction with the system. This subsection discussed subjects’ search behaviors based on the number of search tactics attempted by subjects on two tasks, i.e., task 3 and task 5 which were aimed to collect data on information searching behaviors. The assumption was that the nature of information retrieval determines the dynamic process behind information retrieval, which, in turn, also leads users to shift their tactics in the retrieval process. As it can be seen from table 5, there was a difference between IVA and DTU students in their steps when working with their search tasks for retrieving relevant information.
### Mean n. of search tactics (queries) per subject

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>IVA stud.</th>
<th>DTU stud.</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>02</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>03</td>
<td>1.5</td>
<td>1</td>
</tr>
<tr>
<td>04</td>
<td>1.5</td>
<td>2.5</td>
</tr>
<tr>
<td>05</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>1.3</strong></td>
<td><strong>1.7</strong></td>
</tr>
</tbody>
</table>

Table 5: The mean number of search tactics per subject

The number of shift in search tactics has a mean between 1 and 1.2 for IVA students, while the mean for their counterparts was between 1 and 2.5. DTU students issued the most number of tactics (search queries) with mean slightly higher than that of IVA students. As shown in table 5, DTU students carried out a mean of more 0.4 search tactics than IVA students. However, this increase was due to the heavy use of search tactics by two DTU students, i.e., ID 04 and ID 05 with a mean between 2 and 2.5. Shifts in information seeking are also the product of IIR process. They occur in successive searching. Research has shown that information needs evolve during the search process; this evolution results in dynamic relevance assessments – that is, as subjects learnt more about their information needs, their relevance behaviors changed overtime (Taylor et al 2007; Vakkari & Hakala 2000, p. 541).

Subjects have experienced different levels of change in tactics at different times. This was for example seen in their unfamiliarity with the domain, the completeness of their retrieval and their satisfaction with retrieved results. Frequent cause for the shift of search tactics was the relevance judgments of the examined documents. Those who modified their tactics many times were not
satisfied with the retrieved documents due to their search formulation. Subsequently, those who did not modify their search tactics were satisfied with the retrieved documents. In sum, the above data showed that in terms of search, IVA students with IR knowledge were much precise in their search tactics compared to DTU students.

4. 3. c. Subjects’ number of search terms

As studied by Vakkari et al. (2003) and Wildemuth (2004), the number of search terms subjects employed was examined as a way for investigating the difference between subjects’ search behaviors while evaluating the system performance. In this study, search terms employed by IVA students were compared with those employed by DTU students to study the effects of domain knowledge on the effectiveness of searches. Search terms discussed here were those employed in query formulations that retrieved relevant documents to subjects’ information needs.

<table>
<thead>
<tr>
<th></th>
<th>IVA stud.</th>
<th>DTU stud.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID 01</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>ID 02</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>ID 03</td>
<td>3.5</td>
<td>3</td>
</tr>
<tr>
<td>ID 04</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>ID 05</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td><strong>Mean no. of terms per subject</strong></td>
<td><strong>3.2</strong></td>
<td><strong>4.1</strong></td>
</tr>
</tbody>
</table>

*Table 6: Average number of query terms per subject. (Mean is the sum of the scores in a distribution divided by the total number of scores) (Kelly, 2009, p. 133)*
Statistically, data displayed in table 6 showed that search terms for those two tasks ranged between 2 and 5 for IVA students, while it was between 2 and 9 for DTU students. One of DTU students employed up to 9 terms on each of the two tasks. However, results for her retrieved items were not as relevant to her information needs. IVA students had a mean of 3.2 terms per student, while the DTU students had a mean of 4.1. This difference could be justified by the high number of search terms employed by DTU subject ID 05 with an average of 9 terms. It should be noted that most IVA students employed more specific terms, though they were not familiar with the domain under which the system performs. This might be justified by their knowledge in the domain of information searching (IR).

4.3. d. Summary

Results for this study are similar to Fidel (1991, p. 515) subjects’ characteristics of searching behaviors: individual searchers differ from one another in their degree of interaction during a search, and in their preference for type of move and for type of search key. There was a difference in subjects’ search behaviors. Experienced searchers utilized frequently the system help tools for searching as well as the advanced search functionality than DTU students. Moreover, IVA students employed less search tactics as well as less search terms than DTU students. According to Kelly (2009, p. 4), subjects might have a different cognitive composition and behavioral disposition when performing tasks. And they may vary according to different factors including how much they know about a particular topic, how much they know about searching, how much they know about the particular work or search task that they need to complete, and even what they expect and perceive from the IIR system (Ingwersen & Järvelin, 2005).
Changes in tactics on one hand may be interpreted as the increase in domain knowledge of subjects. Furthermore, the changes in tactics during the processes of information seeking could be explained by the increase in search expertise. The increase in search skills could be the reasons for the increased use of Boolean systems as well as other prefixes provided in the help system. It is empirical that domain knowledge has an impact on users' search tactics (Xie 2009, p.66). However, this was not the case. The increase in search tactics by DTU students did not reflect the increase in use of Boolean systems. Domain knowledge has different effects on users with different levels of information retrieval knowledge. Hsieh-Yee (1993) studied the effects of subject knowledge on search tactics of novice as well as experienced searchers.

On the other hand, performing more searches and trying at the same time more query terms reflect one move of making more effort in seeking relevant information. However, this additional effort did not reflect an effective searching for DTU students. This might be justified by the difference between the domain knowledge and the searching (IR) knowledge, as it was showed by previous researchers (Yee, 1993). These results revealed that the level of domain knowledge may not be important unless the subject has a certain level of searching expertise.

A study by Bhavnani (2002) found that only domain – specific search knowledge is important for effective searching. Based on the average query terms per subject, the use of query terms in this study revealed that subjects, as a whole had a limited familiarity with the technical domain search system under investigation. One subject (IVA student) argued: “I had a harder time formulating queries due to the subject being unfamiliar to me”.

Studies have shown that domain knowledge has an impact on searching assuming that users have sufficient command of the system in use. However, results for this study showed that both domain knowledge and information searching (IR) knowledge are needed to perform search tasks. This
study suggested that both concepts are needed and complementary in the evaluation of the usability and usefulness of IIR systems. It was noted that both domain knowledge and IR knowledge have an effect on searching. The only study on the interaction of both (domain and IR knowledge) by Hsieh-Yee (1993, p. 169) suggested that knowledge has an effect on searching if the searcher is experienced in mastering the system used.

4. 4. Subjects’ evaluation data

This subsection discussed subjects’ evaluation data employed in the study. These data were collected on (1) subjects’ evaluations of search results, thereof the relevance assessments of retrieved documents; (2) criteria for relevance judgments; and (3) the system’s document presentation order.

4. 4. 1. Evaluation of search results.

Data in this subsection were collected on subjects’ relevance assessments of documents. The display strategy of the DTU DL was tested by considering the portion of supposed relevant documents identified out of a set of documents retrieved from the system. To this end, a set of 10 top documents extracted from the system’s ranking were presented to subjects who were then instructed to assess them for their relevancy. Cognitive perspectives played an important role on relevance assessments which varied according to how much subjects were familiar with the domain (Wen et al. 2006, p. 456) compared to the number of relevant documents available to the particular subject (Huang & Wang 2004). The assessments were done on a five point Likert scale with variables strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree. Only
documents on variables agree and strongly agree were considered and combined in this study as
displayed in table 7

<table>
<thead>
<tr>
<th>Doc. no.</th>
<th>IVA stud.</th>
<th>DTU stud.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>4</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 7: Subjects’ evaluation of search results based on document relevance assessments.

Borlund and Ingwersen (1998, p. 325) have called this type of relevance a subjective or human
(user) – based relevance which may refer to the usefulness, usability, or utility of information
objects in relation to the fulfillment of goals, interests, work tasks, or problematic situations. It is
also a situational relevance in that it is concerned with the usefulness of sought objects (Ingwersen

In this study, results showed that documents 1, 2, and 10 were judged as highly relevant by 9
subjects out of ten that participated in the experiment. And 7 out of the ten documents were judged
as relevant by between 5 and 9 subjects. One justification for this agreement could be that subjects
found these documents relevant to their information needs. This high agreement on the degree of
relevance between subjects with different domain knowledge might indicate that when an IIR system believes that a given document is relevant, humans are also likely to agree on that relevancy. Although there was a high degree on the agreement of relevant documents, there was also an indication that the level of relevance assessments differed slightly between IVA and DTU students. While for example all IVA students agreed on 1, 2, and 10 as relevant documents, only four DTU students agreed on the same documents. This degree of agreement on relevance was also compared on documents three, and five respectively. It should be noted that these two classes differed slightly in the ways they assess the relevance of documents based on their different domain knowledge as it is discussed in the following subsection.

4.4.2. Relevance judgments criteria

Data for this subsection were collected by requiring subjects to state their criteria for relevance judgments of documents. This was to highlight whether the criteria they did use in assessing the relevance of documents were also part of the references. But as Ingwersen asked (1992, p. 43): “which objects were to form the base – line assessments: bibliographic details, i.e., titles and/or abstracts, the full - text documents, each single or several semantic entities or passages? Judging the relevancy of documents of course involved some interactions between subjects and the interactive information retrieval system under investigation including scanning text, reading abstracts or other forms of document representations, or even going through the entire document. The judgments made by subjects reflected their cognitive state as they related to their information problem. Data depicted in table 8 revealed that there was a difference on which criteria documents were judged useful between the two classes.
Table 8: Subjects’ criteria for relevance assessments

All five IVA students agreed on abstract as the main criterion while only two DTU students agreed on this criterion. This was also seen on the full-text criterion with four out of five IVA students compared to only one DTU student. In general, students with expertise in the domain of information searching (IR) mostly judged documents based on bibliographic entities, i.e., title, abstract, keywords as well as full-text. However, beyond the bibliographic entities as relevance criteria, results revealed another difference, where DTU students, with their expertise related to the technical domain at the same time judged documents based on other semantic entities, including illustrations, tables, graphs, theories as well as methods. It should be noted that this difference in criteria for relevance judgments between these two classes reflected their differences in their areas of studies. It was found by Toms et al. (2003) that searchers tend to compare information currently available to them with their previous knowledge.

4.4.3. Effects of documents presentation order on system evaluation

This subsection aimed to investigate how the document presentation order affects subjects’ level of satisfaction with the system. The concept of relevance judgment is critical when evaluating an
IIR system. Data for this investigation are displayed in table 9. It compared results of relevance scores of 10 documents ranked from high to low document.

<table>
<thead>
<tr>
<th>Doc. no.</th>
<th>IVA stud.</th>
<th>DTU stud.</th>
<th>Total</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>4</td>
<td>9</td>
<td>4.5</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>4</td>
<td>9</td>
<td>4.5</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>4</td>
<td>9</td>
<td>4.5</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>3.5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>3.5</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>3</td>
<td>7</td>
<td>3.5</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 9: Documents presentation order on relevance assessments

Result from an empirical study has revealed that relevance assessments are influenced by the presentation of documents order. This result has been named “order effects” (Eisenberg & Barry, 1988); a different presentation order of documents brings on different influences.

Results in this study have shown that DTU DL lacks a mode of ranking from high to low relevant documents. They revealed that the degree of relevance was high on documents 1 and 2 with a mean of 4.5; it then decreased on the following documents with a mean between 0 and 3.5; and increased again on document 10 with a mean of 4.5.

For IIR systems, i.e., digital library, usability and usefulness, i.e., collection/system content are two sides of the coin. The ultimate objective of an IIR system is to satisfy users’ needs. Users will normally judge the value of IIR system by evaluating its content. For an IIR system to be effective, its retrieval and usability through the interface has to be perfect as well as display its content with a
rich ranking mode. The DTU DL documents presentation order affected its overall satisfying degree and prompted subjects to judge differently the performance of the system.

One subject argued: “I expected that all relevant documents could be on top ranks in the search results”. Another subject argued: “It is a good system but its ranking mode is not helpful. There should be different functionalities for scoring retrieved results. It is a Google like system”. It should be noted that the retrieval of information and the evaluation of the performance of DTU DL can also rely on the concept of relevance judgment through the system documents presentation order.

4. 4. 4. Summary

Results from this subsection revealed some differences and similarities on the criteria the two classes of subjects employed to judge the relevance of documents. There was a slight similarity between subjects on the judgments of relevant documents. Between 3 and 5 IVA students judged documents based on bibliographic entities compared to only 2 DTU students. However, they differed on other semantic entities. A number of DTU students as technical domain experts employed illustrations, tables, graphs, theories and methods fund in the documents as a supplement for document assessments criteria. For the system performance, result from the study revealed that the implications of ranked output can be useful for the judgments of interactive information retrieval (IIR) systems. Spending time browsing for a desired document beyond the first result page or performing new queries requires more time and effort, which may frustrate users. Therefore, this may hamper the successful use of the whole IIR system.

4. 5. Usability and usefulness evaluations

This section discussed two interconnected evaluation parameters, such as usability and
usefulness. Data on usability of DTU DL were collected through the time lapsed for performing tasks as well as the evaluative feedback elicited from subjects. These data measured the performance of the system based on the following attributes: efficiency, satisfaction including navigation and ease of use. The parameter of usefulness was measured on the subjects’ degree of acceptance of the system content.

4.5.1. Performance data: time

Data for this subsection focused on the calculations of the time lasted for tasks completion compared to the number of tasks successfully performed in order to assess the efficiency of the system. These data were compared between IVA and DTU students as displayed in table 10.

<table>
<thead>
<tr>
<th></th>
<th>Overall time lapsed</th>
<th>Tasks completed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IVA stud.</td>
<td>DTU stud.</td>
</tr>
<tr>
<td>ID 01</td>
<td>59' 21&quot;</td>
<td>82' 09&quot;</td>
</tr>
<tr>
<td>ID 02</td>
<td>95' 36&quot;</td>
<td>57' 03&quot;</td>
</tr>
<tr>
<td>ID 03</td>
<td>42' 09&quot;</td>
<td>62' 13&quot;</td>
</tr>
<tr>
<td>ID 04</td>
<td>58' 27&quot;</td>
<td>81' 15&quot;</td>
</tr>
<tr>
<td>ID 05</td>
<td>71' 28&quot;</td>
<td>49' 15&quot;</td>
</tr>
<tr>
<td>Total</td>
<td>329'</td>
<td>332'</td>
</tr>
<tr>
<td>Mean</td>
<td>65'</td>
<td>66'</td>
</tr>
<tr>
<td>Mean of time/task</td>
<td>7 minutes/task</td>
<td>8 minutes/task</td>
</tr>
</tbody>
</table>

*Table 10: Comparison of mean time per task per subject*

Results revealed a difference in time lasted for tasks completion as well as a difference between the numbers of tasks completed successfully. However, it was clear later in the post – session interview
that this difference was not the cause of the system’s poor state of performance. Three observations were made: (1) Though the mean of time lapsed for performing tasks per student was slightly equal in both two groups (i.e., 65 and 66 minutes), there was a difference on the average number of completed tasks; (2) IVA students completed tasks with a mean of 9, while DTU students have a mean for 8; (3) The mean of time per task was then 7 minutes per IVA student, while it was 8 minutes per DTU student. This was interpreted in their differences on demographic data as well as the experience in IIR. Furthermore, these observations showed that the more IVA students used time the more they retrieved relevant documents and assessed them. It is worth noting that information searching experts were better in both using the time and assessing at the same time the system performance and usefulness. Asked in the post – session interview if the lasted time was reasonable, one IVA participant argued: “It required much more time for to go through the documents looking at different bibliographic entities and other forms of document representations covered by them in order to judge their relevance, hence the effectiveness of the system’s performance”.

4. 5. 2. Subjects’ level of satisfaction

This subsection discussed subjects’ level of satisfaction with the system. Data for this attribute were collected based on system navigation as well as its ease of use as perceived by subjects.

4. 5. 2. a. System navigation

System navigation attribute can affect its interaction with users and consequently with search tasks. For this study navigation was measured based on how much subjects were aware with the location of navigation tools, i.e., help function. As noted by Tsakonas and Papatheodorou (2008, p.
406) easy to navigate systems allow users to conclude their information and work tasks successfully.

Data for this usability measure were collected using the post – task questionnaire including two search tasks that required subjects to rate their satisfaction level toward the system navigation in regard to supporting the accomplishment of tasks, using a five point Likert scale. These data were then calculated and their mean was compared with the overall average of the subjects’ level of satisfaction with the system through a post – session interview at the end of the experiment as displayed in table 11.

<table>
<thead>
<tr>
<th>ID 01</th>
<th>ID 02</th>
<th>ID 03</th>
<th>ID 04</th>
<th>ID 05</th>
<th>Total</th>
<th>Mean</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>IVA</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>18</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>DTU</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>IVA</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>21</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>DTU</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>24</td>
</tr>
<tr>
<td>IVA</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>21</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>DTU</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>20</td>
</tr>
</tbody>
</table>

| Task 2 (navigation) | IVA stud. | 3 | 4 | 4 | 4 | 3 | 18 | 3.6 | 4 |
| Task 2 (navigation) | DTU stud. | 2 | 4 | 5 | 4 | 4 | 19 | 3.8 | 4 |
| Task 4 (System help) | IVA stud. | 3 | 4 | 5 | 5 | 4 | 21 | 4.2 | 5 |
| Task 4 (System help) | DTU stud. | 4 | 5 | 5 | 5 | 5 | 24 | 4.8 | 5 |
| Overall level of satisfaction | IVA stud. | 4 | 4 | 4 | 4 | 5 | 21 | 4.2 | 4 |
| Overall level of satisfaction | DTU stud. | 4 | 4 | 4 | 4 | 4 | 20 | 4 | 4 |

*Table 11: Subjects’ level of satisfaction with the system*

The mean of satisfaction was 3.9 per IVA student, whereas their overall average for satisfaction with the system was 4.2. The mean of satisfaction was 4.3 per DTU student, whereas their overall average for satisfaction with the system was 4. The mode in both classes ranged between 4 and 5.
These data showed correlation on subjects’ level of satisfaction both on system navigation and the overall satisfaction with the system. However, the interview session revealed that there were some problems with the system - although not very serious but that needed some attention. This was discussed later in subsection 4.6. Although there were some complaints, majority of subjects found the system as well to navigate in general. This was revealed by their level of satisfaction which can be translated into their perceptions on system acceptability.

4.5.2. b. System ease of use

This indicator of satisfaction was measured on task 3. Subjects were required to grade the system’s ease of use on locating a known information need employing provided system functionalities.

<table>
<thead>
<tr>
<th></th>
<th>ID 01</th>
<th>ID 02</th>
<th>ID 03</th>
<th>ID 04</th>
<th>ID 05</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>IVA stud.</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>4.2</td>
</tr>
<tr>
<td>DTU stud.</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Table 12: Data on subjects’ perceptions on system ease of use

Data depicted in table 12 showed that most subjects found the system easy to use when searching for a known item. There was a slight difference between IVA students and DTU students which again might be caused by their difference in information searching. However, majority of IVA students with an experience in information searching explained that the system was quite easy to use to find documents when some bibliographic records, i.e., title, author are provided. While the mean was 4.2 for good searchers, i.e., IVA students, it was 3.8 for DTU students. This could be justified by DTU subject ID 05 who was not satisfied with the system ease of use. It should be
noted that this criteria may stimulate and further encourage the use of the system or may frustrate and lead the user to abandon his or her task. Therefore, if a tool is not easy to use, then it is likely to result in inefficient use (Kelly, 2009, p. 120).

4.6. Errors observed under the experiment on the system

Data for this subsection were collected by observing subjects while performing their assigned tasks through some errors and problems encountered during the experiment – errors and problems that were classified into two categories. The first category was concerned with the system design and navigation. Some users were lost in the course of the experiment and this was most observed on the assessment of retrieved documents. At this stage, some subjects found it hard to go back to the search window, thus, the only solution was to close the current window with results and reopen the DTU DL search site. Frustrating! One of the subjects even commented that: “it was difficult for her to start navigate the system”. Another reported that: “the help system was not optimal, thus some users may lack possibilities for searching through the system” and this was regarded as a problem with the system design.

The second category of errors was concerned with some non-informative links. A number of subjects who managed to go through the help function found that the “feedback tool” was not really informative. They lacked some information about the “FAQ, Frequent Asked Questions”. In the same category, some also found that the “DTU blog” was not functional. While the first category of error was regarded as a problem with system design and navigation, the second category was attributed to system information organization.
4.7. Subjects’ perceptions on search tasks

Tasks 8 and 9 aimed to analyze tasks’ impact on users’ search behavior. This subsection collected data on subjects’ personal perceptions on search task complexity/difficulty through post session interview as a follow up. Data is displayed in table 13.

<table>
<thead>
<tr>
<th></th>
<th>ID 01</th>
<th>ID 02</th>
<th>ID 03</th>
<th>ID 04</th>
<th>ID 05</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>T8</td>
<td>IVA stud</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>DTU stud</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T9</td>
<td>IVA stud</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>DTU stud</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 13: The effects of search task on user search behavior (0 stands for not complex, while 1 stands for complex). (NB: 0 represents task non-complexity, while 1 represents task complexity)

Results showed that search task 8 was not difficult for 9 subjects all classes combined. However, and surprisingly, task 9 was complex for all IVA students, while 3 DTU students found the task being not hard to perform. This might lead to the suggestion that domain knowledge and information search knowledge are key factors for performing search tasks. Bell & Ruthven (2004, 9. 59) have suggested that task difficulty can be affected by the following three factors: (1) the difficulty of understanding the required information. It may require specialist knowledge about the task domain. It should be noted that both the knowledge of the system information domain and the information searching knowledge are required to perform such a task. (2) The difficulty of searching; and (3) the difficulty of interpreting the relevance.
4. 8. The usefulness of system content

Usefulness was employed as a criterion for evaluating the DTU DL in order to investigate how the content offered by the system meets users’ requirements (Hartmann, 2006, p. 1756; Savolainen, 2008, p. 278). While usability concentrated on system quality, i.e., the interaction between subjects and the system, usefulness focused on the interaction between the subjects and the system content, i.e., the quality of the information covered by the system. These two concepts have been studied in the area of acceptance and success of information systems. And this was also part of the perspectives of this study in response to the question of what kind of data is covered by the system in order to support users with different domain knowledge, thus for it to be accepted. System quality and information quality (as equivalent of good state usability and usefulness respectively) have an effect on system usage and user satisfaction, which as a consequence may influence individual and organizational performance (Tsakonas & Papatheodorou, 2006, p. 401).

Data on the usefulness of the system content were collected on tasks 5 and 6 and are displayed in table 14. Subjects were asked to evaluate the content of the system based on retrieved information. These data were then combined to generate the overall mean. To this end, usefulness was employed as an extension to the concept of relevance.

<table>
<thead>
<tr>
<th></th>
<th>ID 01</th>
<th>ID 02</th>
<th>ID 03</th>
<th>ID 04</th>
<th>ID 05</th>
<th>Mean</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>IVA stud.</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4.4</td>
<td>5</td>
</tr>
<tr>
<td>DTU stud-</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 14: Comparing the mean of the degree of usefulness of system content
According to Cole et al. (2009, p. 3), the concept of relevance can be used as a usefulness criterion for interaction steps where the immediate objective is to gather topical documents. In this case, it is the aboutness of a document that constitutes its usefulness to advancing the task, so relevance is the appropriate usefulness criterion. The quality of the system content was measured based on various variables, such as the level of the content, i.e., abstracts, full-text, citations, PDF and so forth; as well as the timelines of the content which referred to how current the information resource was. Results showed that majority of subjects accepted the system based on the usefulness and relevance of its content. One subject argued that:

“It looks quite to other modern digital libraries as it is functional with full – text in PDF on found materials. The system performance is quite rich because the objectives of users using it are to find useful and relevant information in different formats.”

Results showed that the mode was equal, 5 for both sides. However, the mean was higher per DTU student compared to that of per IVA student, 5 and 4.4 respectively. Only the targeted users of an IIR, i.e., digital library can determine the usefulness of its content. Therefore, if the content of an IIR is of no use for its intended end – users, the users will have little or no reason to even browse the content of the system.

5. DISCUSSIONS AND CONCLUSION

The thesis discussed about some issues for understanding the IIR evaluation of system’s usability and usefulness in relation to subjects with different domain knowledge. The evaluation of IIR has become a more cognitive process with the utilization of various measures based on user – involvement and situational approaches. The experiment was designed to investigate the effects of subjects’ domain knowledge on the evaluation of usability and usefulness of an IIR. Data on
subjects’ similarities/differences with respect to interaction with the system and the system content were analyzed.

5. 1. Evaluation criteria

On one hand, usability is measured toward effectiveness, efficiency, and user satisfaction (ISO 9241, 1998). On the other hand, usefulness in user–centered evaluations is the degree to which a specific information item will serve the information needs of the user (Tsakonas & Papatheodorou, 2005, p. 402). Qualitative measures, i.e., observations and interview are important to evaluate the usability of an IIR from users’ subjective perceptions. Quantitative criteria, i.e., questionnaires are important for the objectivity of usability and usefulness evaluation. The search of the system and the display of its output are key factors in IIR evaluations. However, information behaviors which differ from user to user can affect the evaluation of an IIR. These behaviors consist of various activities, such as information seeking, searching, retrieval, use, and interaction with system interfaces (Ju, 2007, p. 2007). Subjects with different domain knowledge evaluate an IIR differently. Criteria for evaluation can differ from subject to subject in respect with his/her domain knowledge.

The quantitative evaluation criteria employed in the experiment were analyzed based on some factors: (1) subjects with different information searching behaviors can affect the evaluation of an IIR. These behaviors were measured through different dependent variables including (a) tasks completion time, (b) the number of search tactics issued by a subject; (c) number of search terms employed by a subject; (d) the relevance judgments of system content; (e) the criteria for relevance judgments; (f) as well as the order of document presentations; (2) the relative relevance can be employed to measure the usefulness of the system content for the system to be accepted. Relevance in the context of system usability and usefulness is associated with the issue of how well users are
able to accomplish tasks with the system as well as how well the retrieved information contributes to the user information needs. The design of the results display of an IIR system output can be one of the reasons that my influence users to judge the system’s performance.

5.2. The design of system’s usability and usefulness

Two classes with different domain knowledge were employed in the experiment, i.e., masters in library and information science and students in a technical domain. The former with information retrieval experience lacked the domain knowledge under which the system works. The latter has knowledge on the domain of the system under evaluation but lacked the information retrieval experience. Data from observations and interview in the experiment indicated that there was a difference between subjects’ use of DTU DL search facilities. Experienced searchers seemed to use more advanced search strategy while the other (DTU students in the experiment) used the simple search mode. The goal of the usability and usefulness design of IIR systems should be to accommodate various patterns as well as satisfy users with different domain knowledge. As a solution to this issue, DTU DL provides different search interfaces, i.e., simple as well as advanced searches like other modern IIR systems. Both observations and results in the experiment showed that the advanced search was real. Furthermore, subjects who interacted with advanced features and consulted other search features provided by the system made their searches productive and more accurate.

A good design of an IIR system consists of varied features that can influence users to accept it. Although the design of DTU DL provides a number of good features, i.e., choice of searching field, search within results through keywords, citation links and so forth, its results ranking mode as well
as its non–informative links need to be adjusted. Routinely information searchers felt that some users could take long going through pages in order to find relevant information. It was suggested that the system should provide users with different features for presenting relevant documents.

Regarding system usefulness, subjects liked the DTU DL content. They were satisfied on the fact that the system provided documents with abstracts, full text, PDF formats, descriptors and so forth. Furthermore, subjects liked the ability for navigating the system through different specific facets (search fields), i.e., articles, books and journals in order to limit their search. Some noted that “the display of information was quite nice as one could look at the hard covers of the displayed documents”.

The usability and usefulness of an IIR system requires the knowledge in query formulation, the organized system for search as well as the system display of the content. Qualitative data provided the overall perceptions on subjects’ satisfaction with the system. Satisfaction was found as a good attribute for measuring the system usefulness (subjects’ interaction with the system content) as well as system usability (subjects’ interaction with the system). In short, the system was found supportive for both classes of subjects.

5.3. Conclusion

Designing IIR systems in order to help users express their information needs and understand the returned results as well as evaluating resulting systems are core of the HCI community (Anick, 2003). However, users evaluate an IIR system with respect to how efficient and effective the system can provide them with useful information contents in order to satisfy their information needs. Two main issues are concerned with the IIR system design and development: (1) the interaction between the user and the system as users focus mainly on the process of searching and finding useful
information; and (2) the interaction between the user and the system content. The former leads to usability studies of IIR systems, whereas the latter leads to the study of system usefulness. While the goal of usability of an IIR system is to study users’ interaction behaviors with the system, i.e., information searching as well as the system searches so that the system can be improved or redesigned, that of its usefulness is conducted through relevance judgments of retrieved documents displayed on its search screen.

The evaluation of usability and usefulness of the DTU DL was done through some improvements to the traditional Cranfield experiment, such as the employment of predefined search tasks, simulated work tasks, as well as subjects with different profiles and domain knowledge. Though precision and recall as traditional measures in the Cranfield experiment were not part of the experiment, results revealed that usability and usefulness of an IIR system is related to the individual cognitive process. Users’ subjective perceptions, opinions as well as behaviors and system objectivity are factors that may affect the system evaluation. And a number of measures are needed to characterize the interaction between the user and the IIR system and the users’ search behaviors. These measures include the number of users’ search tactics; that of users’ search terms as well as observations of users’ search behaviors. On the other hand, measures for the usefulness of the system consist of the relevance judgments of documents as well as criteria for relevance judgments for system content. Results in this study have shown that system content relevance judgment is a worth measure in order to evaluate the design of an IIR system results display.

Despite some observed errors, the DTU DL was accepted as a usable and useful IIR system by majority of subjects. Although they were satisfied with usability and usefulness of the system through some features, like the system navigation, system ease of use, system content, i.e., abstracts, full – texts, PDF and so forth; the easy way provided by the system to save searched
results for further reuse; search help tools; and so forth, some areas marked their discomfort which need to be taken into consideration.

In regard with usability and usefulness evaluation of an IIR system through users with different domain knowledge, the thesis found that both the concepts of domain knowledge and that of IR knowledge are needed. Previous studies have shown that domain knowledge has an impact on searching assuming that users have sufficient command of the system used (Vakkari, 2003). However, results in this study revealed that these two concepts are complementary and have an effect on searching, hence the effects on all the evaluation process. It has been suggested that knowledge has an effect on searching only if the searcher is experienced mastering the system in use.

5.4. Limitations of the study

One of the limitations observed in this study was on the number of subjects who participated in the experiment. Subjects were recruited for availability sampling. The numbers of subjects eligible to participate in the experiment were relatively small because the time of their recruitment coincided with the period of their exams. Data from this small number of subjects could not provide substantive results for the evaluation of the system. Another limitation of this study was that the DTU HCI facilities lacked the possibilities for videotaping and that of data logs in order to more validate results. However, despite these limitations, the comparison of the performance of two classes of subjects with different domain knowledge expertise helps better understand the information problem solving process, which can, in turn, be employed to design and develop more usable and useful IIR systems.
6. REFERENCES


**Home pages**

Adobe Connect: [www.connect.forskningsnettet.dk](http://www.connect.forskningsnettet.dk)

DTU Digital Library: [http://www.digitallibrary.dtu.dk](http://www.digitallibrary.dtu.dk)

Userminds: [http://my.intrateam.dk/da/anbefalet-link/userminds-brugervenlighedstest-m-m](http://my.intrateam.dk/da/anbefalet-link/userminds-brugervenlighedstest-m-m)
7. APPENDIX

Appendix A: Pretest questionnaire: demographic data

Subject ID-------

I would like to thank you for accepting to participate in this experiment which is scheduled to last approximately 60 minutes. For the ethic reasons, all your personal data you provide to me are to be treated with confidentiality, viewed only by the experimenter and the supervisor of the thesis. Another consideration is that the entire test is not about you, rather the system under evaluation. With this regard, I would ask you to provide me with information about you and your background to enable me to better interpret (1) your use of the system, and (2) your reactions (perceptions) to the system.

1. Subject’s name: -----------------------------------------------

2. Subject’s gender:

   - Male: ---

   - Female: ---

3. Subject’s age: ---

4. University name:

5. What is your current status?
• Undergraduate ------------------------------------------

• Master student ---------------------------------------

6. How many years have you been at your University? ---

7. Your area of education (faculty) -----------------------

8. What is your level of computer knowledge ----1 --- 2 ---- 3 ---- 4 ---- 5
(5 being very excellent)

9. Do you have any knowledge about the DTU Digital Library? ------

10. How often do you use the DTU DL? (To be filled by DTU students only)

• Once or twice a week ----

• Once or twice a month ----

• Once or twice a semester ----

• Every day ----

• Never ----

Date: ------------------
Appendix B: Post - task questionnaire on usability and usefulness evaluation

Subject ID--------

The aim of this test is to evaluate the performance and usability of the DTU DL. To this end, I would ask you to perform some tasks and try to talk aloud what you are doing. There is also a room for your comments and I would also ask you to talk aloud when you answer the question on the questionnaire.

As said, remember the entire test is not about you, rather the system under investigation, and your answers will be anonymous. The test is to last … and after you will be asked to provide with a general overview of the navigation and performance of the system.

I. Introductory tasks

Task 1: Enter the DTU DL search portal. Search for a topic of your interest and save your results in MyLibrary.

1. What do you expect MyLibrary contains?

2. Is MyLibrary useful for you?

   Yes----------------------------------------------------------------------------------------------------------------------

   No------------------------------------------------------------------------------------------------------------------------

Task 2: Navigating the system: if you need some help to DTU DL, is there any function for that?
If yes, is the information in the help system clear and useful?

Very unsatisfied ------1 ------2 ------3 ------4 ------5 Very satisfied

Please give your comment: --------------------------------------------------

II. Search tasks

Task 3: Finding a known item: the easy for use of the system


Please talk out loud about your search strategy. How easy was it to use the system in order to find your information need?

Difficult to use ------1 ------2 ------3 ------4 -- ------5 Easy to use

Please give your comment--------------------------------------------------

Task 4: Assess one of the relevant results in search 4 by clicking on these two icons (+) and (-).

What do these two icons tell you? --------------------------------------------------

What is your level of satisfaction here?
Task 5: Usefulness of the system content


Please talk out loud and tell me about your search strategy. How useful is the retrieved information according to your information needs?

Not useful -----1 -----2 -----3 -----4 -----5 Very useful

Please give your comment: -------------------------------------------------------------------------------------------------------------------------------------

Task 6: Usefulness of the system content

Search for the topic” Exploring preoperational features in chronic depression” where the search results include books, e-books and articles.

Assess the results:

Please talk out loud and tell me about your search strategy. How useful is the retrieved information according to your information needs?
Not useful -----1 ------2 ------3 ------4 ------5 Very useful

a. How would you find the relevant information? -----------------------------------------------

b. What criteria that makes you to judge them relevant? -----------------------------------------------

Give your comment-------------------------------------------------------------------------------------

III. System ranking task, task complex and relevance assessments

Task 7: You are writing your thesis on “cognition”. A friend has just told you about an author “Daniel Reisberg” who has written some books on this topic.

a. Please find books on “Cognition”

b. Then search for cognition and Reisberg.
Assess retrieved information and tell me why you think you didn’t find the Reinsberg books during the first search (search a)

Please give your comment: ---------------------------------------------------------------------------

Task 8: Your friend has been working overtime and now he is suffering from a disease which may be a consequence of overworking. Interestingly, you have learnt how devastatingly hard it can be to cope with some of these kinds of diseases.
Find information that provides advices on how to cope with it and how it can be treated without drugs.

1. Find 10 top documents for this task. Are they relevant to your information need?

<table>
<thead>
<tr>
<th></th>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Neither agree or disagree</th>
<th>Agree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc. 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. What are the criteria for judging them?

Please give your comment

3. Was the task understandable to you?

Please give your comment
**Task 9:** While on a tour in USA, one of your brother’s guests who is working with an International Organization was complaining about the high price of commodity and all the situations that influence this phenomenon.

When back to Copenhagen, you decided to research and find information on worldwide factors that are important in deciding the price of commodity.

1. Find 10 top documents for this task. Are they relevant to your information need?

<table>
<thead>
<tr>
<th></th>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Neither agree or disagree</th>
<th>Agree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc. 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc. 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. What are the criteria for judging them relevant?
Please give your comment

3. Was the task understandable to you?

Please give your comment

Appendix C: Post – session questionnaire

Subject ID-------

Thank you for you time you took for the test and the feedback you provided. The last session of the test is an interview about your overview of the system.

Please tell me what you think about the DTU DL by talking aloud while you answer to the following questions

The system and the interface

- Is there any thing you liked/disliked about the DTU DL search portal?
- Was the system useful to you and easy to learn?
- Is there anything that makes the portal hard to navigate?
- Where exactly occurs the problem: give any suggestions on how the search portal could be improved?

Task completion
- Have you been able to accomplish your tasks?
- If you have not, what was the problem according to you
  a. Difficult tasks?
  b. Provided time?
  c. The system itself?
- What kinds of changes would you come with in order to assist users?

Please assess the performance of the system:
- Have you easily located needed information?
- Were the search results relevant to your queries?
- What about links to different source of information? Were they sustainable?
- Please give your last comments on your level of satisfaction.