EU-IST Project IST-2003-506826 SEKT

SEKT: Semantically Enabled Knowledge Technologies

D8.4.1 Results of User Tests and Completed Use Case Studies

Tom Bösser (KEA), Pompeu Casanovas (UAB), Nuria Casellas (UAB), Elke-Maria Melchior (KEA), Ian Thurlow (BT), Joan-Josep Vallbé (UAB)

Abstract

The procedures and results of the user tests and validation of the SEKT case study applications are described. Tests with representatives of the future user population of the BT digital library application, and of the IURISERVICE application were conducted. The results show that users are highly positive about the expected benefit of the SEKT enabled applications, as compared to their traditional tools available to them now. We see decisive performance increases as well as improvements in the quality of the information obtained in the search process.

On the basis of these results we can state with confidence that the addition of SEKT functionality will be welcomed by users and will produce significant benefits.

Keyword list: IURISERVICE, BT digital library, case study, applications, user tests

WP8 Usability and Business Benchmarking

Report	RE
Contractual date of delivery: 31/12/2006	Actual date of delivery: 01/03/2007

SEKT Consortium

This document is part of a research project partially funded by the IST Programme of the Commission of the European Communities as project number IST-2003-506826.

British Telecommunications plc.
Orion 5/12, Adastral Park
Ipswich IP5 3RE
UK
Tel: +44 1473 609583, Fax: +44 1473 609832
Contact person: John Davies
E-mail: john.nj.davies@bt.com

Empolis GmbH
Europaallee 10
67657 Kaiserslautern
Germany
Tel: +49 631 303 5540
Fax: +49 631 303 5507
Contact person: Ralph Traphöner
E-mail: ralph.traphoener@empolis.com

Jozef Stefan Institute
Jamova 39
1000 Ljubljana
Slovenia
Tel: +386 1 4773 778, Fax: +386 1 4251 038
Contact person: Marko Grobelnik
E-mail: marko.grobelnik@ijs.si

University of Karlsruhe, Institute AIFB
Englerstr. 28
D-76128 Karlsruhe
Germany
Tel: +49 721 608 6592
Fax: +49 721 608 6580
Contact person: York Sure
E-mail: sure@aifb.uni-karlsruhe.de

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
UK
Tel: +44 114 222 1891
Fax: +44 114 222 1810
Contact person: Hamish Cunningham
E-mail: hamish@dcs.shef.ac.uk

University of Innsbruck
Institute of Computer Science
Techikerstraße 13
6020 Innsbruck
Austria
Tel: +43 512 507 6475
Fax: +43 512 507 9872
Contact person: Jos de Bruijn
E-mail: jos.de-bruijn@deri.ie

Intelligent Software Components S.A.
Pedro de Valdivia, 10
28006
Madrid
Spain
Tel: +34 913 349 797
Fax: +49 34 913 349 799
Contact person: Richard Benjamins
E-mail: rbenjamins@isoco.com

Kea-pro GmbH
Tal
6464 Springen
Switzerland
Tel: +41 41 879 00
Fax: 41 41 879 00 13
Contact person: Tom Bösser
E-mail: tb@keapro.net

Ontoprise GmbH
Amalienbadstr. 36
76227 Karlsruhe
Germany
Tel: +49 721 50980912
Fax: +49 721 50980911
Contact person: Hans-Peter Schnurr
E-mail: schnurr@ontoprise.de

Sirma Group Corp., Ontotext Lab
135 Tsarigradsko Shose
Sofia 1784
Bulgaria
Tel: +359 2 9768 303, Fax: +359 2 9768 311
Contact person: Atanas Kiryakov
E-mail: naso@sirma.bg

Vrije Universiteit Amsterdam (VUA)
Department of Computer Sciences
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands
Tel: +31 20 444 7731, Fax: +31 84 221 4294
Contact person: Frank van Harmelen
E-mail: frank.van.harmelen@cs.vu.nl

Universitat Autonoma de Barcelona
Edifici B, Campus de la UAB
08193 Bellaterra (Cerdanyola del Vall` es)
Barcelona
Spain
Tel: +34 93 581 22 35, Fax: +34 93 581 29 88
Contact person: Pompeu Casanovas Romeu
E-mail: pompeu.casanovas@uab.es

Siemens Business Services GmbH & Co. OHG
Otto-Hahn-Ring 6
81739 Munich
Germany
Contact person: Dirk Ramhorst
Tel: +49 (89)63640225; Fax: +49 89 63640233
Email: Dirk.Ramhorst@siemens.com

Executive Summary

The main objective of the implementation of the case study applications in the SEKT project is to demonstrate the capabilities of semantic knowledge technology to adopters and customers.

Application prototypes which incorporate the SEKT technology components were developed in the three case studies with a sufficient degree of maturity to allow conclusive testing with potential users under realistic conditions.

Valid and meaningful field tests were planned and conducted with the BT digital library application for BT employees (WP11) and the IURISERVICE application for legal professionals (WP 10). Disruptive organizational developments in one of the partner organisations with a key role in the WP 9 case study did not allow meaningful testing with users.

The **SEKT prototype for the BT digital library** integrates the *Squirrel* search and browse application and the *SEKTagent* application as a new front-end to the existing BT digital library, which provides access to several million of documents in a number of databases and publishers to which BT has subscriptions. Twenty BT knowledge workers tested the new components against the current search engine in use in BT's digital library. The goal was to evaluate user performance and user satisfaction.

The new SEKT semantic search and browse functions were assessed as clearly positive. In particular, the improved functionality for searching web content stands out by the very positive evaluation by subjects.

Users consistently rated the quality of information higher when using the SEKT application to conduct their information search, but they did not progress faster towards the goal of their search.

The **SEKT prototype for the IURISERVICE application** is targetted at the needs of judges in their first position. The specific conditions of work require that these judges are “on duty” for 24 hours during part of their time, when they are required to make decisions on urgent legal cases presented to them by the police. They have to make important decisions under time pressure, and often discussing open questions with senior colleagues is not possible.

A representative group of ten judges in their first position, employed in the legal system of Catalunya, and ten legal experts from the law faculty of the UAB participated in the field test. The test consisted in elaborating solutions to legal cases which are representative of the cases to which judges have to provide an immediate answer when on duty. The goal was to evaluate user performance (task solved, duration, and assessment of the quality of the decision and the legal argumentation by independent senior experts), and user satisfaction with IURISERVICE.

All subjects solved the legal cases significantly faster when using IURISERVICE, compared to the traditional manner of work.

The quality of the solution to each case was assessed by an independent senior legal expert. The result does not indicate a systematic effect of the use of IURISERVICE on the quality of the decisions and justifications of the legal experts. The subjects attain the same high level of quality in their decisions and legal argumentation.

The judges and legal experts rated the effect which they expect that IURISERVICE will have on their work very positively. They are satisfied with the usability of IURISERVICE.

The results of the tests are decisively encouraging for the implementation of semantic knowledge technology for information-intensive work. The subjective assessment of all subjects is highly positive. The subjects expect significant advantages for the efficiency and the quality of their work from the use of applications with semantic technology. The SEKT functionality in the application prototypes for the BT digital library and IURISERVICE is assessed as clearly desirable by almost all users participating in the tests.

On the basis of these results we can state with confidence that the addition of SEKT functionalitiy will be welcomed by users.

Contents

1 Introduction

The main objective of the implementation of the case study applications in the SEKT project is to demonstrate the capabilities of semantic knowledge technology. Only realistic tests with prospective users can demonstrate to adopters and customers that they can obtain significant benefits from the use of semantic technology. The results of the case study development are intended to feed back into research and generate new ideas, and also to act proactively by demonstrating to prospective adopters which benefits SEKT can provide. For this purpose, application prototypes were developed in the three case studies which incorporate the SEKT technology components, and which achieve a sufficient degree of maturity to allow conclusive testing with potential users under realistic conditions.

Important aspects of SEKT-enabled applications (such as the partial automation of annotation of information objects, and semantic search) remain hidden from users. The users see the results of their search-and-browse activities which the technology helps to produce. For users, the visible result should be that the knowledge which they require for their work is available easily, comfortably, providing more comprehensive results than the existing solutions based on traditional technology, such as keyword search and manual annotation with metadata.

Three prototypes were developed, but disruptive organizational developments in one of the partner organisations with a key role in the WP 9 case study did not allow meaningful testing with users. In two case studies valid and meaningful tests were conducted: the BT digital library for BT employees (WP11), and the IURISERVICE application for legal professionals (WP 10). The results of the final field tests, which were planned and carried out in the specific contexts of these case studies, are described in this report.

2 User-centred activities in the SEKT development process

Prospective users selected from the population of employees in the user organisation were involved in the SEKT development cycle from the start.

In the initial development phase, before the detailed specifications of the applications were completed and implementations were initiated, the needs and preferences of users were analyzed. The results were fed back to the development teams and did help to shape the applications. The results of user needs analyses also helped to define assessment criteria and measurement procedures for the field tests with realistic user groups. These results were reported in D8.2.1 User needs and opportunities for business process improvement with knowledge technology and D8.2.2 User needs v2.

One of the important conclusions was that for users the quality of the information obtained by search-and-browse, and the precision of the information presented are significant quality aspects according to which they evaluate the applications. Quality of information from the user-perspective is not sufficiently defined by recall and precision metrics in the information-retrieval sense, but must take into account the application context of the knowledge worker. Users want to be able to access information comfortably and efficiently, reducing their workload and obtaining the most relevant and useful information first. A goal was to evaluate user performance and user satisfaction with respect to these dimensions. As a consequence, appropriate approaches were conceived to measure information quality from the perspective of the user, in addition to the use of established and proven methods.

User validation plans were drawn up and updated in accordance with project progress to prepare the user tests. The full plans for each case study were reported previously in D8.3.1 Methods for user analysis in the SEKT use cases.

Early prototypes were tested by experts (using checklists and guidelines), and with the help of users (using heuristic evaluation, cognitive walkthroughs, and focus groups). The results of these tests helped to improve the user interface and the functionality of the application according to the stated needs and requests of the users and recommendations from experts.

The approach to user validation and the involvement of users was organized in three phases:

(1) user needs analysis (mainly year one of the SEKT project)

(2) usability inspection and testing (year 2 and year 3)

(3) field tests with mature prototypes (year 3)

This report describes the procedures and results of the third phase, the field tests of application prototypes.

3 Field tests of the SEKT prototype applications

The principles applied to the testing of the SEKT applications are derived from the project objectives and the specific application context. From the users perspective the question is whether the development objectives were achieved. The use of semantic knowledge technology is expected to:

- help knowledge workers find information more efficiently and effectively,
- make relevant information much easier to access,
- provide information with a higher level of quality than alternative means to access information

These variables were quantified by defining appropriate measures. Tests were conducted under realistic conditions which are representative for the future application context.

Objective and subjective measures were selected according to the definition of the success criteria of the users and of the test conditions. It should be noted that the users are professionals who normally work mostly autonomously. As part of a complex task (such as patent search, preparation of an RTD project, solving a legal case) the users determine themselves which quality criteria they apply to the search under the specific conditions, i.e. they determine which results are satisfactory. We base the analysis on subjective assessments by the test users. Although preferable in principle, it would not be a fair and meaningful test to use external criteria, which may not correspond to the criteria which the subjects apply in the test.

The main assessment criteria are:

- Task completion
- Time to complete task (if meaningful)
- Quality of the information obtained (by different measures)

The main objective is to obtain a quantitative assessment of the application in terms of benefits and added value of semantic knowledge technology. In addition, users were asked to rate the quality of the SEKT enhanced system in comparison to their traditional application. Several measures were applied to assure the validity of the results. These include:

- Rating of the specific functionality in terms of expected benefit and value
- Rating of the SEKT-enhanced application in comparison to the traditional tools which are currently in use
- Assessment of the user-perceived quality of the SEKT application with the standardised SUMI scale.

To carry out the evaluation, a systematic and controlled approach (i.e. essentially an experimental plan) was defined. This approach provides the highest reliability and validity of results, but requires experimental control of the test conditions, which means that not all aspects of the realistic context of work are reproduced. An alternative, considered earlier, would have been the collection of data in a naturalistic setting, using sophisticated online sampling procedures. For several reasons this did not turn out to be possible: Firstly the decision was to select a sample of professional users, to obtain the highest quality of data (rather than, for example, students or hired subjects). These subjects tend to be available for very restricted periods of time only, and it is rather hard to control the conditions under which they carry out the testing. It turned out to be impossible to foresee sufficient time to implement this procedure.

In addition, the limited amount of data in the databases used for the tests required that tasks, which are answerable with the specific data in the databases, had to be carefully defined and presented to the subjects.

4 Test procedures

The preparation of each of the field tests was carried out according to standard procedures. The experimental conditions were determined for each of the case studies, including the following components:

- Agreements with user organisations
- Test environment and setup
- Subjects
- Test procedure and tasks
- Instruction
- Data collection
- Results
- General observations

4.1 BT digital library prototype

The SEKT prototype for the BT digital library integrates the *Squirrel* search and browse application and the *SEKTagent* application in a new front-end to the existing BT digital library, which provides access to several million of documents in a number of databases and publishers to which BT has subscriptions. The new components were tested against the current search engine in use in BT's digital library. In the case of *Squirrel*, the aim was to assess whether the information search is more efficient in terms of information quality and time. The main objective is to determine whether the new technology helps people to obtain the information which they need in their work easier and faster.

Another important factor is the quality of the information which is obtained as a result of the search and browse process: From analyses of the user needs earlier in the project we know that it is important for users to obtain information selectively. They want the right information for their work, and they would like it to be presented in an easily accessible manner. Semantic knowledge technology should provide specific advantages in this respect: Rather than just finding more information, the information should be relevant to the context, and presented in a comfortable format.

The tests were designed to show which benefits SEKT produces in terms of efficiency and comfort, and whether the new functionality is desirable and acceptable to users.

Agreements with user organisations

The tests were carried out with users of the BT digital library in agreement with the management of the digital library.

Test environment and setup

The tests were carried out on BT premises in a separate room reserved for testing purposes. One subject at a time was instructed and tested per session, lasting around 3 hours. A PC connected to the BT network was available for subjects, very similar to their normal working environment.

A test environment, which comprised the existing BT DL search engine and the new SEKT search and browse application, was configured to give access to approximately 37,000 bibliographic records and 2,000 web documents in a limited technical domain in the telecommunications area.

Tests were carried out over a period of five weeks in November and December 2006.

Subjects

Twenty users with a wide range of experience using information search tools were invited to take part in the tests. They are representative of the user population of the BT digital library, which comprises knowledge workers from the predominantly technical areas of work in the BT research and development departments.

Procedure and Tasks

Two separate tests were conducted with each user:

(1) The key search functions of both search and browse applications (existing BT digital library tool and SEKT application) were demonstrated to the users who subsequently carried out similar procedures as those demonstrated on their own. The users were asked to assess the impact of the new SEKT-enabled functionality on their work.

(2) The subjects completed self-guided information seeking tasks using both search and browse applications.

4.1.1 Part 1: User assessment of the SEKT functionality in the BT digital library

The intention of the first part of the test was to familiarize the subjects with the main functions of each search and browse tool in a systematic and controlled manner, and at the same time to obtain assessments of the subjects in a controlled manner. A comparison was requested between the use of the new functionality and solving the tasks with the current functionality of the BT digital library.

The following functions enabled by SEKT were demonstrated to the users, practiced in examples, and then assessed by each subject:

(1) named entity recognition in the new search and browse application,

(2) navigation and browsing using the topic ontology in the new search and browse application compared to the use of controlled indexing terms for navigation and browsing using the current technology,

(3) search refinement in both applications,

(4) integration of Web content in the new search and browse application, and

(5) the *SEKTagent* semantic search agent function (compared with the use of the current information spaces implementation).

After each function had been demonstrated to the subject, and after the subject had spent some time using the functions of both applications to complete a simple search task, the subject was asked to provide an assessment of this specific functionality, as used in this task.

The following questions were asked. Answers were given as a rating on a scale with 4 or 5 values.

- Do you expect to find information faster with the new search and browse application in comparison with the current search engine?
- Does the information search task become easier to complete with the new search and browse application in comparison with the current search engine?
- Do you expect to find better quality information with the new or the current technology (where the relevance of results was to be taken as the main measure of quality)?
- Does the new function offer an improvement compared to functions available for solving the task in the current search engine?

A rating scale with 4 values was used for questions where we wanted to force the subjects to give decisive answers.

Results

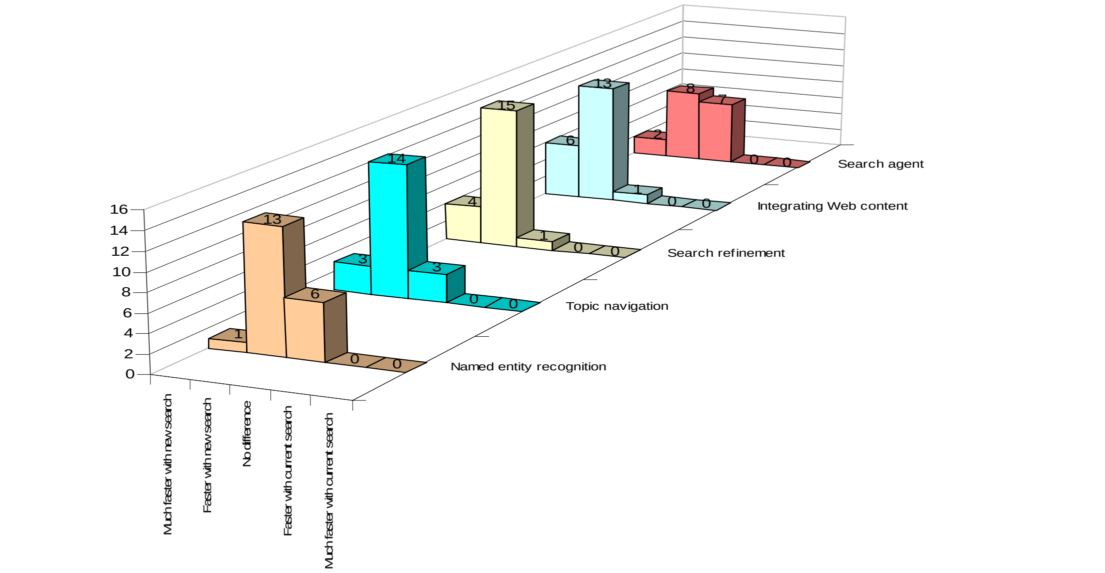
Twenty subjects completed the tests. All data are included in the analysis. A small number of data points are missing (subjects forgetting to answer questions).

Assessment of the SEKT functionality

The results of the assessment of the SEKT functionality are shown in Figure 1 to Figure 4. Twenty subjects answered the questions, the frequency of answers to each question is shown. Overall, there was a clearly positive response to the new functions provided in the semantic search and browse application. The almost complete absence of negative assessments of the new functionality provided by the SEKT prototype should be noted.

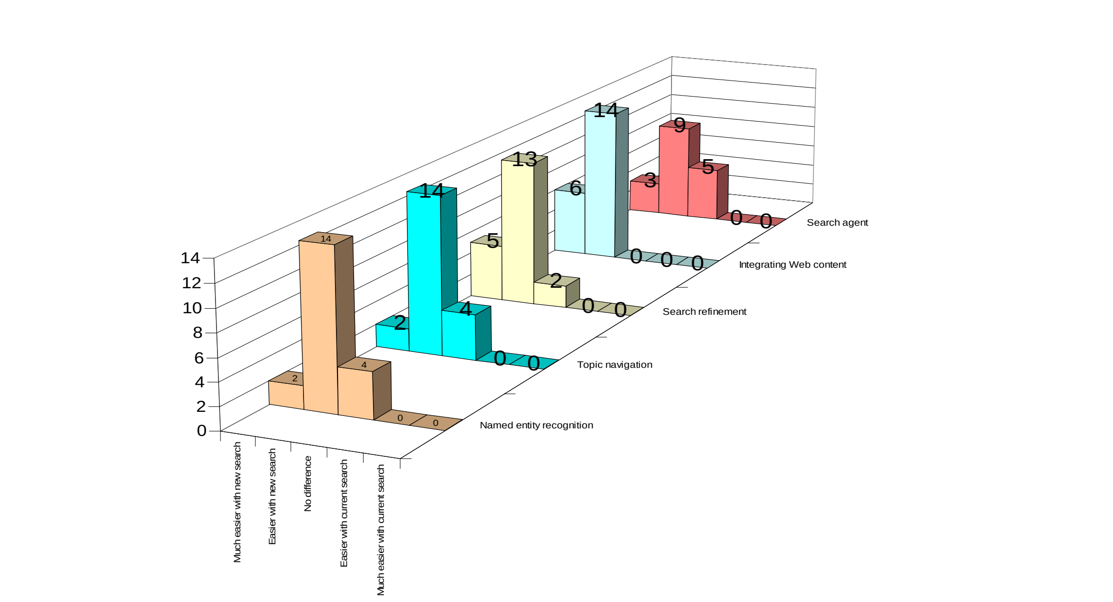
The improved functionality for topic navigation, search refinement, and for searching web content stands out by the very positive assessment by subjects.

Figure 1: Do you expect to find information faster with the new search and browse application?



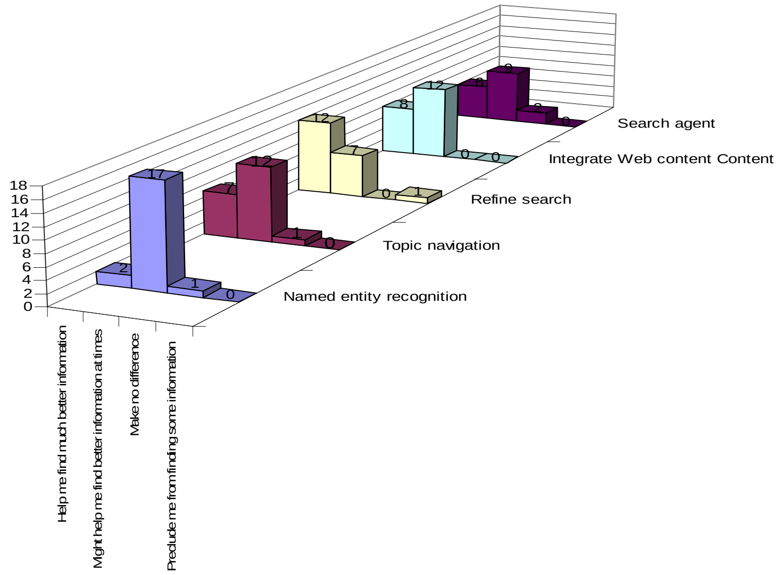
	Much faster with new search	Faster with new search	No difference	Faster with current search	Much faster with current search
□ Named entity recognition	1	13	6	0	0
□ Topic navigation	3	14	3	0	0
□ Search refinement	4	15	1	0	0
□ Integrating Web content	6	13	1	0	0
□ Search agent	2	8	7	0	0

Figure 2: Does the information search task become easier to complete with the new search and browse application?



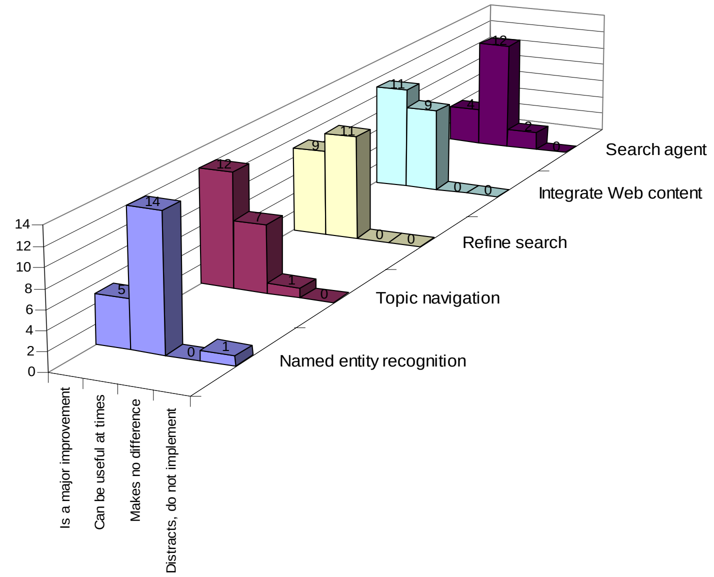
	Much easier with new search	Easier with new search	No difference	Easier with current search	Much easier with current search
□ Named entity recognition	2	14	4	0	0
□ Topic navigation	2	14	4	0	0
□ Search refinement	5	13	2	0	0
□ Integrating Web content	6	14	0	0	0
□ Search agent	3	9	5	0	0

Figure 3: Do you expect to find better quality information with the new application?



	Help me find much better information	Might help me find better information at times	Make no difference	Preclude me from finding some information
■ Named entity recognition	2	17	1	0
■ Topic navigation	7	12	1	0
□ Refine search	12	7	0	1
□ Integrate Web content Content	8	12	0	0
■ Search agent	6	9	2	0

Figure 4: Does the new function offer an improvement compared to functions available for solving the task in the current search?



	Is a major improvement	Can be useful at times	Makes no difference	Distracts, do not implement
■ Named entity recognition	5	14	0	1
■ Topic navigation	12	7	1	0
□ Refine search	9	11	0	0
□ Integrate Web content	11	9	0	0
■ Search agent	4	12	2	0

4.1.2 Part 2: Self-paced search and browse task

For users the value of a search and browse application is determined by the impact on the effectiveness of their search process. The impact may be that the search process takes less time, or that it leads to higher quality of information, or both. In appendix 8.1 we explain the measurement procedure which we have developed to capture this aspect quantitatively.

For the purpose of this study, we have employed the approach which uses a quality assessment of the information obtained in the search process by the subjects themselves. This seems most appropriate in this instance because the professionals use the digital library to search information for a specific purpose with which they are familiar, and where they determine the quality goals of their search and browse activity themselves.

Users are motivated to carry out the search task efficiently, i.e. successfully and timely, and without undue cognitive effort. How these factors are weighted against each other is determined by the person conducting the search. She may decide to be satisfied with a minimum level of quality and use less time, or aim for a higher quality. It is even possible that time and quality criteria remain the same level, but that the subject aims to reduce the cognitive effort expended.

We obtained assessments by the subjects at points in time during the search process which the subjects were asked to choose themselves. They were asked to state what the quality of the information that they had obtained up to that moment is, and how far they assume to have progressed in their information search process (relative both to the start and the expected time to completion). A 7-point scale was used for information quality and a 15-point scale for progress in the search process, and the time was recorded as well.

This measure is expected to show us whether the subjects obtain higher quality information (according to their own assessment), and whether they approach the search result fast. The end is when they decide that they have obtained sufficient information, or that further search does not give more information of value.

We expect to see in the data whether SEKT in comparison with traditional search and browse tools improves speed of the search process, or information quality, or both.

Tasks

Each user was asked to complete two similar, but separate, information seeking tasks from a set of six. One task was completed using the semantically enabled search and browse application, the other was completed using the search and browse application currently in use in BT's digital library. The order in which a subject used the applications was varied randomly, so that half of the tasks were performed using the keyword-based search engine first, and the other half using the semantically enabled search engine first. An example of the tasks used is given below.

Task description. The database contains a number of documents related to the topic of new telecommunications services. We are particularly interested in the use of telecommunications to support health services, e.g. health monitoring and wellbeing. Please take some time to find a set of documents that are relevant to this subject. The set of documents should provide a reasonable overview of the state of technical developments and services in this field. The list should be as concise as possible, yet still cover the field to a reasonable extent (e.g. find a list of approximately 10 documents that you consider most relevant). Assume that the documents you select will serve as an introduction to the field that you propose as a reading list to other technical and managerial experts. The managerial experts are also interested in finding out additional information about companies and organisations (e.g. Universities) working in this field. Please try to find information of this type as you search.

The users were presented with rating scales on paper, and were asked to assess the quality of the results returned from the application as they completed their search, and secondly how far they considered to have progressed in their search. Assessments of information quality were given by users at points in time that they determined themselves, i.e. subjects were asked to give assessments at meaningful points in their search, e.g. the moment before they submit a modified query, rather than at set time intervals. This was done in order to minimize interruptions during the search process.

Results

There is no significant difference in time between the users' search tasks with the two search engines.

SEKT search engine	Old BT dl search engine
16.7 minutes	14.7 minutes
SD = 4.71	SD = 4.83

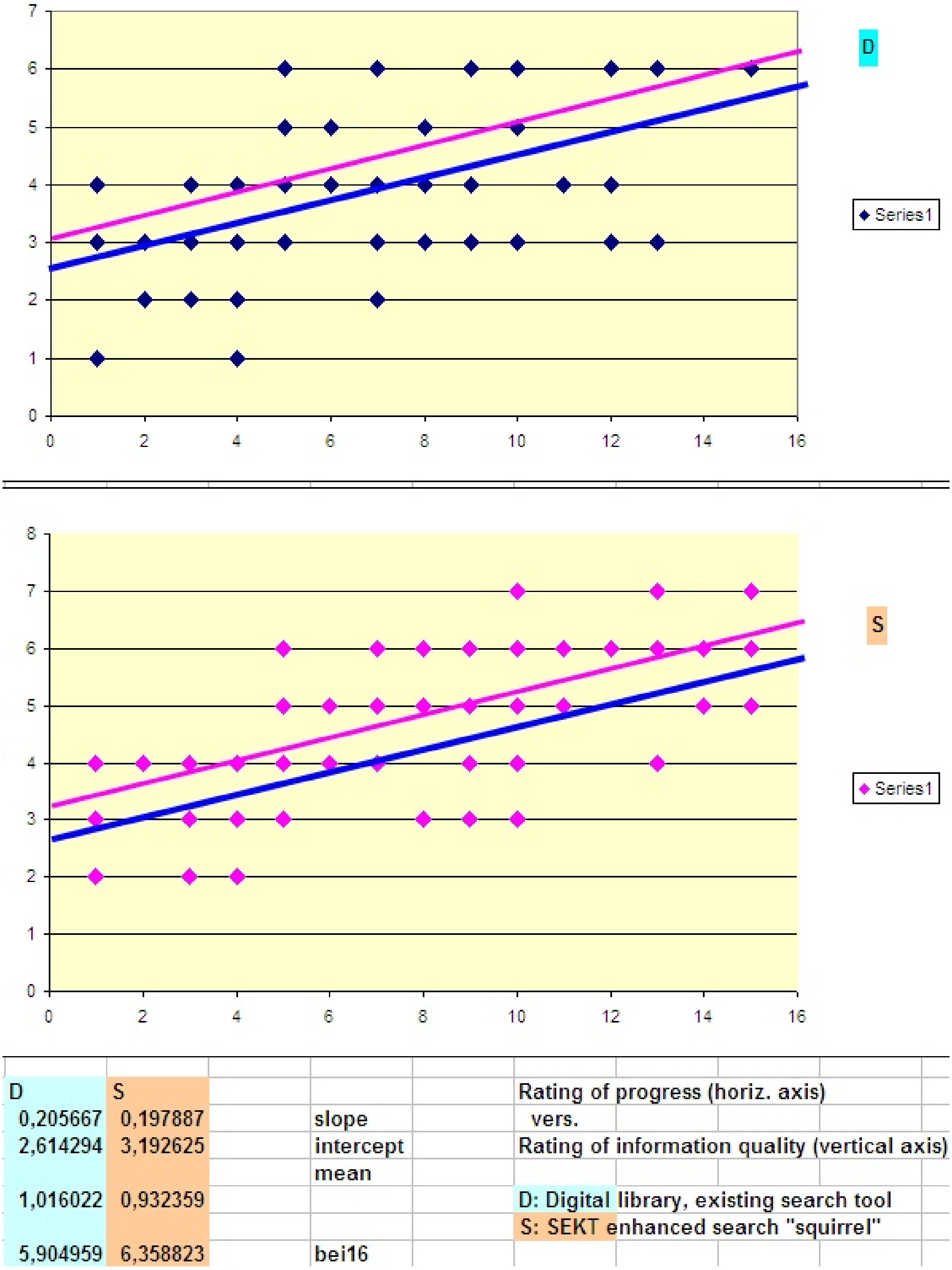
Users gave an average of 4.6 assessments per search task, i.e. about every 3 minutes. To the extent that the subjects followed the instruction this means that one new search cycle (with a reformulated query) was carried out about every 3 minutes.

The average rating of Information Quality using the existing library system was 3.99 against an average PIQ of 4.47 using the semantically enabled search and browse application. The sign test was applied to the data, comparing the average values for each subject. It shows the difference to be significant ($p<0.01$).

For each subject the average value for information quality was compared for both experimental conditions (with SEKT or with the old search engine). The rating for information quality is significantly higher under the search condition with SEKT (sign test, two-sided, $p<0.01$).

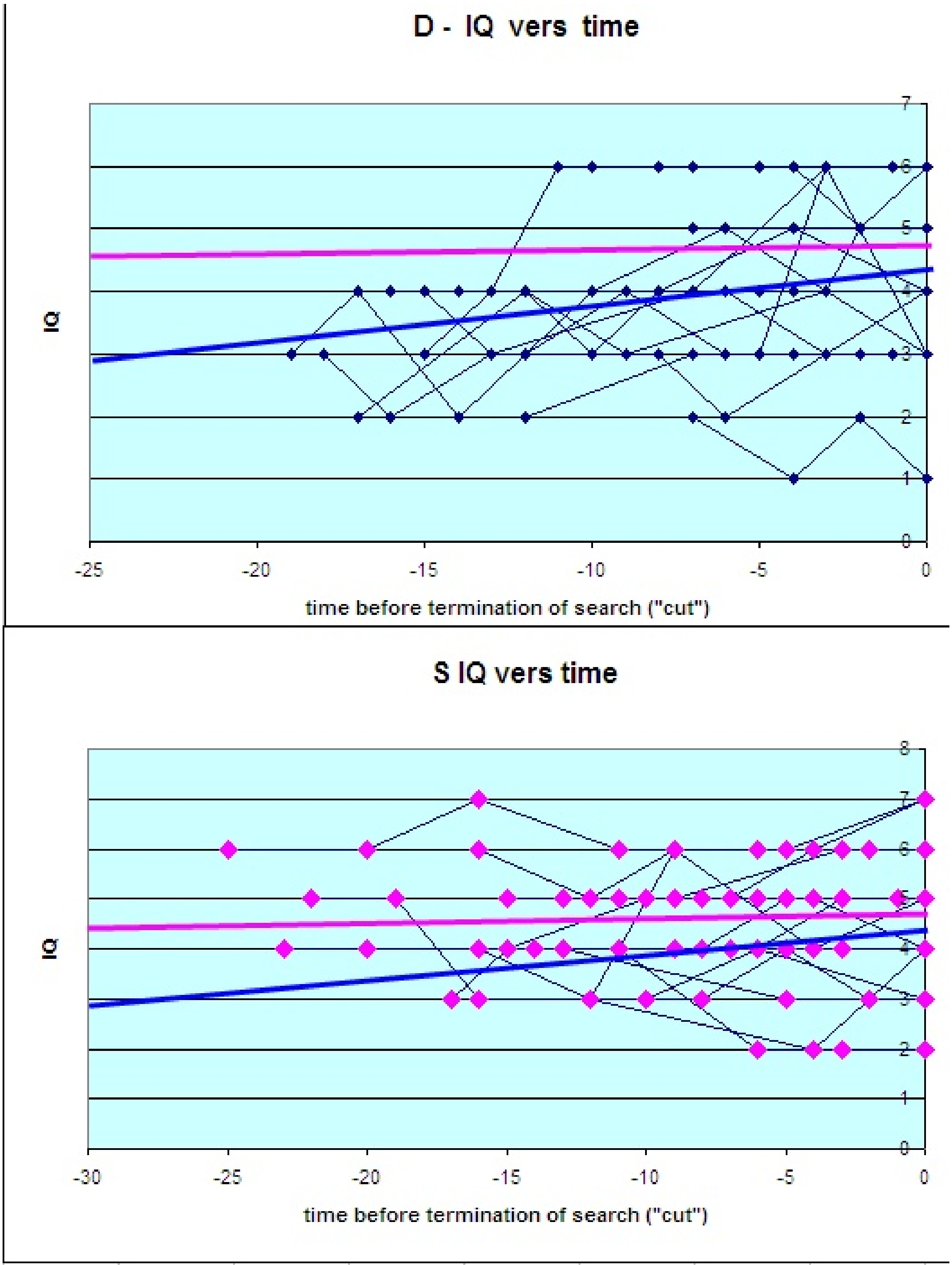
Information quality is plotted against progress of search. Progress is plotted retrospectively: When users have found sufficient information they terminate the search. The plot shows how subjects progress towards their goal.

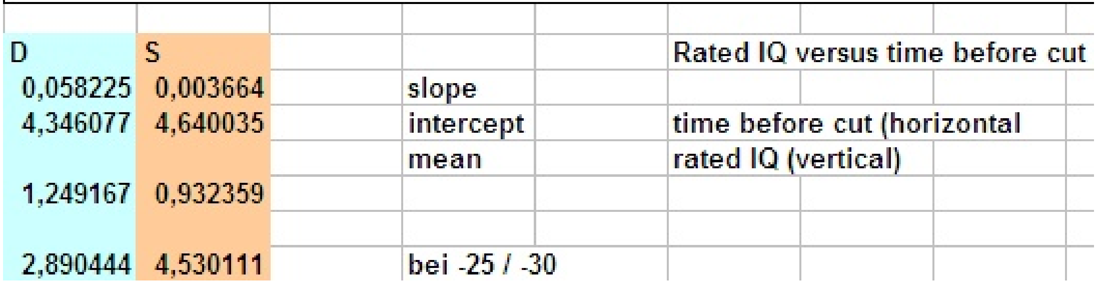
Because the duration of the search is not fixed, an anchor point on the time-axis is needed to compare the data from different persons. The anchor point could be the start or the end-point of search. We use the end point of search, i.e. the point in time when each subject terminates the search (the “cut”) as anchor point. The information state at the end point is defined more precisely, while in the starting condition a variable state of prior knowledge may exist, which we do not know.



D	S				
0,205667	0,197887		slope		Rating of progress (horiz. axis) vers.
2,614294	3,192625		intercept		Rating of information quality (vertical axis)
			mean		
1,016022	0,932359				D: Digital library, existing search tool
					S: SEKT enhanced search "squirrel"
5,904959	6,358823		bei16		

Figure 5: Information quality against progress





D	S				Rated IQ versus time before cut
0,058225	0,003664		slope		
4,346077	4,640035		intercept		time before cut (horizontal
			mean		rated IQ (vertical)
1,249167	0,932359				
2,890444	4,530111		bei -25 / -30		

Figure 6: Information quality as a function of time before search is terminated.
D – using the existing digital library search interface
S – using the SEKT search and browse interface

The plot of information quality versus rated progress (Figure 5) in the search progress shows that information quality (averaged over 20 subjects) is higher throughout the search progress. The slope of the linear regression lines calculated for the two conditions does not differ significantly for both experimental conditions (“old” digital library system vers. SEKT enabled search and browse application).

The results show that users consistently rated the quality of information higher when using the SEKT tools to conduct their information search, but they did not progress faster towards the goal of their search. This is also confirmed by looking at information quality as a function of time before search is terminated (Figure 6).

The advantage of the SEKT application in terms of information quality is visible in the IQ versus time plot as in the other views of the same data. Although this is not statistically significant, there is a tendency that information quality is not just higher with the SEKT application throughout the search process, but that it also starts at a higher level. This would mean that the SEKT application delivers higher quality information right from the first query.

The results show that SEKT delivers information with a higher quality for users. Under the conditions of this study the users did not terminate search earlier with SEKT, but continued the search. When comparable tests would be carried out under time pressure, we could expect that users prefer to terminate search earlier with satisfactory information quality, rather than aiming for higher information quality.

4.1.3 Evaluation of the search and browse application with SUMI

After the two tests were completed, the subjects were assumed to be familiar with SEKT search and browse. To conclude the test, the Software Usability Measurement Inventory (SUMI) was administered to each subject. SUMI gives a detailed view of the subjective assessment of the usability of the SEKT semantically enabled search and browse application,.

SUMI measures five independent factors of user satisfaction:

- Efficiency refers to the user's feeling that the software enables them to perform their tasks in a quick, effective and economical manner.
- Affect refers to the positive user feeling of the user being mentally stimulated and pleased as a result of interacting with the software.
- Helpfulness refers to the user's perceptions that the software communicates in a helpful way and assists in the resolution of operational problems.
- Control refers to the feeling that the software responds in an expected and consistent way to input and commands.
- Learnability refers to the feeling that the user has that it is relatively straightforward to become familiar with the software.

The result of the SUMI analysis is shown in Figure 7. (One subject did not complete one page of the questionnaire. The data had to be excluded.)

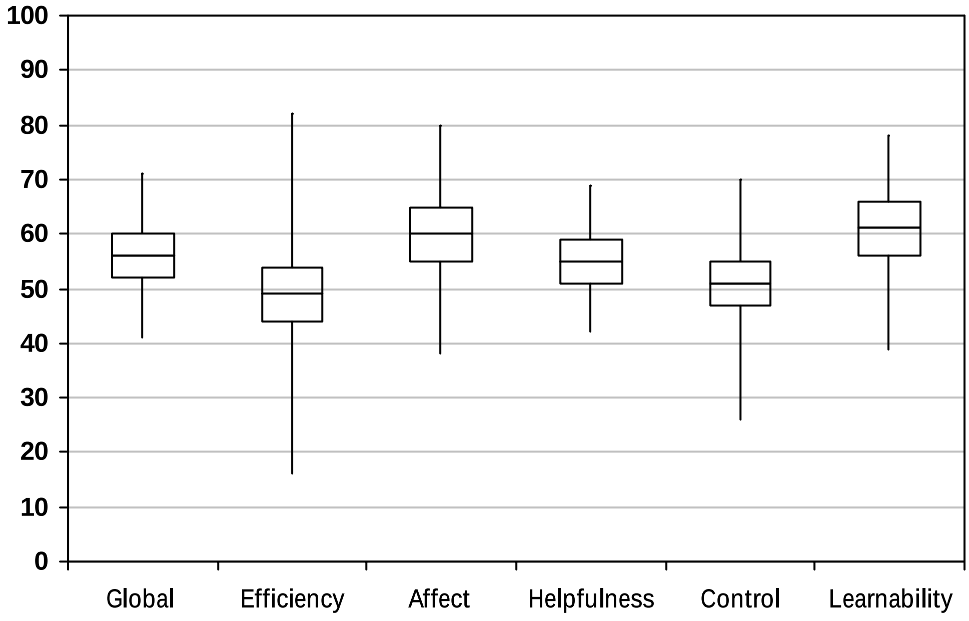


Figure 7: Results of the SUMI profile analysis for the SEKT prototype of the BT DL, 19 subjects. The graph shows medians, upper and lower 95% confidence intervals and upper and lower limits of the data distributions.

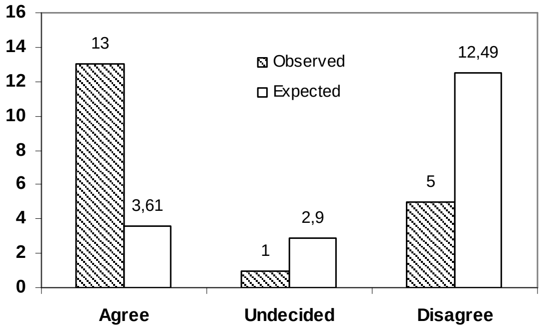
SUMI Profile Analysis for the SEKT prototype of the BT DL, 19 subjects					
Usability Scales	**Upper Limit**	**Upper 95% Confidence Limit**	**Median**	**Lower 95% confidence Limit**	**Lower Limit**
Global	71	60	56	52	41
Efficiency	82	54	49	44	16
Affect	80	65	60	55	38
Helpfulness	69	59	55	51	42
Control	70	55	51	47	26
Learnability	78	66	61	56	39

Results show that the overall assessment of users is positive. They find the application easy to learn, and see it as pleasing. The efficiency of work with SEKT, and the ability to exercise control over the application are just seen as average.

The Goodness of Fit (by Item Consensual Analysis) between the observed and expected answers to the 50 SUMI questions was analyzed using Chi Square statistics. The results below show where the nineteen subjects made pronounced statements about the application under analysis, which illustrates in more detail on which judgements the test users agree strongly.

Efficiency

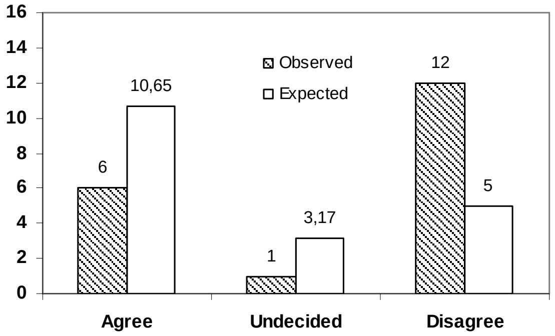
More users than expected **agree** that the SEKT DL responds too slowly to inputs (99,99 % confidence).



This software responds too slowly to inputs.
(Chi Square value = 30,18)

Control

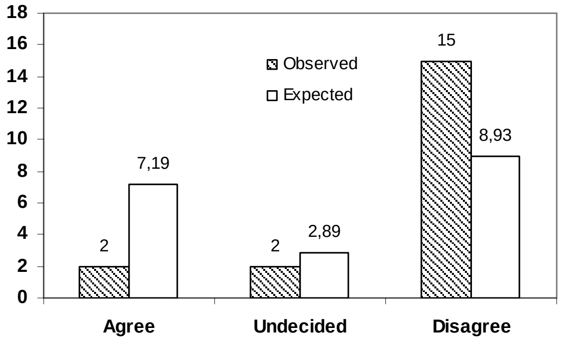
More users than expected **disagree** that the speed of the SEKT DL is fast enough (99 % confidence).



The speed of this software is fast enough.
(Chi Square value = 12,51)

Affect

More users than expected **disagree** that there have been times in using the SEKT DL when they have felt quite tense (95 %).



There have been times in using this software when I have felt quite tense.
(Chi Square value = 8,15)

The users who completed the SUMI assessment are reasonably satisfied with the usability of the Squirrel search and browse application (Global usability was rated 4 scale points higher than the average). Users liked using *Squirrel* (Affect was rated 8 scale points higher than the average), but do not seem to find the presentation of the user interface particularly attractive. The SEKT DL is considered easy to use (Learnability was rated 11 points higher than the average and Helpfulness was rated 4 points higher than the average). Control of *Squirrel*, the feeling that the software is responding in an expected and consistent way, is considered average (Control was rated 1 point below the average). The majority of users considered Squirrel to be no more efficient than other comparable software systems (Efficiency was rated 4 scale points below the average). Users found the response to inputs, i.e. the time to display a set of results, too slow.

Overall, users rate *Squirrel* positively and believe that it has attractive properties, but are also unhappy about some properties, especially performance and speed.

[Note: The SUMI test is scaled to a mean of 50, while 10 scale points correspond to one standard deviation.]

4.2 IURISERVICE application

The IURISERVICE application is targetted at the needs of a specific user group, judges in their first position. The specific conditions of work require that these judges are "on duty" for 24 hours during part of their time, when they are required to make decisions on urgent legal cases presented to them, for example by the police. They have to make important decisions under time pressure, and often they cannot resort to discuss cases with senior colleagues. It is in this case especially important to obtain the participation of a representative group of test users from this population of future users. A main constraint was the need to design the tests in such a way that judges were able to participate in the tests.

Agreements with user organisations

IURISERVICE in its initial form is designed to meet the requirements of judges in their first position. Therefore it is most important to assure the participation of these users in tests of the system. It was necessary to agree with the Judicial Authorities to authorize the participation of these judges in the SEKT project. This was achieved after extensive negotiations, and is considered a major advantage for the project.

Test location, environment and conditions

IURISERVICE will be available at the location of work of judges. For the purpose of the tests a number of workplaces were made available in the Escuela Judicial in Barcelona, and in the Faculta de Dret of the UAB, where the tests took place. It was assured that IURISERVICE was operating efficiently on all workplaces used. Because IURISERVICE at this stage of implementation contains a limited amout of legal documentation only, the tests were designed to relate to the legal domains which are represented in the database and covered by the ontology as available at this stage.

Time

The tests were carried out between January 29, 2007, and February 16, 2007.

Subjects

The subjects were selected from two groups:

- judges in their first position, employed in the legal system of Catalunya,
- legal experts from the law faculty of the UAB.

The subjects were invited personally to participate in the field tests. The judges participated in two sessions of one half day each, and the legal experts in one session with two separate parts. A total of 10 judges and 10 legal experts from the UAB participated. Subjects were compensated financially for their time.

Tasks

The test consisted in elaborating solutions to legal cases which are representative of the cases to which judges have to provide an immediate answer when on duty. It was also assured by prior tests that sufficient relevant information of the jurisdiction on each one of these cases is contained in the database.

Procedure and instruction of the subjects

The subjects were firstly informed about the purpose of the test and made familiar with the test environment. The first half of the tasks were carried out using the normal environment of work, i.e. without the use of IURISERVICE, but with availability of all sources of information accessible to the judges in their work, among these the legal databases LaLey, Aranzadi, El Derecho and also Google. Each subject received a package of documents on paper containing all information about the first case, including forms and questionnaires to record results. According to their normal working procedures, the subjects recorded their decisions and justification on paper or used a text editor. After completing a case, the documents were collected, and the subjects were handed out a package with the next case. All subjects solved the same six cases. The cases were distributed in a randomized order to the judges. The first case on each day was one of two cases considered easier than the others, serving as a “warm up” to familiarize the subjects with the entire process. When three cases were completed on the first day, the subjects were allowed to leave.

The second phase of the test was conducted on the following day for the judges. On the second day of testing the judges were introduced to IURISERVICE, including one worked example. Then all judges worked through one case on their own with IURISERVICE. Following this, the judges solved three cases, randomly selected from the same cases as in the first part of the test. Thus, each subject solved all six cases, but in a different order. After completing the tests, each subject was interviewed individually.

The procedure for the legal experts was slightly modified: The subjects solved two cases on each day only, the first “warm up” task was excluded, otherwise the tasks were drawn from the same sample of cases as those for the judges. The subjects were instructed individually. They carried out the first half of the test (without IURISERVICE), took a break, and then were introduced to IURISERVICE and solved two further test cases.

One aspect should be noted: A fully controlled experiment would have required a further control group which did solve the second part of the test without IURISERVICE. This was not possible for two reasons: Firstly a considerably larger number of subjects would have been required, and secondly it must be taken into account that the subjects are highly qualified professionals, who are motivated by participating in a meaningful test. It was not considered possible to ask them to carry out a seemingly not meaningful task (which the control task would have been).

Data collection and analysis

The data collected were:

- Performance: Task solved, duration, and assessment of the quality of the decision and the legal argumentation by independent senior experts.
- Assessment by the subjects: After completion of the second part of the tests subjects were asked to assess IURISERVICE in comparison to their normal way of working. SUMI was administered with the request to assess IURISERVICE.
- Interviews were carried out individually with each subject.

Results

Performance: Time to provide solutions to legal cases

All subjects were able to solve all tasks, both with and without IURISERVICE. The comparison of the time taken for each task (providing a solution for a legal case) shows a significant (sign test, $p<0.01$, two-dided test) difference: All subjects did solve two cases much faster with the use of IURISERVICE.

Mean time to solve cases with and without IURISERVICE

Data of 9 judges and legal experts each without IURISERVICE (data missing from one subject in each group). Data of 10 judges and legal experts each with IURISERVICE.

	Judges	**Legal Experts**
With IURISERVICE	9 minutes per case	5 minutes per case
Without IURISERVICE	23 minutes per case	20 minutes per case

Quality of the solution

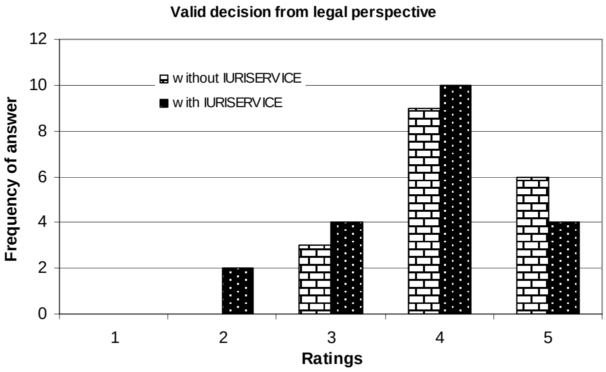
The quality of the solution to each case was assessed by an independent senior legal expert. The criteria of assessment were

- Is the decision valid from a legal perspective?
- The soundness of the decision
- The richness of the argumentation
- The number of legal sources used

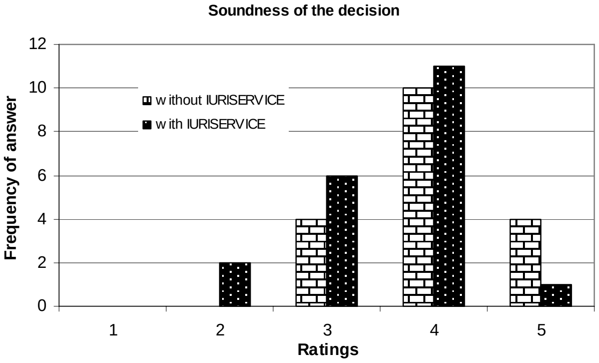
(1 = not satisfactory, 5 = excellent)

Results of rating the quality of the solution

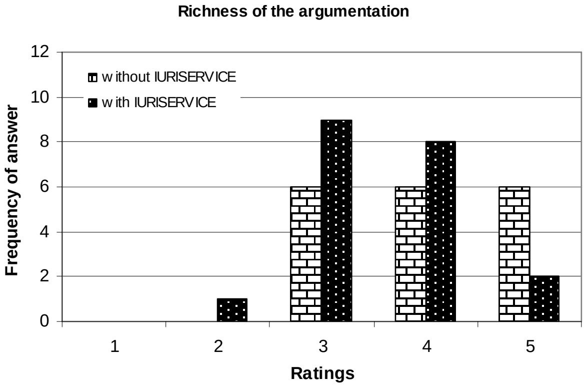
Valid decision from a legal perspective



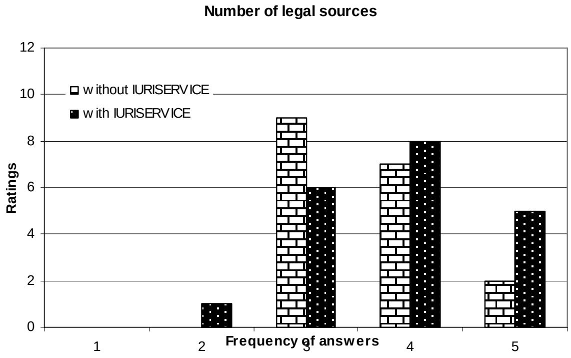
Soundness of the decision



Richness of the argumentation



Number of legal sources



[The ratings of the legal experts’ quality of solution are in progress and not yet available.]

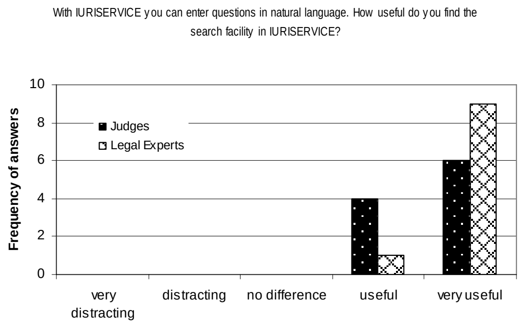
The data do not indicate a systematic effect of the use of IURISERVICE on the quality of the decisions and justifications of the legal experts. The subjects attain the same high level of quality in their decisions and legal argumentation.

Assessments of IURISERVICE by the subjects

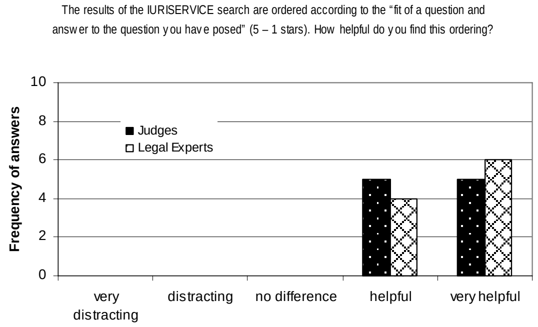
The subjects rated which effect they expect that IURISERVICE will have on their on their work. Five aspects of IURISERVICE, on a 5-value scale. The rating is clearly positive by all subjects.

Results of the assessment of IURISERVICE by the subjects

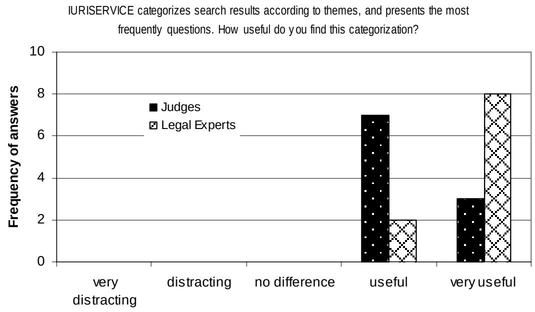
Both, judges and legal experts, find the search facility in IURISERVICE which allows them to enter questions in **natural language (**very) useful.



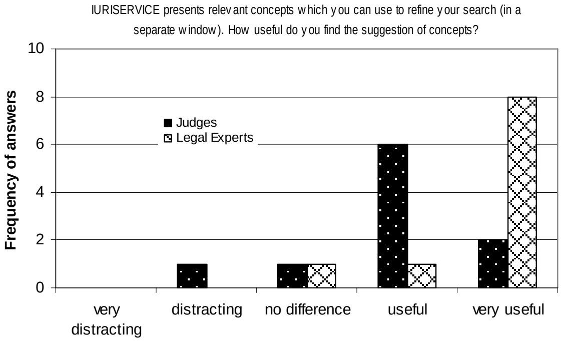
Both, judges and legal experts, find the **ordering of results according to the fit** of the question to the question they have posed helpful or very helpful.



Both, judges and legal experts, find the categorization of the most frequent results according to **themes** helpful or very helpful.



Most of the judges and legal experts find the **suggestion of concepts to refine the search** useful or very useful. Only one judge found this aspect distracting and another judge and one legal expert each did not see a difference compared to using other databases.



Assessment of IURISERVICE with SUMI

SUMI is a validated and proven questionnaire which measures the satisfaction of users with a software application on five empirically identified factors, which are described as follows:

- “The Affect subscale measures the user's general emotional reaction to the software - it may be glossed as Likeability.
- Efficiency measures the degree to which users feel that the software assists them in their work and is related to the concept of transparency.
- Helpfulness measures the degree to which the software is self-explanatory, as well as more specific things like the adequacy of help facilities and documentation.
- The Control dimensions measures the extent to which the user feels in control of the software, as opposed to being controlled by the software, when carrying out the task.
- Learnability, finally, measures the speed and facility with which the user feels that they have been able to master the system, or to learn how to use new features when necessary.”

SUMI is scaled such that the mean of the scale is 50, and the standard deviation is 10.

The results shown in figures 8 and 9 for the two groups of subjects (judges and legal experts) show a highly positive assessment. The high degree of consistency of the two groups tested independently lends a high degree of reliability to these results. The factors which stand as highly positive are Affect and Efficiency.

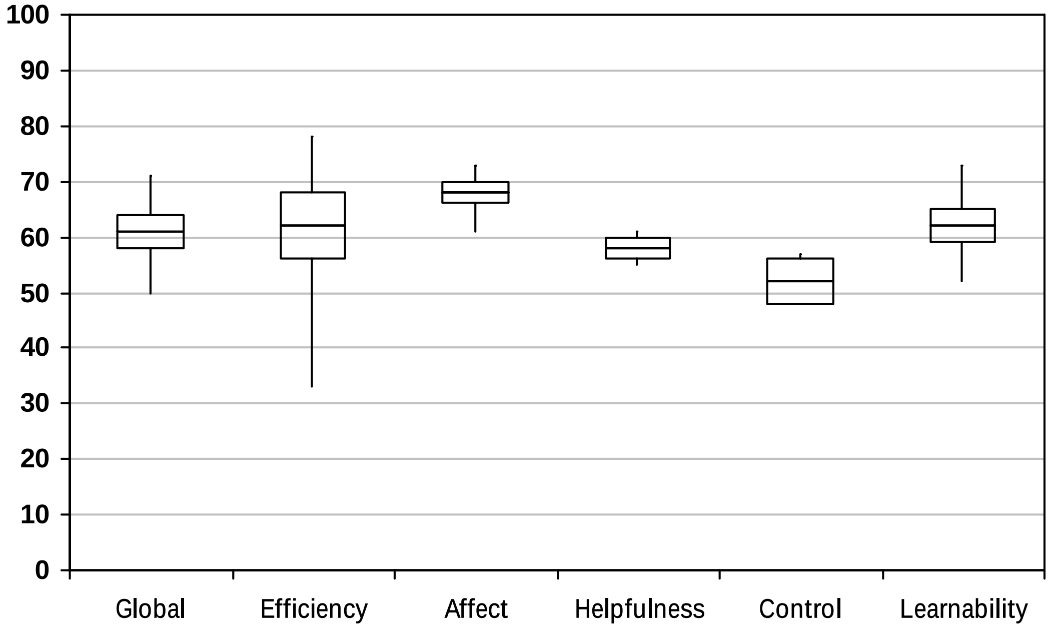


Figure 8: Results of the SUMI profile analysis of the SEKT prototype for IURISERVICE, 9 judges. The graph shows the median, the upper and lower 95% confidence intervals, and the upper and lower fences.

SUMI Profile Analysis for the SEKT prototype for IURISERVICE, 9 judges					
Usability Scales	**Upper Fence**	**Upper 95% Confidence Limit**	**Median**	**Lower 95% Confidence Limit**	**Lower Fence**
Global	71	64	61	58	50
Efficiency	78	68	62	56	33
Affect	73	70	68	66	61
Helpfulness	61	60	58	56	55
Control	57	56	52	48	48
Learnability	73	65	62	59	52

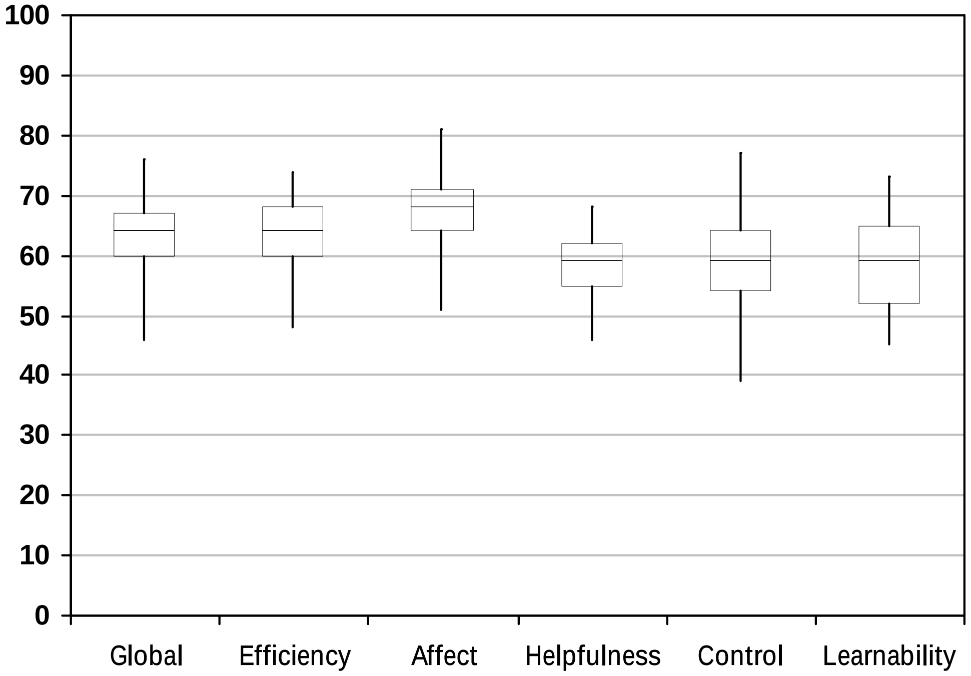


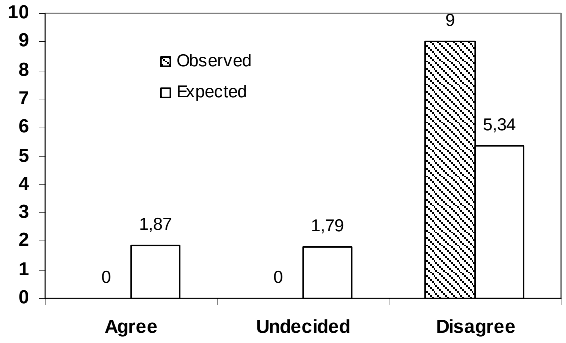
Figure 9: Results of the SUMI profile analysis for the SEKT prototpye for IURISERVICE, 10 legal experts. The graph shows the median, the upper and lower 95% confidence intervals, and the upper and lower fences.

SUMI Profile Analysis for the SEKT prototype for IURISERVICE, 10 legal experts					
Usability Scales	**Upper Fence**	**Upper 95% Confidence Limit**	**Median**	**Lower 95% Confidence Limit**	**Lower Fence**
Global	76	67	64	60	46
Efficiency	74	68	64	60	48
Affect	81	71	68	64	51
Helpfulness	68	62	59	55	46
Control	77	64	59	54	39
Learnability	73	65	59	52	45

The Goodness of Fit (by Item Consensual Analysis) between the observed and expected answers to the 50 SUMI questions was analyzed using Chi Square statistics. The results SUMI Item Consensual Analysis for IURISERVICE, 10 judges and 10 legal experts, show where the subjects make pronounced statements about the application under analysis.

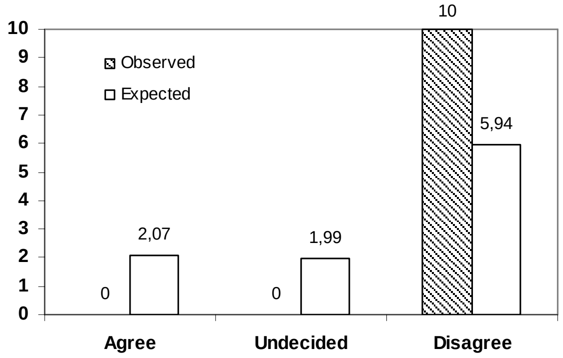
Efficiency

More **judges** and than expected **disagree** that "there are too many steps required to get something to work" (95 % confidence).



Item 36: There are too many steps to et something to work (Chi Square value = 6,16)

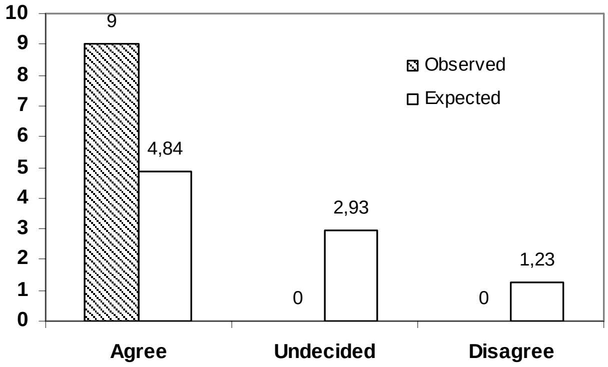
More **legal experts** than expected **disagree** that "there are too many steps required to get someting to work" (95 % confidence).



Item 36: There are too many steps to et something to work (Chi Square value = 6,84)

Affect

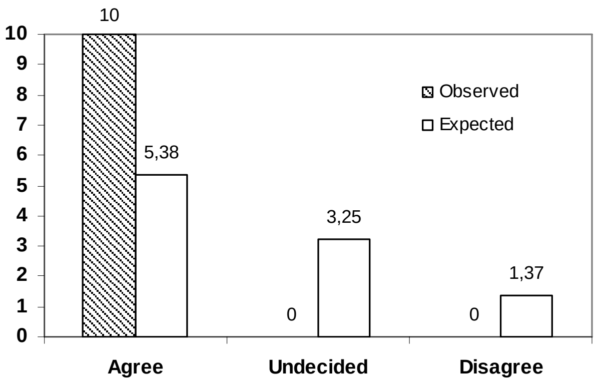
More **judges** than expected agree that "working with IURISERVICE is satisfying" (95 % confidence).



Item 12: Working with this software is satisfying (Chi Square value = 7,74)

Affect

More **legal experts** than expected agree that "working with IURISERVICE is satisfying" (95 % confidence).



Item 12: Working with this software is satisfying (Chi Square value = 8,6)

The judges and legal experts who completed the SUMI assessment are satisfied with the usability of IURISERVICE (Global usability was rated 11 scale points higher than the average by judges, and 14 scale points higher by legal experts).

Both, judges and legal experts, liked using IURISERVICE (Affect was rated 14 scale points higher than average by both groups of users). IURISERVICE is considered efficient to use (Efficiency was rated 12 scale points higher than average by judges and 14 scale points higher by legal experts).

IURISERVICE is considered easy to use (Learnability was rated 12 scale points higher than average by judges and 9 scale points higher by legal experts; Helpfulness was rated 8 scale points higher than average by judges, 9 scale points higher by legal experts). Control, the feeling that the IURISERVICE is responding in an expected and consistent way is considered average by the judges (Control was rated 2 scale points higher than average) and better than average by the legal experts (who rated Control 9 scale points higher than average).

Observations made by subjects

After the conclusion of the tests the subjects were interviewed informally and asked about their further observations and opinions on IURISERVICE. The important observations were:

Comments referring to a single case which had just been solved

Subject 1: "The way I wrote down the question in Iuriservice was almost the same in which the question appeared in the system, so the exercise was easy to solve."

Subject 3: "In this question I have found the solution of the problem very quickly. I wish IURISERVICE will be available in the Courts pretty soon!"

Subject 4: "Iuriservice offers more quickness and more precise answers to the case solving."

Subject 5: “The system happened to have stored the very same question I posed.”

Subject 6: “I found the answer with IURISERVICE (the question was very similar to the one I posed), but I just used a part of this answer to solve the case.”

Subject 9: “In this case, the database has worked better than in the previous one, because it has offered me the answer even though I posed a question in different words than those which appear in the question stored in the system.”

Subject 9: (in the last case) “The database has worked very well again.”

“What is the aspect of Iuriservice which you find most attractive?”

Subject 1: “That you can write down the questions in the very way you think them and that those questions are referred to practical questions that have not any answer in the law (in some occasions).”

Subject 5: “That there are related questions and the extended answers.”

Subject 6: “The use of natural language.”

Subject 8: “That you may find quick solutions to urgent problems.”

Subject 9: “The cases the system tries to solve are very well selected because they are very frequent and they usually appear during the on-duty period (when one has not time to find the solution at all).”

Subject 10: “The aesthetics of the system and that you can pose complex questions in an easy way.”

Results of the evaluation of IURISERVICE

The prospective users assess IURISERVICE in comparison to their traditional way of working (using legal databases) as highly positive. They see considerable benefits in the functionality which is added by SEKT, and they find the system easy to learn and pleasant to use. This is shown both by the questions posed after the tasks were carried out, and by the results of the standardised SUMI questionnaire.

The performance of the IURISERVICE shows that significant performance advantages can be gained by implementing SEKT functionality for specific user groups and their needs.

5 Assessment of the potential business benefit of SEKT

The business benefit of SEKT is estimated by comparing the cost of a SEKT implementation to the benefits generated by SEKT for the users and the user organisation.

Cost of SEKT

The cost of knowledge management system implementations was estimated by brainstorming with a number of experts (from SEKT partners SIRMA, empolis, Siemens SBS, and iSOCO) with a background in estimating the cost of large scale knowledge management implementations on the basis of commercial project experience. Estimates were produced individually and separately, discussed, and ranges considered reasonable for the characteristic cost factors determined. The details of the estimation procedure was discussed, but the underlying detailed parameters are not publishable due to their commercial character.

The cost parameter which we tried to estimate was total cost per user per year, based on a five year life span of the system implementation (annual cost per user), which was determined for organisations of different size. The size was considered either for organisations of 3000 users or 30000 users. For smaller organisations insufficient relevant experience existed to come to a reliable estimate, but the common view was that the cost will be somewhat higher. (Not surprising in view of the rationale of knowledge management, which is to exploit the economy of scale.)

Knowledge-management (KM) systems using traditional technologies and semantic technology compete. The case considered is the implementation of a new KM system. Under most conditions a varying degree of transfer of resources from a legacy system would be expected, but this can only be estimated for a concrete case. The cost for licensing and configuration is not considered to be significantly different for a knowledge management based on SEKT per se. The main difference is due to the comparative internal effort required in the user organisation to maintain the organisation-specific taxonomy or ontology. This accounts for roughly 50% of the cost of the KM system, the estimates were ranging from 30% to 60%.

Assessment of the potential cost of SEKT implementation		
	3000 users	30.000 users
Total lifecycle cost over a 5-year lifetime (€ per user per year)	50 30 - 100	35

The differential between a taxonomy based KM system (which may be a bare bones system, but could also have extended functionality) and an ontology based system was estimated to range between 0 and 20 Euro. An application integrating SEKT would include more functionality, so a comparison of strictly equivalent systems would not be possible.

Benefits

The benefits which SEKT can produce are

- Reduced task time
- Satisfaction (feel good factor) motivation
- Reduced workload and stress
- Higher quality of the task execution

Which of the results from the SEKT evaluation support that these effects can be expected?

Reduced task time		Users use considerably less time when executing tasks with the use of IURISERVICE. This would be the most direct business benefit.
Satisfaction (feel good factor), motivation		All users find the applications including SEKT helpful and pleasant. The advantage is higher motivation of the users for obtaining maximum benefit from applications.
Reduced workload and stress		Users rated the SEKT applications positive with respect to workload and learning. This is an indirect benefit which would be expected to lead to a higher rate of usage of the SEKT application.
Higher quality of the task execution		Users stated clearly that they expect to obtain information with higher quality when using the SEKT applications. The professionals which we see as future users of SEKT would benefit strongly from the availability of better information for their tasks. The complementarity of search time and information quality should be kept in mind. [IURISERVICE??]

Potential negative sideffects

The availability of a technical and organisational infrastructure does not guarantee that the expected benefits, in particular the economic benefits, are observed in practice. A decisive factor is the acceptance by a sufficiently large share of the user population. Knowledge management as envisaged by SEKT implies active participation by users, in the form of supporting the semi-automated development of metadata and knowledge sharing. We have observed that strong organisational stereotypes concerning collaboration are in effect.

The possible emergence of new, undesirable forms of behaviour is only visible in realistic tests. An area of concern would be the responsibility for the maintenance and quality of knowledge assets, or the emergence of in-groups insufficiently connected to the mainstream knowledge management process.

These factors must be taken into account when planning profound changes in the IT support for knowledge management. Future RTD could study these factors in more depth, but would require considerable more time after the availability of a technically mature system.

Cost / benefit of SEKT

The cost benefit of SEKT in comparison to traditional knowledge management systems is clearly positive under the assumptions and parameters which we used. The added cost is rather limited and would deliver added functionality and indirect benefits, especially improved quality of information, and improved user satisfaction.

In quantitative terms a reduction of process time (as shown in the IURISERVE case), even if much smaller than observed, would more than offset the cost of a SEKT enabled knowledge management system in a sizeable organisation. The low learning cost for users of SEKT, which users clearly saw in the studies, is an additional quantitative benefit.

The results clearly support the use of SEKT in knowledge management systems for professional users. The challenge for adopters, after assuring a competent and efficient implementation for their particular organisations, is to assure that the solution corresponds to the needs of the organisation, and is widely accepted. This issue is however beyond the scope of our research.

6 Conclusions

The results of the tests are decisively encouraging for the implementation of semantic knowledge technology for information-intensive work of professionals. The subjective assessment of all subjects is highly positive. The subjects expect significant advantages for their personal efficiency, and the quality of their work from the use of applications with semantic technology. The application prototypes for the BT digital library and IURISERVICE for judges in the Spanish legal system demonstrate highly desirable features for the users.

The results also show that semantic knowledge technology can help to generate quantitative advantages – these can be improved performance or higher quality of work. An application which makes this task easier is a clear advantage for the individual user, and for the organisation.

Professionals and knowledge workers are themselves responsible for monitoring the quality of their own work. One aspect of their responsibility is to make sure that the information used is of high quality, representing the state-of-the-art. User behaviour in that respect is flexible, and the context determines which benefits (time saving, better quality, or reduced workload) are obtained. We have seen this in the two studies: Legal decisions may be characterized by the need to achieve no less than a high quality (based on a comprehensive understanding of the background information), while the digital library search may be a case where a certain time budget is allocated by users (based on their prior experience), and when this is exhausted they accept the level of information attained.

We can not exclude the possibility that the search in the DL might result in an increase of information quality if the database contains more documents.

The cost-benefit consideration is that SEKT implementation has costs in the same range as conventional knowledge-management systems based on the use of taxonomies. (The average cost is estimated at 50€ per user per year in large organisations, depending upon size of the organisation and complexity of the service, the range being realistically between 20 and 100 € per user per year.) The added cost of SEKT depends on the complexity of ontology development and maintenance, and ranges between 0 and 20 €.

This cost is offset by added functionality, performance increase, and by increased acceptance of KM solution.

In conclusion, the validation of the SEKT prototypes shows that the claims made for SEKT technology – that it will deliver information to users more efficiently, easier, and with better quality – are strongly supported. The results are highly encouraging and strongly support the commercial exploitation of SEKT in applications for professional knowledge management.

7 References

[1] Bösser, T., and Melchior, E.-M. (2002) User-Centred Product Creation. Best Practice in Interactive Electronic Publishing. ISBN 3-00-009652-3 2002. http://www.vnet5.org

[2] Constantine, L.L., and Lockwood, L.A.D. (1999) Software for Use. A Practical Guide to the Models and Methods of Usage-Centered Design. New York: Addison Wesley.

Dumas, J.S. and Redish, J.C. (1994) A practical guide to usability testing. Norwood: Ablex Publishing.

[3] Kirakowski, J., and Corbett, M. (1993) SUMI: the Software Usability Measurement Inventory, British Journal of Educational Technology 24 (3), 210–212.

[4] Rubin, J. (2001) Handbook of Usability Testing. How to plan, design, and conduct effective tests. Chichester: John Wiley & Sons.

[5] Van Rijsbergen, C. J. (1980) Information Retrieval. London: Butterworths.

8 Appendices

8.1 Information quality measurement

As described in D.8.3.1 User validation methods for SEKT, a method was devised to determine empirically the “information foraging function” for each subject. Such a method was developed with two main components:

- Firstly, a sampling method was developed to obtain on-line measurements from subjects who engage in search-and-browse processes. A set of small programs was developed which can be used to implement studies by configuring appropriate pop-up windows with a variety of information-collection functions.
- Secondly, a factor analysis study was carried out to determine the factors which have to be captured, which is described in the next section below.

The specific methods developed for user validation in the SEKT project are designed to assess information quality. The conceptual basis is the concept of “information foraging” developed by Card and Pirolli.

The goal ist to estimate empirically an information foraging function of the following form:

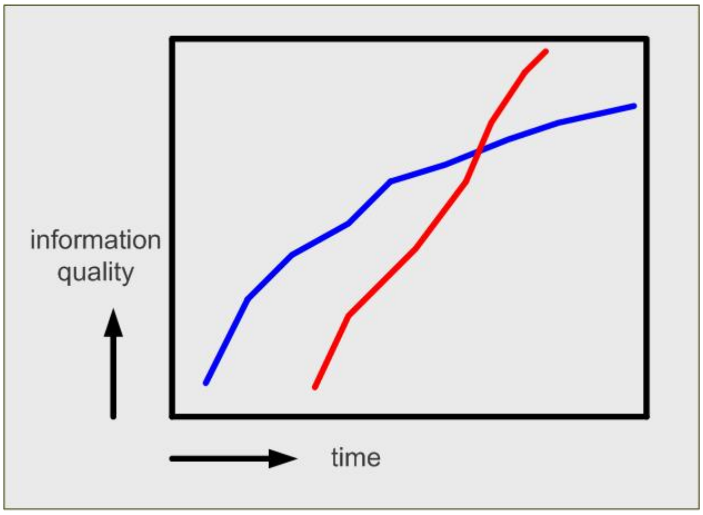


Figure 10: Information search function describing two search processes

The principle to realize this measurement is to carry out measurements of information quality at meaningful points in time, which do not need to be equidistant. These could be stages in an information search process.

The resulting characteristic information search function indicates which quality level is achieved, and how fast the goal of search is approached.

There are different options for defining appropriate measures:

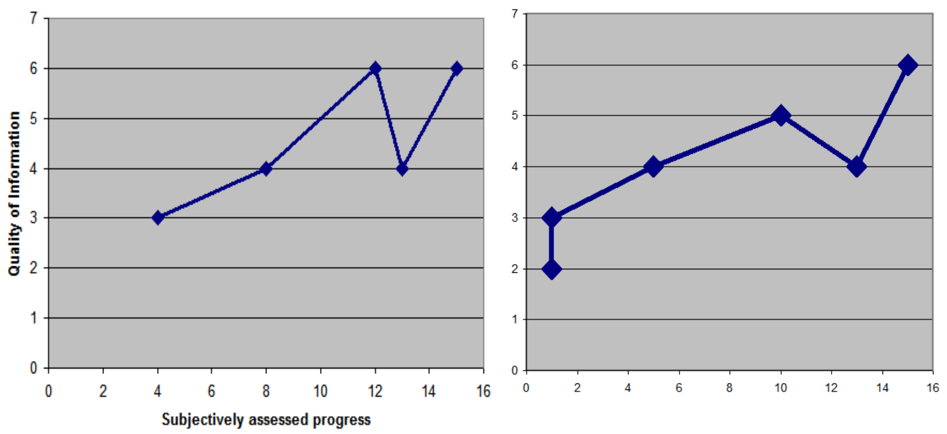
- Subjective assessments by the person which carries out the search

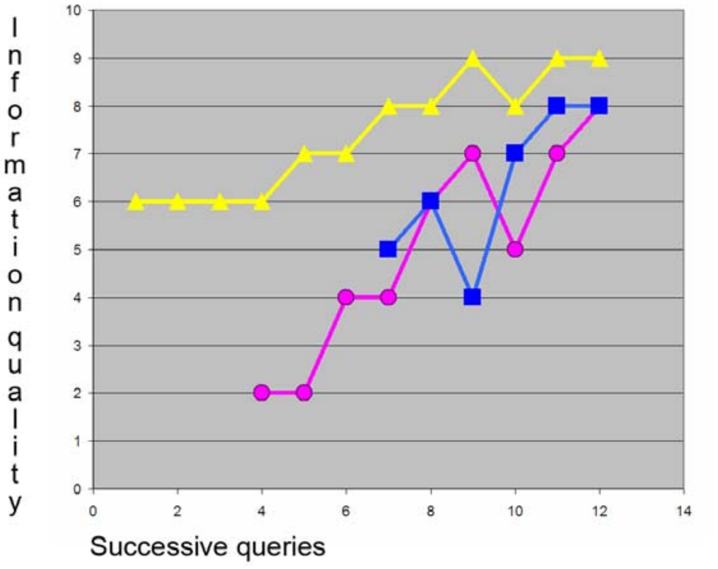
- Subjective assessment by external experts
- Objective measures collect from the search results, or other data

The ability to collect appropriate measures and the purpose of the investigation determine which measures are most meaningful. We will aim at a subjective rating of information quality, because the SEKT applications are offered to professionals with a high degree of autonomy in the organisation of their personal work. Additional objective measures will be used where practicable. The information will be collected with the use of user feedback forms which pop up at system-generated points in time.

The rating scales will be constructed according to proven psychometric principles.

2 search processes, same subject (SBS application K)





Tests with different configurations of the method did show that results as expected can be obtained. Two examples are shown, individual data from a test of the SBS application K (which would have been compared to SEKT), and search processes in a university library with a very extensive set of accessible digital resources (around 30 million documents in total).

8.2 Measures of Information Quality (Factor Analysis)

In order to understand which factors users employ to assess information quality, a study was conducted with 284 subjects who completed a questionnaire with a total of 24 questions to evaluating information quality from different perspectives during a search-and-browse process.

The subjects were visitors of a large digital library of a university, persons working in call center which provides technical support, and persons searching travel information.

The results did show that most of the variance was explained by a general quality factor (40%). The second important factor is “progress” (42% of the variance). A third factor could be called “quality of information presentation”, accounting for 9% of the information. This factor is statistically not reliable with 284 subjects, however.

The data confirm that it makes sense to describe information quality in search processes by a limited number of factors, and allows us to configure effective tests, using the most reliable and valid items from the questionnaire.

8.3 BT digital library data

BT DL - SUMI Scoring Report from SUMISCO 7.38

Profile Analysis

```
Scale             UF    Ucl   Medn  Lcl   LF
Global            71    60    56    52    41
Efficiency        82    54    49    44    16
Affect            80    65    60    55    38
Helpfulness       69    59    55    51    42
Control           70    55    51    47    26
Learnability      78    66    61    56    39
```

Note:

The Median is the middle score when the scores are arranged in numerical order. It is the indicative sample statistic for each usability scale.

The Ucl and Lcl are the Upper and Lower Confidence Limits. They represent the limits within which the theoretical true score lies 95% of the time for this sample of users.

The UF and LF are the Upper and Lower Fences. They represent values beyond which it may be plausibly suspected that a user is not responding with the rest of the group: the user may be responding with an outlier.

Individual User Scores

```
User  Globa Effic Affec Helpf Contr Learn
1     52    35    71    60    38    65     one
2     51    49    51    58    51    34     two  (L)
3     58    57    56    51    56    62     three
4     57    47    62    55    50    60     four
5     62    60    60    67    52    66     five
6     67    62    69    64    65    61     six
7     55    60    62    53    45    59     seven
8     38    38    52    43    33    37     eight  (GL)
9     52    41    60    54    36    61     ten
10    60    52    63    56    48    60     eleven
11    33    20    41    23    45    42     twelve  (GH)
12    56    44    56    48    68    65     thirteen
13    65    68    56    65    66    65     fourteen
14    40    34    34    57    41    32     fifteen  (GAL)
15    43    38    53    46    40    65     sixteen
16    53    55    40    57    51    52     seventeen
17    61    57    66    53    51    64     eighteen
18    56    44    71    51    56    69     nineteen
19    65    65    71    64    53    63     twenty
```

Any scores outside the interval formed by the Upper and Lower Fences are potential outliers. The user who produced an outlier is indicated in the right hand column. The initial letters of the scales in which outliers are found are indicated in parentheses.

Item Consensual Analysis

In the following table, the numbers in the row labelled ‘Profile’ are the observed responses of the actual users to each item.

The numbers in the row labelled ‘Expected’ are the number of responses expected on the basis of the standardisation database.

The Goodness of Fit between the observed and expected values is summarised using Chi Square, and these statistics are presented on the line below the expected values.

The number at the end of the Goodness of Fit line is the total Chi Square which applies to that item. The greater the value of the total Chi Square, the more likely it is that the obtained values differ from what is expected from the standardisation database.

Each total Chi Square marked with

*** is at least 99.99% certain to be different

** is at least 99% certain to be different

* is at least 95% certain to be different

Total Chi Square values without asterisks are not likely to differ much from the standardisation database.

In this output, the SUMI items which differ most from the standardisation are presented first.

```
This software responds too slowly to inputs.
Item 1      Agree Undecided   Disagree
Profile     13    1     5
Expected    3,61  2,9   12,49
Chi Sq      24,44 1,25  4,49  30,18***

Getting data files in and out of the system is not easy.
Item 49     Agree Undecided   Disagree
Profile     0     17    2
Expected    2,62  7,8   8,58
Chi Sq      2,62  10,86 5,05  18,53***

The software documentation is very informative.
Item 15     Agree Undecided   Disagree
Profile     0     18    1
Expected    6,62  9,48  2,9
Chi Sq      6,62  7,67  1,25  15,53***

If this software stops it is not easy to restart it.
Item 9      Agree Undecided   Disagree
Profile     0     15    4
Expected    3,08  7,33  8,6
Chi Sq      3,08  8,04  2,46  13,57**

The software has a very attractive presentation.
Item 42     Agree Undecided   Disagree
Profile     4     12    3
Expected    10,7  5,16  3,14
Chi Sq      4,2   9,06  0,01  13,26**

The speed of this software is fast enough.
Item 29     Agree Undecided   Disagree
Profile     6     1     12
Expected    10,65 3,17  5,18
```

Chi Sq	2,03	1,48	8,99	12,51**

I think this software has made me have a headache on occasions.

Item 37	Agree	Undecided	Disagree	
Profile	0	2	17	
Expected	4,6	3,97	10,43	
Chi Sq	4,6	0,98	4,13	9,71**

There have been times in using this software when I have felt quite tense.

Item 32	Agree	Undecided	Disagree	
Profile	2	2	15	
Expected	7,19	2,89	8,93	
Chi Sq	3,74	0,27	4,13	8,15*

8.4 IURISERVICE Data

IURISERVICE – Judges - SUMI Scoring Report from SUMISCO 7.38

Profile Analysis

Scale	UF	Ucl	Medn	Lcl	LF
Global	71	64	61	58	50
Efficiency	78	68	62	56	33
Affect	73	70	68	66	61
Helpfulness	61	60	58	56	55
Control	57	56	52	48	48
Learnability	73	65	62	59	52

Note:

The Median is the middle score when the scores are arranged in numerical order. It is the indicative sample statistic for each usability scale.

The Ucl and Lcl are the Upper and Lower Confidence Limits. They represent the limits within which the theoretical true score lies 95% of the time for this sample of users.

The UF and LF are the Upper and Lower Fences. They represent values beyond which it may be plausibly suspected that a user is not responding with the rest of the group: the user may be responding with an outlier.

Individual User Scores

User	Globa	Effic	Affec	Helpf	Contr	Learn		
1	58	60	68	56	43	62	1	(C)
2	61	63	63	52	41	66	2	(HC)
3	61	48	69	59	51	55	3	
4	64	62	68	57	54	66	4	
5	55	41	65	57	51	55	5	
6	66	69	71	61	53	62	6	
7	57	62	66	58	60	59	7	(C)
8	56	48	61	59	52	61	8	(A)
9	66	65	71	61	56	71	9	(H)

Any scores outside the interval formed by the Upper and Lower Fences are potential outliers. The user who produced an outlier is indicated in the right hand column. The initial letters of the scales in which outliers are found are indicated in parentheses.

Item Consensual Analysis

In the following table, the numbers in the row labelled ‘Profile’ are the observed responses of the actual users to each item.

The numbers in the row labelled ‘Expected’ are the number of responses expected on the basis of the standardisation database.

The Goodness of Fit between the observed and expected values is summarised using Chi Square, and these statistics are presented on the line below the expected values.

The number at the end of the Goodness of Fit line is the total Chi Square which applies to that item. The greater the value of the total Chi Square, the more likely it is that the obtained values differ from what is expected from the standardisation database.

Each total Chi Square marked with

```
***   is at least 99.99% certain to be different

**    is at least 99% certain to be different

*     is at least 95% certain to be different
```

Total Chi Square values without asterisks are not likely to differ much from the standardisation database.

In this output, the SUMI items which differ most from the standardisation are presented first.

```
The speed of this software is fast enough.
Item 29     Agree Undecided   Disagree
Profile     3     5     1
Expected    5,05  1,5   2,45
Chi Sq      0,83  8,15  0,86  9,84**

Error prevention messages are not adequate.
Item 38     Agree Undecided   Disagree
Profile     0     8     1
Expected    2,24  3,64  3,12
Chi Sq      2,24  5,22  1,44  8,89*

Getting data files in and out of the system is not easy.
Item 49     Agree Undecided   Disagree
Profile     0     8     1
Expected    1,24  3,69  4,07
Chi Sq      1,24  5,02  2,31  8,57*

Working with this software is satisfying.
Item 12     Agree Undecided   Disagree
Profile     9     0     0
Expected    4,84  2,93  1,23
Chi Sq      3,58  2,93  1,23  7,74*

There have been times in using this software when I have felt quite
tense.
Item 32     Agree Undecided   Disagree
Profile     0     1     8
Expected    3,4   1,37  4,23
Chi Sq      3,4   0,1   3,36  6,87*

There is never enough information on the screen when it's needed.
Item 18     Agree Undecided   Disagree
Profile     0     5     4
Expected    1,55  2,01  5,44
Chi Sq      1,55  4,43  0,38  6,35*

I sometimes wonder if I am using the right command.
Item 11     Agree Undecided   Disagree
```

```
Profile     1     4     4
Expected    3,2   1,43  4,38
Chi Sq      1,51  4,64  0,03  6,18*

There are too many steps required to get something to work.
Item 36     Agree Undecided   Disagree
Profile     0     0     9
Expected    1,87  1,79  5,34
Chi Sq      1,87  1,79  2,5   6,16*

I would recommend this software to my colleagues.
Item 2      Agree Undecided   Disagree
Profile     9     0     0
Expected    5,38  2,41  1,21
Chi Sq      2,43  2,41  1,21  6,05*
```

IURISERVICE – Legal experts - SUMI Scoring Report from SUMISCO 7.38

Profile Analysis

Scale	UF	Ucl	Medn	Lcl	LF
Global	76	67	64	60	46
Efficiency	74	68	64	60	48
Affect	81	71	68	64	51
Helpfulness	68	62	59	55	46
Control	77	64	59	54	39
Learnability	73	65	59	52	45

Note:

The Median is the middle score when the scores are arranged in numerical order. It is the indicative sample statistic for each usability scale.

The Ucl and Lcl are the Upper and Lower Confidence Limits. They represent the limits within which the theoretical true score lies 95% of the time for this sample of users.

The UF and LF are the Upper and Lower Fences. They represent values beyond which it may be plausibly suspected that a user is not responding with the rest of the group: the user may be responding with an outlier.

Individual User Scores

User	Globa	Effic	Affec	Helpf	Contr	Learn		
1	64	63	71	58	58	60	11	
2	69	69	71	63	63	64	12	
3	65	60	66	69	49	55	13	(H)
4	63	66	69	59	60	63	14	
5	59	57	58	60	64	65	15	
6	66	66	66	57	69	57	16	
7	48	44	61	50	41	29	17	(EL)
8	68	66	69	60	65	67	18	
9	56	54	51	53	57	47	19	
10	56	65	71	48	51	55	20	

Any scores outside the interval formed by the Upper and Lower Fences are potential outliers. The user who produced an outlier is indicated in the right hand column. The initial letters of the scales in which outliers are found are indicated in parentheses.

Item Consensual Analysis

In the following table, the numbers in the row labelled ‘Profile’ are the observed responses of the actual users to each item.

The numbers in the row labelled ‘Expected’ are the number of responses expected on the basis of the standardisation database.

The Goodness of Fit between the observed and expected values is summarised using Chi Square, and these statistics are presented on the line below the expected values.

The number at the end of the Goodness of Fit line is the total Chi Square which applies to that item. The greater the value of the total Chi Square, the more likely it is that the obtained values differ from what is expected from the standardisation database.

```
Each total Chi Square marked with

***   is at least 99.99% certain to be different

**    is at least 99% certain to be different

*     is at least 95% certain to be different

Total Chi Square values without asterisks are not likely to differ
much from the standardisation database.

In this output, the SUMI items which differ most from the
standardisation are presented first.


This software seems to disrupt the way I normally like to arrange my
work.
Item 16     Agree Undecided   Disagree
Profile     2     7     1
Expected    0,97  2,39  6,64
Chi Sq      1,08  8,92  4,79  14,79***

The software hasn't always done what I was expecting.
Item 41     Agree Undecided   Disagree
Profile     0     2     8
Expected    4,65  2,26  3,09
Chi Sq      4,65  0,03  7,81  12,49**

I find that the help information given by this software is not very
useful.
Item 8      Agree Undecided   Disagree
Profile     0     8     2
Expected    2,22  3,21  4,57
Chi Sq      2,22  7,14  1,44  10,8**

I keep having to go back to look at the guides.
Item 30     Agree Undecided   Disagree
Profile     5     4     1
Expected    1,99  2,23  5,78
Chi Sq      4,55  1,41  3,95  9,91**

The software has at some time stopped unexpectedly.
Item 4      Agree Undecided   Disagree
Profile     0     2     8
Expected    4,69  1,06  4,25
Chi Sq      4,69  0,84  3,3   8,83*

Working with this software is satisfying.
Item 12     Agree Undecided   Disagree
Profile     10    0     0
Expected    5,38  3,25  1,37
Chi Sq      3,98  3,25  1,37  8,6*

I sometimes wonder if I am using the right command.
Item 11     Agree Undecided   Disagree
Profile     0     4     6
Expected    3,55  1,59  4,86
Chi Sq      3,55  3,68  0,27  7,49*

If this software stops it is not easy to restart it.
Item 9      Agree Undecided   Disagree
Profile     0     8     2
Expected    1,62  3,86  4,53
```

```
Chi Sq      1,62  4,45  1,41  7,48*

There are too many steps required to get something to work.
Item 36     Agree Undecided   Disagree
Profile     0     0     10
Expected    2,07  1,99  5,94
Chi Sq      2,07  1,99  2,78  6,84*

I would recommend this software to my colleagues.
Item 2      Agree Undecided   Disagree
Profile     10    0     0
Expected    5,98  2,68  1,35
Chi Sq      2,71  2,68  1,35  6,73*

This software occasionally behaves in a way which can't be
understood.
Item 46     Agree Undecided   Disagree
Profile     0     2     8
Expected    3,22  2,52  4,26
Chi Sq      3,22  0,11  3,28  6,61*

Using this software is frustrating.
Item 27     Agree Undecided   Disagree
Profile     0     0     10
Expected    1,71  2,14  6,15
Chi Sq      1,71  2,14  2,41  6,26*

There have been times in using this software when I have felt quite
tense.
Item 32     Agree Undecided   Disagree
Profile     0     2     8
Expected    3,78  1,52  4,7
Chi Sq      3,78  0,15  2,32  6,25*
```

8.5 Inspection of the SBS SEKT application

The aim of user tests with the SBS SEKT application was to demonstrate the benefit created by SEKT in the form of improvements over the existing *knowledgemotion* system. User tests were prepared by enabling a comparison of the existing ´knowledgemotion´ versus the system enhanced with SEKT components (K vers S).

Tests were prepared by building an ontology for the S system, selecting and indexing a relevant dataset. After inspecting the system, a number of tasks were defined, taking account of the specifics of the installation, and additional data were uploaded.

In a second iteration, tests were carried out with the set of defined tasks, by three experts. Results of querying the knowledge base were: All queries which retrieved documents from the S set of information provided clearly shorter lists of results than the K system – higher "precision" as demanded by users in the user needs studies. It could be seen that this is due to the datamodel in S (ontology). For the user, clearly reduced user time results (fewer queries, reduced browsing in the results list), which, as we know from the user needs analyses, is a high priority for the users.

The support for preparing knowledge assets in a semi-automatic fashion is a SEKT specific functionality which is not available in the K system. It is assessed as highly positive (reducing annotation effort and enhancing the quality of metadata), but would require a realistic test context to be assessed quantitatively.

Tests were not continued further at this stage, because

- Profound organisational changes were taking place
- The organisation was undergoing changes with consequences expected for the knowledge management organisation and processes
- Very large datasets would have to be integrated (and used) in order to make further tests meaningful

We concluded from the expert evaluation that the SEKT application for SBS corresponded to the needs of users, as they were determined in the earlier SEKT user needs analysis.