Are *Semantic* Desktops Better?

Summative Evaluation Comparing a Semantic against a Conventional Desktop

Thomas Franz, Ansgar Scherp, and Steffen Staab
University of Koblenz-Landau, Universitätsstr. 1, 56070 Koblenz, Germany
{franz,scherp,staab}@uni-koblenz.de

ABSTRACT

Semantic desktop environments aim at improving the effectiveness and efficiency of users carrying out daily tasks within their personal information management (PIM) infrastructure. They support the user by transferring and exploiting the explicit semantics of data items across different PIM applications. Whether such an approach does indeed reach its aim of facilitating users' life and — if so — to which extent, however, remains an open question. In this paper we address this question with the first summative evaluation of a semantic desktop. We have developed a test environment to evaluate two semantic PIM applications against standard PIM tools. As result, we have found significant efficiency and satisfaction improvements for typical PIM tasks.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology*; H.4.1 [**Information Systems Applications**]: Office Automation

General Terms

Experimentation, Management

1. INTRODUCTION

Significant efforts have been invested in research on the semantic desktop in recent years (e.g. Nepomuk project (12 million EUR funding), Calo (200 million US$)) and a lot of achievements have been made through such research, including the availability of various semantic desktop prototypes [16, 14, 5, 13, 10]. Unlike approaches to automate personal information management (PIM) processes, e.g. [6, 11], semantic desktops establish information linkage and reuse across the boundaries of PIM tools. Yet, the very important question whether the ultimate goal of semantic desktop research, namely improved support for PIM, has been achieved, still remains unanswered. Formative [12] evaluations have been conducted, giving inside into *how* particular semantic desktop tools are used, e.g. [16]. Yet, to the best of our knowledge, no publications are available that complement formative results with summative [12] evaluations answering the question whether semantic desktops perform *better* than state-of-the-art desktops. Accordingly, it is neither known if the use of a semantic desktop results in superior PIM effectiveness, efficiency, or satisfaction, nor do we know which kind of PIM tasks semantic desktops support best and which semantic desktop approaches are most promising to pursue. In this paper, we shed light on these questions, presenting the first summative evaluation of a semantic desktop. The core contribution of this paper is threefold:

1. We introduce the design of a reproducible, summative, and task-based evaluation for comparing semantic desktops with state-of-the-art desktops (Sect. 3).
2. We report on encouraging results gathered from the conduction of such an evaluation with 18 test persons (Sect. 4), showing that a semantic desktop indeed improves PIM. We have measured significant efficiency and satisfaction improvements of users working with a semantic desktop compared to users working with an off-the-shelf desktop, particularly when working on complex PIM tasks.
3. We provide insight into potential future directions of semantic desktop research by extracting implications from the observations made throughout the evaluation (Sect. 5).

The evaluation comprises 10 tasks, created by the instantiation of common task types defined in PIM research [1, 8, 2, 17]. The tasks have been executed based on a real world scenario and corresponding original data, extracted from two event managers of a conference-like event at the University of Koblenz. The evaluation compares COSIMail and COSIFile, two tools of the X-COSIM semantic desktop, with off-the-shelf PIM tools, i.e. the Thunderbird email client and the KDE file manager. COSIMail and COSIFile implement core features of a semantic desktop, supporting information linkage and reuse among the two well examined PIM domains of email and file management [7, 4, 15, 3].

2. EVALUATING SEMANTIC DESKTOPS

The eventual goal of semantic desktop research is to improve PIM support. Accordingly, the goal of our evaluation is to

K-CAP'09, September 1–4, 2009, Redondo Beach, California, USA.

investigate whether users benefit from the utilization of a semantic desktop and if so, what the benefits are for which kinds of tasks. We summarize these goals by the following goal hypothesis for the evaluation: *The utilization of a semantic desktop leads to improved PIM support for common PIM tasks, when compared to a conventional desktop.* In the following, we elaborate on the meaning of this hypothesis and derive requirements for an evaluation design.

Improved PIM support: Improvement of PIM support can only be measured by the comparison with state-of-the-art PIM support. Accordingly, a summative instead of a formative evaluation needs to be conducted. While formative evaluations enable to investigate on the usability of a system, i.e. give answers to the question *how good* a system is, summative evaluations give answers to the question *which is better* [12]. Commonly accepted indicators for improvement should be considered for the comparison of systems such as user effectiveness, user efficiency, and user satisfaction.

Semantic Desktops: Results of an evaluation targeting the above hypothesis are most valuable if they can be generalized for a class of systems instead of being restricted to a specific system only. Accordingly, the evaluated system should provide features common for the class of systems and should build upon technology shared by the majority of systems in that class. For semantic desktop systems, formal ontologies, and RDF as a flexible data structure represent the shared technology. It enables the semantic linking of information to mediate between PIM applications, the common feature of semantic desktops [16, 14, 5, 13, 10].

PIM Tasks: Based on the analysis of research on PIM, we distinguish between three dimensions of PIM tasks: The first dimension defines the types of PIM tasks when interacting with information systems for PIM, e.g. as described by Barreau [1], including the *acquisition*, *organization*, *storage*, and *retrieval* of information, as well the task of system *maintenance*. The second dimension describes application types that support the execution of such tasks for specific information items, e.g. email management, file management, text processing. The third dimension represents application domains in which application types and task types are combined with respect to a specific domain, e.g. biology, conference management. Semantic desktop tools, e.g. [16, 14, 5, 13, 10], predominantly focus on supporting email management and file management as core and well examined application types, with email even considered as a "habitat" for PIM [7]. An evaluation of a semantic desktop should accordingly consider these prominent application types and build upon known and examined tasks defined by research on PIM. In addition, the application domain of an evaluation needs to be understood by all test persons to ensure that results are not flawed by a lack of understanding.

3. EVALUATION DESIGN

In the following, we briefly[1] explain the design of an evaluation by considering the requirements posed above. We have designed a task-based, summative evaluation where two groups of test persons each execute the same set of tasks, however, using two different desktops loaded with an identical data corpus. One group using a semantic desktop, the other group working on a standard desktop environment. To ensure that observed user behavior is not flawed by the confrontation with an unfamiliar data corpus, the corpus has been extracted from real data about the night of computer science[2] at the University of Koblenz, an event in which all test persons actively participated as hosts. The corpus consists of 119 original emails and 66 files, including reports, spreadsheets, and images extracted from the desktops of two organizers. 18 test persons from the computer science department of the University of Koblenz have participated in the evaluation, 15 PhD students and 3 graduate students. They were randomly assigned into the two groups and had not used a semantic desktop before.

3.1 Evaluation Process

The evaluation consists of three phases, namely *introduction*, *observation*, and *feedback*. The introduction phase familiarizes test persons with the data corpus and the tools used in the observation phase. For all of the involved test persons, this introduction has taken no more than 5 minutes. In the observation phase, the actual evaluation is carried out, i.e. test persons execute the tasks of the task set on either the semantic or the conventional desktop. Here, no further assistance to test persons is given while their behavior is observed using screen recordings and by taking notes. The observation phase is followed by a feedback phase where test persons fill-in a questionnaire, and where they express further subjective feedback in a short interview. The evaluation process is supported by a graphical tool (cf. Fig. 1) we have developed to guide test persons through the different tasks of the evaluation and for presenting a description and contextual information for each task. The tool also presents a feedback questionnaire and logs indicators for user efficiency and effectiveness.

3.2 PIM Tasks

Test persons have executed 10 different PIM tasks. The original (German) and translated task instructions as presented to the test users are available online[3]. Following the design rationale introduced in Sect. 2, selected tasks are instances of task types defined in PIM research (cf. [1, 8, 2, 17]). The task set includes the task of organization, retrieval, and storage as these are the tasks commonly supported by semantic desktops. The application domains of these tasks are email and file management, the core domains supported by todays semantic desktop tools. The first task T1 requires to *organize* files and emails into folder hierarchies, resembling organization tasks as defined in [1]. It has the additional purpose of familiarizing test persons further with the data corpus and the tools used throughout the evaluation. The tasks T2 and T3 are lookup tasks as defined in [8], where a piece of information inside some document (e.g. an email, or spreadsheet)

[1] Please see the technical report [9] for more details.

[2] http://nacht-der-informatik.uni-koblenz.de/

[3] http://isweb.uni-koblenz.de/Research/DataSets

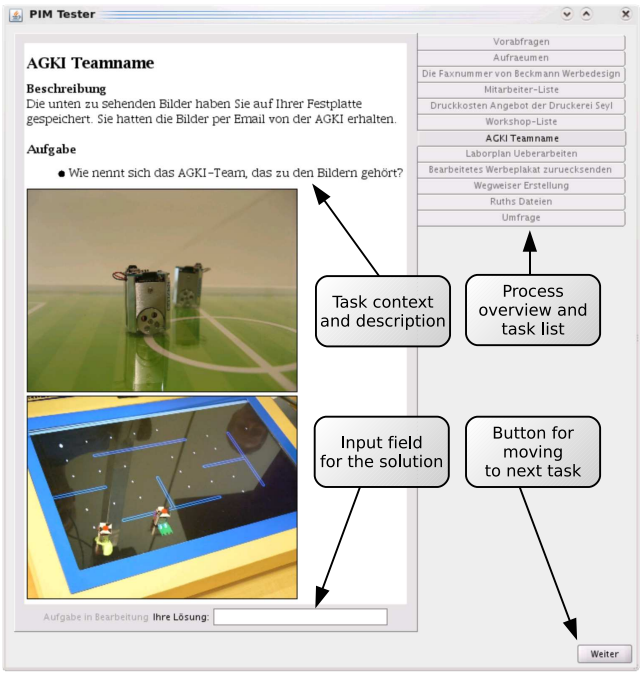

Figure 1: Screenshot of the Evaluation Wizard

needs to be found (e.g. a phone number) while the document itself may be exactly known or not. Lookup tasks are well supported by search functionalities of PIM tools (e.g. search in Thunderbird) while information linkage and reuse, as provided by semantic desktops, is not expected to be of any help. However, we included T2 and T3 as baseline tests in order to check the equality of the two user groups of the evaluation. Tasks T4-T10 are used to compare the semantic and the conventional desktop. Task T4-T6 are multi-item tasks (also defined by [8]), i.e. tasks where information from multiple documents needs to be collected or processed, e.g. T6 asks for the name of a team working with miniature robots given two images that have been received as attachment to an email containing the name (Fig. 1 illustrates how the task is presented to test persons). Task T7-T9 are more complex tasks combining lookup and multi-item tasks, additionally requiring to edit, compare, and send files or emails. They resemble collaborative document editing via email as defined in [2, 17], e.g. T9 asks to insert a logo into a file sent by a colleague and to return it. The final task T10 is about information collation, combining the tasks of organization and retrieval [1]. Test persons are asked to identify files retrieved by two colleagues and to put them into a specific folder.

3.3 Semantic Desktop

As representatives for semantic desktop tools, COSIMail and COSIFile have been selected for the evaluation. They provide common features of a semantic desktop, supporting information linkage and reuse for the dominant PIM domains of email and file management. A screencast illustrating COSIMail and COSIFile is available online[4].

[4]http://vimeo.com/2278034

Both tools build upon the X-COSIM [10] semantic desktop which defines the following layered architecture (cf. Fig. 2): The bottom layer provides an RDF store where all metadata is kept. The metadata is described by the X-COSIM ontology (X-COSIMO) that enables formally precise descriptions in spite of context-dependent conceptualizations as commonly employed by end-user applications. The sec-

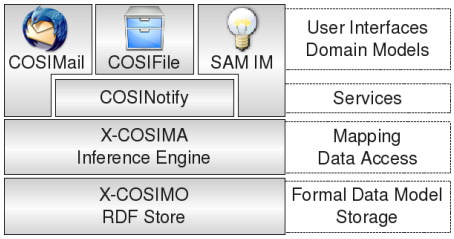

Figure 2: X-COSIM Architecture

ond layer, X-COSIMA, maps between applications on upper layers by inferring contextualized views of information. For instance, if an email attachment is stored to the file system, then it is inferred that the information object representing it can be mapped to two different contextualized representations, namely the concept `Attachment` and `FileSystemRealization`. Easy-to-use programming objects built upon the contextualized representations and enable programmatic access to stored metadata while hiding the complexity of the underlying conceptual model given by X-COSIMO. The third layer provides generic services such as COSINotify, a software component that tracks system operations to update the semantic desktop store, e.g. updating location information for a moved file. On the top layer are semantic PIM tools which employ the domain specific programming objects to exploit semantic metadata about the information they deal with, e.g. email messages, filesystem information, and to-do descriptions.

COSIMail is one PIM tool of the top layer, built as an extension for the Thunderbird email client. COSIMail contributes and retrieves metadata to and from the semantic desktop store, similar to the Nepomuk Foxtrott[5] extension for the Firefox browser. Among others, this metadata includes an email's body and subject, sender and recipients, sent date, and information about attached files. While a number of exploitations of such metadata can be envisioned, e.g. enhanced email search as implemented by the Seek[6] extension for Thunderbird, COSIMail exploits the metadata in an additional user interface widget that enables to access stored email attachments or view their location on the filesystem. The widget displays stored attachments as a blue link, similar to links on web pages (see Fig. 3). By left and right mouse clicks, users can open the saved file and the containing folder, respectively.

COSIFile is another semantic PIM tool, available as an extension for file managers of the KDE (http://kde.org) and the GNOME desktop (http://gnome.org).

[5]http://code.google.com/p/nepomuk-mozilla/
[6]http://simile.mit.edu/seek/

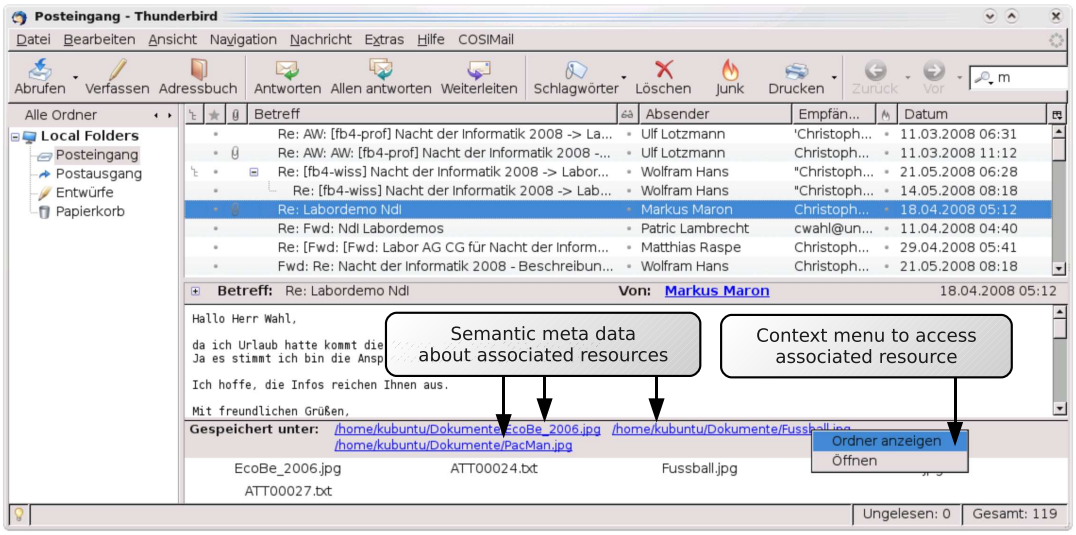

Figure 3: Screenshot of COSIMail showing Emails of the Evaluation Data Corpus

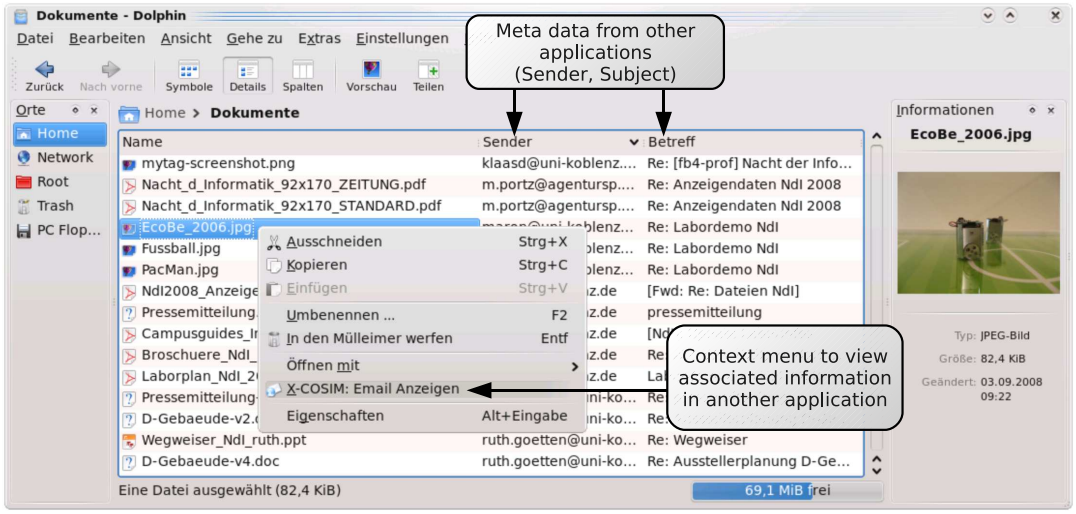

Figure 4: Screenshot of COSIFile showing files of the Evaluation Data Corpus

COSIFile augments file managers by a means to switch from the file manager view of information to the view implemented by another PIM application. For this, it provides an additional menu entry in the file context-menu, e.g. for viewing the message to which a file was attached (cf Fig. 4). Moreover, COSIFile improves support for manual search by browsing the filesystem, users preferred search strategy [4]. Like the Nepomuk dolphin extension (http://nepomuk.kde.org) that supports comments and tags as additional file properties, COSIFile also enables additional file attributes (cf Fig. 4). Attributes that are considered useful by human users, e.g. the sender of a file [3], can be used to sort files accordingly. File managers of conventional desktops do not support such file attributes.

4. RESULTS

For each test person, we have measured indicators of effectiveness and efficiency in the observation phase, while we gathered indicators on satisfaction thereafter (cf. Sect. 2).

4.1 Effectiveness

We have measured the effectiveness of users by observing whether test persons were able to finish a task. Additionally, the evaluation wizard logged answers that test persons entered into the input field presented for each task (cf. Fig. 1). All 18 test persons indicated that they were able to complete all tasks. By an analysis of the answer logs we recognized three wrong answers, two of them by test persons using the semantic desktop tools.

4.2 Efficiency

For each task, we measured the time, number of window switches, mouse clicks, and distance of mouse movements that test persons spent for completing the task. Table 1 lists the measurements per task and group of test persons. The table also lists the relative performance of the group working with X-COSIM tools, shown as percentage of the performance of the other group, e.g. a value of 50% indicates that test persons using the X-COSIM desktop needed 50%

Table 1: Average and Median Execution Times, Window Switches, Mouse Clicks, and Mouse Movements

			Intro	Baseline		Evaluation Tasks						
			Organization	Lookup		Multi-Item			Doc. Driven Coll.			Collation
			T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Execution Time	Conventional	Avg.	13:14	01:10	01:18	03:31	02:04	02:28	04:37	04:16	05:00	06:00
		Med.	13:30	01:17	00:57	03:48	02:02	02:10	04:42	04:10	05:40	07:14
	X-COSIM	Avg.	10:52	01:00	01:28	03:52	01:25	01:37	04:14	03:34	03:07	03:29
		Med.	06:44	01:01	01:03	03:03	01:27	01:43	03:53	03:50	02:44	03:34
	rel. Perf.	Avg.	82%	86%	113%	110%	69%	66%	92%	84%	62%	58%
		Med.	50%	79%	111%	80%	71%	79%	83%	92%	48%	49%
	T-Test		0.305	0.259	0.373	0.376	**0.014**	**0.042**	0.319	0.178	**0.004**	**0,019**
Wind. Switches	Conventional	Avg.	27.13	5.50	6.63	14.75	13.38	7.86	18.00	14.75	18.14	26.67
		Med.	14.00	4.00	6.50	13.00	13.00	7.00	15.00	14.50	18.00	27.00
	X-COSIM	Avg.	12.75	4.63	5.88	18.63	10.88	5.63	13.00	10.88	10.71	5.29
		Med.	7.00	4.00	6.00	14.50	9.00	5.00	10.50	13.00	11.00	4.00
	rel. Perf.	Avg.	47%	84%	89%	126%	81%	72%	72%	74%	59%	20%
		Med.	50%	100%	92%	112%	69%	71%	70%	90%	61%	15%
	T-Test		0.306	0.228	0.266	0.191	0.159	0.114	0.187	0.113	**0.003**	**<0.001**
Mouse Clicks	Conventional	Avg.	562.57	19.00	256.00	21.43	90.14	77.83	98.17	84.57	108.00	135.67
		Med.	353	13	331	17	69	78.5	86	74	111	144
	X-COSIM	Avg.	315.56	20.67	222.33	22.89	38.88	30.75	88.10	88.50	53.00	136.14
		Med.	202	22	232	20	32	29	70.5	67	55	125
	rel. Perf.	Avg.	56%	109%	87%	107%	43%	40%	90%	105%	49%	100%
		Med.	57%	169%	70%	118%	46%	37%	82%	91%	50%	87%
	T-Test		0.121	0.384	0.338	0.400	**0.006**	**0.014**	0.373	0.439	**0.002**	0.495
M. Movements	Conventional	Avg.	126260	16364	14146	34937	30408	23922	67536	52042	70852	82991
		Med.	106927	8490	13786	39113	29627	17052	61083	56344	70673	74892
	X-COSIM	Avg.	83647	12029	13733	40957	16582	16862	42236	39477	45448	47745
		Med.	44405	9957	11808	30395	15749	15989	40032	39521	38505	45190
	rel. Perf.	Avg.	66%	74%	97%	117%	55%	70%	63%	76%	64%	58%
		Med.	42%	117%	86%	78%	53%	94%	66%	70%	54%	60%
	T-Test		0.149	0.224	0.433	0.327	**0.017**	0.130	**0.064**	0.143	**0.018**	**0.041**

of the time, window switches, mouse clicks or mouse movements that test persons on the conventional desktop needed. We also analyzed the statistical significance of the measured performance differences by the computation of t-tests for each task. Values where we measured statistically significant ($P(T \leq t) \lesssim 5\%$) differences are printed in boldface. Fig. 5

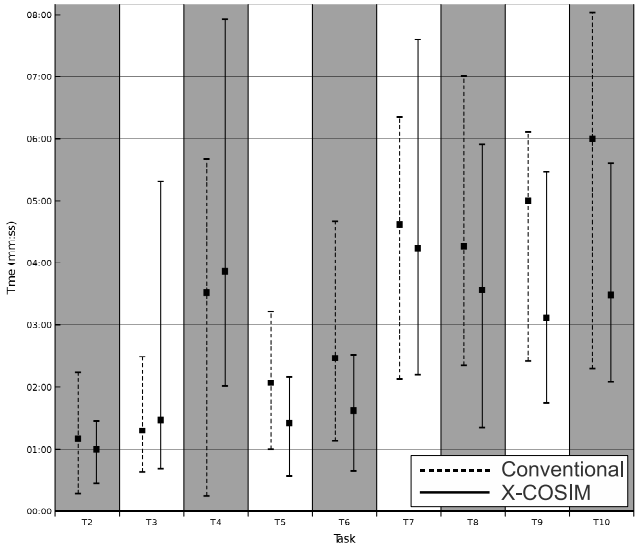

Figure 5: Min., Max., and Avg. Execution Times

adds further details about the distribution of the measured execution times for task T2-T10. It presents the minimum, average, and maximum execution times for the two groups of test persons.

T1 Organization: On average and median, test persons spent most of their efforts on the initial organization task (cf. Table 1). Two test persons (one using the X-COSIM desktop, one using the conventional desktop) decided not to organize files and emails at all.

T2-T3 Lookup: For the two baseline tasks, test persons of both groups needed the least execution times among all tasks (T1-T10) (cf. Table 1). Test persons using the X-COSIM tools needed less time to execute T2, however, needed more time to execute T3, yet, the differences are not statistically significant.

T4-T6 Multi-Item: Among the three multi-item tasks, both groups spent the most time, largest amount of mouse movements, and most window switches on the first task T4 (cf. Table 1). For T4, the average performance of X-COSIM users is weaker then for the other group, yet, not significantly. Conversely, for T5 and T6, the group testing the X-COSIM tools was more efficient with respect to all efficiency indicators. Execution times and mouse clicks are significantly reduced for T5 and T6, and mouse movements are significantly reduced for T5.

T7-T9 Document-Driven-Collaboration: Over all these tasks, test persons using X-COSIM tools required less time, less mouse movements and less window switches (cf. Table 1). Their reduction in mouse movements has been significant for task T7 and T9. Additionally, significant reductions for all indicators have been measured for T9.

T10 Information Collation: Also for T10, we have measured significant performance improvements by users of the X-COSIM desktop for all efficiency indicators except the number of mouse clicks.

4.3 Satisfaction and Usability

We have measured the satisfaction of test persons by statements derived from the IsoMetrics^s questionnaire. Test persons were asked to rate the statements using the IsoMetrics rating scale: 1=*strongly disagree*, 2=*disagree*, 3=*so-so*, 4=*agree*, 5=*strongly agree*. The group using the X-COSIM desktop was asked to rate 8 statements about the usability and utility of COSIMail and COSIFile (a detailed listing of all statements and ratings can be downloaded at http://isweb.uni-koblenz.de/Research/DataSets). Test persons responded positively to these statements with an average score of 4.01, i.e. *agreeing* that the presented features of the semantic desktop are useful. Additionally, test persons of both groups were asked to rate 3 general statements on the perception of the task set. The results are presented in Table 2. Test persons working with X-COSIM tools considered the task set less tedious (S9, reduction by 0.7 and 1.0 avg/median) and felt significantly less difficulties in dealing with the data corpus (S11). Both groups responded equally about the statement that the task set corresponds to tasks in their own work (S10), responding *so-so* (3.4) on average and *agree* (4.0) in the median.

5. DISCUSSION

As presented above, significant efficiency improvements have been measured for users of the semantic desktop. In the following, we detail on specific results based on further observations made and conclude the discussion with a derivation of high level implications.

5.1 Effectiveness

Concerning effectiveness, none of the 18 test persons indicated to be unable to solve the presented tasks and only 3 wrong answers have been measured. Accordingly, we cannot conclude that the X-COSIM tools result in increased effectiveness of their users. Yet we infer that the tasks chosen for the evaluation have been adequately presented and resemble typical PIM tasks so that test persons had no difficulties in understanding and executing the tasks.

5.2 Efficiency

We observed that test persons working with the X-COSIM desktop put less effort into the organization task T1 (cf. Table 1). The difference, however, is insignificant. Yet, from comments during the feedback interview we presume that the use of a semantic desktop results in the impression that less organizational effort is needed, e.g. one test person explained that *" ... such features reduce the need for organization ... "*. Another test person stated that *"... having more of such features, I would not organize at all ..."*.

Task T2 and T3 have been included in the task set as baseline tasks. Insignificant performance differences and partly even lower performance of X-COSIM users on these tasks suggest that both groups of test persons have been comparable with respect to their PIM skills.

For multi-item task T5 and T6, users of the X-COSIM desktop have been significantly faster than users of the conventional desktop system. Only for the first multi-item task T4, users working on the conventional desktop have been faster, however, not significantly. We observed that some test persons of the X-COSIM desktop initially misinterpreted the semantic connection of files and emails provided by COSIMail (cf. Fig. 3) which may explain their weaker performance on T4. Multi-item task T6 turned out to be an illustrative example of the benefits of information linkage provided by semantic desktops. T6 has been presented as shown in Fig. 1. It requires to find the name of the research team working with the robots depicted on the two images. The group working on the X-COSIM desktop solved T6 by starting to search for the images on the filesystem, exploiting the image-preview feature of the file manager. Having found the images, they exploited the link to the email message that contained the team name. Most persons of the other group also started T6 by using the same strategy, however, got stuck once they had found the images as they could not easily switch to the associated email.

For all of T7-T9, the group using X-COSIM tools outperformed the other group with respect to all efficiency indicators, partly even significantly (cf. Table 1). From this observation we conclude that features of our semantic desktop become more beneficial with increasing task complexity.

On the final task T10, test persons using the X-COSIM tools have also been significantly more efficient than test persons using the conventional tools with respect to execution time, mouse movements, and window switches (cf. Table 1). We explain the highly significant (at the 1% level) reduction of window switches by the fact that test persons using X-COSIM tools could solve this task using COSIFile exclusively. Instead, solving the task with conventional tools required to switch back and forth between the email client and the file manager due to the lack of email metadata available on the standard file manager.

5.3 Satisfaction

Test persons working on the semantic desktop felt less stressed (S9 reduction by 0.7/1.0 avg/med) and perceived the tasks as significantly less difficult (S11 reduction by 1.3/1.5 avg/med). Accordingly, we believe that they enjoyed the PIM tasks more than test persons working with conventional tools. This interpretation also maps to the positive subjective feedback about the usefulness and usability of the semantic PIM tools. However, in the feedback interview, test persons also remarked the slow response times of the test systems (due to running on a virtual machine) and initial confusions about the semantic connections presented by COSIMail.
Test persons using the conventional desktop expressed a slight tendency (3.5 on average and median, S11, Tab. 2) to-

Table 2: Subjective Feedback from both Groups of Test Persons

	Statement	Conventional Avg	Conventional Med	X-COSIM Avg	X-COSIM Med	T-Test
S9	Working on the tasks was tedious/cumbersome	3.0	3.0	2.3	2.0	0.149
S10	The tasks correspond to task types that I also need to do for my work	3.4	4.0	3.4	4.0	0.493
S11	Tasks were significantly harder to do, since I was required to work with unknown data	3.5	3.5	2.2	2.0	**0.023**

wards accounting difficulties in the task execution to a lack of knowledge about the data corpus. Instead, test persons of the X-COSIM desktop rather disagreed (Avg 2.2, Median 2.0), indicating that they considered their limited knowledge on the data corpus as unproblematic. These answers are significantly different. We conclude that the additional support of a semantic desktop improved users' perception on their ability to execute the task-set successfully. Our measurements on their efficiency, as presented above, show that their perception was right. We see identical ratings for S10 as further indicators for the balance of both groups. Nevertheless, we expect an even stronger identification with the presented tasks for users with a less technical work background then test persons in our evaluation who are used to special IT infrastructure. This impression is backed by a test persons stating that tasks like document-driven-collaboration occur only rarely as "... for such things, there is SVN ...".

5.4 Lessons Learned and Implications

From the results presented and discussed above, we may conclude several lessons learned and implications.

Data-Driven Desktop: As we have seen above, the utility of a semantic desktop depends on the task type. It performed the strongest, where relations between information are essential for solving a task quickly (cf. Task T10). While on the semantic desktop, relations are inferred, conventional desktops provide only application-centric user interfaces requiring the user to build relations manually. We notice that semantic desktop extensions open-up the conventional desktop towards a data-centric desktop where applications rather act as view mechanisms that can be adapted, e.g. by additional file attributes, or changed, e.g. by enabling to switch easily between applications based on selected information.

Creative Exploitation and Behavioral Changes: We observed that test persons got more acquainted to the additional features of the X-COSIM tools with increasing usage time. They started to invent their own strategies of using them, e.g. we observed a test person who used the link to saved attachments as provided by COSIFile (cf. Fig. 3) to support the selection of the file in the attach-file-dialog of the email client. Such unexpected utilization comes along with a change of user behavior. As we found out in Sect. 4.2, test persons using the semantic desktop have seen less need for organizing information. Noticing such exploitation and adaptation, we believe that constant use of a semantic desktop would improve PIM even more than measured in our evaluation.

Intuitiveness: The relations made explicit by our semantic desktop seem to be intuitive to use as they support human associative thinking. For instance, as discussed for T6, the connections provided by our semantic extensions have enabled test persons to pursue the strategy they had in mind first. Lacking semantic linkage of information items, users of the conventional tools experienced unsatisfying situations that required them to re-think their solution strategy and to spent more effort.

Interaction Design: Functional and efficient technology is only one aspect of a useful semantic desktop. We have learned that good interaction design is crucial for users to benefit from provided enhancements when executing everyday PIM tasks. As shown by the evaluation results for Task T4, little irritations cause tremendous efficiency loss. Our strategy of extending existing tools rather than building new tools turns out to be a strong requirement for developing a usable semantic desktop that can compete with state-of-the-art productive software for PIM.

Training: Our test persons are well-versed computer users who have received less than 5 minutes for the introduction to the tools as used in the evaluation. We have additionally conducted the same evaluation with two secretaries, measuring much weaker performance compared to the 18 test persons. In subsequent interviews we found out that the unfamiliar KDE environment (both use a Windows environment in daily work) resulted in strong distraction and irritation. We conclude that a comparable evaluation with less-versed users would require a lot more up-front training or working environments that are closer to the one used by those users.

Acceptance and Readiness: Users of our semantic desktop system accepted the enhancements provided, enjoyed their utilization, and have been more productive than users of the conventional desktop. Furthermore, throughout the evaluation, we experienced no system failure in any of our semantic desktop components. These results show both the technological readiness as well as the user interface maturity of our desktop. Moreover, the stability of both systems enabled to measure high-quality results as test persons could work through the task set without interruptions caused by system failures.

6. RELATED WORK

While there is a large body of work related to the enhancement of PIM support, e.g. by approaches to automate PIM processes [6, 11], we point to semantic desktop tools and the like in the following. As this paper is *not* about the comparison of semantic desktops we focus on the evaluations conducted for these systems.

Haystack [14] is a semantic desktop that provides a monolithic application for personal information management such as email management and calendaring. Unlike Haystack, IRIS [5] integrates existing applications, e.g. the Mozilla suite for web browsing and email management, into one user interface. It was developed within the CALO research

project[7] and serves there as knowledge store. To the best of our knowledge, an evaluation of Haystack or IRIS towards user effectiveness, user efficiency, and user satisfaction has not been published.

Gnowsis, partly supported by the Nepomuk semantic desktop project, provides a user front-end for connecting files, contacts and further PIM assets. A study reporting about the long-term utilization of Gnowsis by two users has been published [16]. While the study gives insight into the specific utilization by those users, it does not reveal any information about the benefits of such systems compared to conventional desktop systems.

UMEA [13] is a PIM application supporting the manual association of documents, folders, URLs, and contacts to *projects*. The UMEA user interface provides project views for accessing associated information quickly. Unlike semantic desktops, UMEA does not build upon an open and formal model to describe the information it deals with. Thus, it cannot exploit reasoning capabilities and is restricted to the mentioned information items and one kind of relation among them. A formative evaluation of UMEA was conducted by eight users over a period of two to six weeks were users responded positively about the general idea behind UMEA.

Conventional desktops are currently beginning to establish information linkage. For instance, the Mac OS X desktop indexes emails and filesystem documents, including provenance information about a file, e.g. an identifier of an email it was sent with. As of today, such information cannot be utilized effectively by users as appropriate user interface widgets, e.g. as provided by COSIMail and COSIFile, and machine-interpretable descriptions of the information are missing. However, our evaluation has shown that such features are desireable and provide significant improvement for PIM.

With respect to the semantic desktops mentioned above, formative evaluations have been carried out partly, providing valuable insight into the utilization of semantic desktops. The summative evaluation presented here enables to compare semantic desktops with conventional desktops. To the best of our knowledge such a kind of evaluation has not been published before for any semantic desktop.

7. CONCLUSION

We have presented a reproducible summative evaluation that fills the gap of summative evaluations of semantic desktops that are missing so far. Significant efficiency and satisfaction improvements by users of a semantic desktop are encouraging results that provoke continued research efforts and an expansion towards further PIM domains and tasks other than email and file management, e.g. system maintenance [1]. Moreover, the evaluation has shown that semantic web technology can be exploited successfully on the desktop where it enables mature and useful PIM tools. Thus, semantic desktop technology is ready to be adopted by developers of productive desktop systems, e.g. as started for the KDE desktop (http://nepomuk.kde.org/).

[7] http://caloproject.sri.com/

Acknowledgements

This work was funded by the X-Media project (www.x-media-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978. We thank David-Paul Mann for implementation work and Christoph Thul for supporting the video analysis.

8. REFERENCES

[1] D. K. Barreau. Context as a factor in personal information management systems. *JASIS*, 46(5):327–339, 1995.

[2] V. Bellotti, N. Ducheneaut, M. Howard, I. Smith, and R. E. Grinter. Quality versus quantity: E-mail-centric task management and its relation with overload. *Human-Computer Interaction*, 20:89–138, 2005.

[3] T. Blanc-Brude and D. L. Scapin. What do people recall about their documents? implications for desktop search tools. In *Intelligent User Interfaces*, pages 102–111, 2007.

[4] R. Boardman and M. A. Sasse. "Stuff goes into the computer and doesn't come out": A cross-tool study of personal information management. In *CHI*, pages 583–590, 2004.

[5] A. Cheyer, J. Park, and R. Giuli. IRIS: Integrate. Relate. Infer. Share. In *Workshop On The Semantic Desktop, ISWC*, 2005.

[6] M. Dredze, T. Lau, and N. Kushmerick. Automatically classifying emails into activities. In *IUI*, 2006.

[7] N. Ducheneaut and V. Bellotti. E-mail as habitat: An exploration of embedded personal information management. *Interactions*, 8(5):30–38, 2001.

[8] D. Elsweiler and I. Ruthven. Towards task-based personal information management evaluations. In *SIGIR*, pages 23–30. ACM, 2007.

[9] T. Franz, A. Scherp, and S. Staab. Does a semantic desktop facilitate your daily tasks? Technical Report 12, Universität Koblenz-Landau, 2008. ISSN (Online) 1864-0850.

[10] T. Franz, S. Staab, and R. Arndt. The X-COSIM Integration Framework for a Seamless Semantic Desktop. In *K-CAP*, 2007.

[11] M. Freed, J. G. Carbonell, G. Gordon, J. Hayes, B. Myers, D. P. Siewiorek, S. Smith, A. Steinfeld, and A. Tomasic. Radar: A personal assistant that learns to reduce email overload. In *AAAI*, pages 1287–1293. AAAI Press, 2008.

[12] G. Gediga and K.-C. Hamborg. Evaluation of software systems. *Encyclopedia of Computer Science and Technology*, 45, 2001.

[13] V. Kaptelinin. Umea: Translating interaction histories into project contexts. In *CHI*, pages 353–360, 2003.

[14] D. R. Karger, K. Bakshi, D. Huynh, D. Quan, and V. Sinha. Haystack: A General-Purpose Information Management Tool For End Users Based On Semistructured Data. In *CIDR*, pages 13–26, 2005.

[15] P. Ravasio, S. G. Schär, and H. Krueger. In pursuit of desktop evolution: User problems and practices with modern desktop systems. *ACM Trans. Comput.-Hum. Interact.*, 11(2):156–180, 2004.

[16] L. Sauermann and D. Heim. Evaluating long-term use of the gnowsis semantic desktop for pim. In *ISWC*, 2008.

[17] S. Whittaker and C. L. Sidner. Email Overload: Exploring Personal Information Management of Email. In *CHI*, pages 276–283, 1996.