Authors:	Kumpulainen Sanna, Järvelin Kalervo
Name of article:	Information interaction in molecular medicine : integrated use of multiple channels
Name of work:	IIiX '10 Proceeding of the Third Symposium on information interaction in Context (New Brunswick, New Jersey, USA, August 18 - 21, 2010)
Year of publication:	2010
ISBN:	978-1-4503-0247-0
Publisher:	ACM
Pages:	95-104
Discipline:	Natural sciences / Computer and information sciences
Language:	en

URN: http://urn.fi/urn:nbn:uta-3-874
DOI: http://dx.doi.org/10.1145/1840784.1840800

This is a preprint of the paper: Kumpulainen, S. & Järvelin, K. (2010). Information Interaction in Molecular Medicine: Integrated Use of Multiple Channels. In: Belkin, N. & al. (Eds.), Proc. of the Information Interaction in Context conference (IIiX 2010), New Brunswick, NJ, August 2010. New York, NY: ACM, pp. 95-104. ISBN: 978-1-4503-0247-0. DOI=10.1145/1840784.1840800
http://doi.acm.org/10.1145/1840784.1840800

Information Interaction in Molecular Medicine: Integrated Use of Multiple Channels

Sanna Kumpulainen
Department of Information Studies and Interactive Media, University of Tampere
FIN-33014 University of Tampere, Finland.
sanna.kumpulainen@uta.fi

Kalervo Järvelin
Department of Information Studies and Interactive Media, University of Tampere
FIN-33014 University of Tampere, Finland.
kalervo.jarvelin@uta.fi

ABSTRACT
Task-based information access is a significant context for studying information interaction and for developing information retrieval (IR) systems. Molecular medicine (MM) is an information-intensive and rapidly growing task domain, which aims at providing new approaches to the diagnosis, prevention and treatment of various diseases. The development of bioinformatics databases and tools has led to an extremely distributed information environment. There are numerous generic and domain-specific tools and databases available for online information access. This renders MM as a fruitful context for research in task-based IR. The present paper examines empirically task-based information access in MM and analyzes task processes as contexts of information access and interaction, integrated use of resources in information access and the limitations of (simple server-side) log analysis in understanding information access, retrieval sessions in particular. We shed light on the complexity of the between-systems interaction. The findings suggest that the system development should not be done in isolation as there is considerable interaction between them in real world use. We also classify system-level strategies of information access integration that can be used to reduce the amount of manual system integration by task performers.

1. INTRODUCTION

Task-based information access is an important viewpoint in analyzing user-system interaction in information retrieval [10][28]. Work tasks trigger the actual information needs which are transformed into search tasks that lead to information access.

In real life, information access (IA) depends on the task at hand, personal practices, and available resources [10]. The present paper focuses on IA in research tasks in molecular medicine (MM), which is an information-intensive and rapidly growing domain aiming at providing new approaches to the diagnosis, prevention and treatment of various diseases. There are numerous generic and domain-specific tools and databases available for online information access. Therefore it is not surprising that the Web is reported in surveys as the most important information source for MM researchers [21]. However, in surveys "the Web" may be a convenient and unanalytical label for many kinds of channels and resources accessed

through the Web. The NAR online Database Collection alone lists 1230 selected databases in the domain [6]. The information environment consists not only of biological databases, but also of literature databases, and other types of websites as well, for facilitating work tasks in the domain. The present paper shows that such resources are used in integrated ways with search engines and tools on PC. This renders MM as a fruitful context for research in task-based IR.

The focus of the paper is on task-processes, task complexity and the integrated use of information channels in order to seek, retrieve and manage information during work task performance. Our work is based on rich observational and logging data on six researchers ('users') in MM. Thereby it relates to earlier research in IR focusing on personal search histories and search systems' query logs but differs from them in essential aspects.

1.1 Related research

Typical web user studies focus on large datasets over considerable time periods, but are limited to a single system or engine [9]. The results of such studies seem applicable to other search engines, while the joint use of resources in various types of channels remains unseen. In contrast to general Web use, the task-based context necessitates the study of task performers ('users'), their tasks and the organizational and social context of task performance. We argue that traditional log analysis has serious limitations in the study of task-based IA.

Obtaining contextual information for Web search engine logs is difficult. It is therefore challenging to make inferences about the individual search engine user's search goals and satisfaction on the basis of log data. The logged queries may be analyzed as sessions but remain as weakly contextualized labels, which may be proxies for more or less muddled or explorative information needs; factual, known item, or topical needs; or be informational, navigational or transactional needs [4][8]. Teevan & colleagues [26] present a model for query personalization, which seeks to predict query intent from click data but is, nevertheless, a server side analysis and in the generic rather than task-based domain. Xie [29] analyzed interactive intentions and search strategies of library users (both physical browsing and OPAC sessions). In both studies, just one (type of) system was in focus. The subjects were not observed with their work tasks. However, work task goals direct searching behavior and help to make sense of sessions and queries.

A number of studies have looked at task-based IA but, typically, the tasks are seen as search tasks or information seeking tasks [13] [16]. While one learns about, e.g. Web browser use in tasks like Fact Finding, Information Gathering, Browsing, and Transactions [13], the integration of various resources to work tasks remain unobserved. Other studies define tasks as the goals of information-seeking behavior [1][7][14][22], but do not study work tasks as *processes*.

Some studies have focused on the biomedical field, studying work tasks and processes as the context of information access/ behavior [2][17][25]. However, these studies focus on a specific task [2] or on abstract task classes [17][25] rather than real life task performance as the context of IA.

1.2 The Present Study

We focus on work task sessions where multiple tools are integrated for IA. Thus, we see IA as a subtask of work tasks. We analyze work task processes (sessions) based on shadowing field notes and search logs in MM at three levels: the query/navigation level (providing statistics on query types), the session/interaction level (transitions in information access), and task level (task complexity affecting information interaction patterns in work task performance sessions). Our study is based on extensive data on work tasks (24 task sessions) of six researchers in molecular medicine. We employed interviews, shadow-

ing, and logging in data collection. The data are very rich and, as a consequence, a further goal is to develop a method for analyzing this kind of data on task-based information access.

We use the concept *information access* (IA) as a broader term to traditional query-based retrieval in text databases or the Web, but at the same time as a term of narrower scope than task-based information seeking. IA covers retrieval, browsing and navigation in all kinds of collections of data, such as bio-databases and literature databases, or other media including printed material and handwritten notes.

Channels are information resource categories. Information is accessed via channels during task performance processes. We used five distinct channel categories: Web Search Engines (WE), General Web Sites (WS), Literature Databases (LDB), Bio-Databases (BDB) and Other (O) which included all printed material and human information resources used.

Toms [27] has defined *information interaction* as "the process that people use in interacting with the content of an information system". Information interaction can be examined at three levels: as (a) interaction between humans and system content, (b) integrated use of systems (between channels), (c) integrating systems in terms of data exchange and semantics. We focus on information interaction as interaction between different channels, which describes elaborately the integrated use of various tools and databases. The various systems used have an effect upon one another and are not used in isolation, and information is transferred between systems.

A *session* can be conceptualized in several ways. There are three types of session definitions. Firstly, sessions are semantic in nature. As an example, He and Göker [9] proposed that the start and end of a session are the points where the intent behind a query changes. Secondly, sessions can be defined as non-semantic constructs, defined as series of queries within a small range of time [23] or in terms of the granularity at which the log data is gathered [12]. Thirdly, a session can refer to a work task performance session. We assumed the last approach. Our concept of session is based on semantic bounds of work task performance period. A work task thus consists of sessions within which IA in several channels may be performed. In the entire span of a work task, its sessions may be interleaved with sessions of other work tasks.

Work tasks are of varying *complexity*, from routine to highly complex [5]. We will classify the work tasks in our data on the basis of the task performers' knowledge on the inputs, performance process, and results of the tasks when beginning the tasks.

2. RESEARCH DESIGN

2.1 Research Questions

We shall focus on the following research questions:

a) Work task sessions: how do work task sessions differ at different levels of task complexity?

b) Interaction: how are various channels / information integrated in interaction?

c) Queries: what types of keys or queries are used in interaction?

d) Methodology: how to analyze task-based IA using rich data on task processes for answering the above questions a-c?

2.2 Data collection

In this study, we used interaction log data, interview data and shadowing field notes collected during six months in years 2007 and 2008. Shadowing is a method, which involves a researcher closely following a member of an organization over a period of time to uncover actions performed in the real-time context of

an organization [18]. Throughout the shadowing period, the researcher asks questions which will prompt a commentary from the shadowee. Some questions will be for clarification, other questions to reveal a purpose, or, as in this study, the intentions behind the tasks. Field notes were continuously written during the shadowing.

Session logs were collected with the PLogger tool [15], which was installed on a proxy server and the subjects' browsers were set to direct all traffic through it. PLogger is a tool for collecting visited http URLs in chronological order and for analyzing them. It consists of two components, the ProxyLogger for logging the URLs and the LogBrowser for analysis of the logs. The users are able to edit their logs before submitting them to the investigator. This is an important feature for user acceptance of logging while not all subjects were sensitive about their logs. The logged URLs were collected into a relational database for initial pre-processing and filtering.

The shadowees were interviewed at the beginning of the shadowing to find out about their research processes, current tasks and their perceptions of their information access. The field notes were collected by shadowing six molecular medicine researchers for an average of 24 hours per person over periods from three to eight weeks. The shadowees were selected from two research groups, three from a cancer genetics group and three from a bioinformatics group. Shadowing took place in naturalistic settings and supported the investigator in constructing the shadowees' normal tasks. They were performing real tasks under familiar conditions and employing their current practices and orientations. The only artificiality was caused by the unavoidable presence of the shadower at times and the interference of data collection instrumentation. Any problems in task performance could be immediately observed. When clarification was needed immediate interaction was possible. This permitted to study activities that people might have been unable (e.g. due to forgetting) or unwilling to report (failures). Indeed, the interaction of tasks, information goals and the integrated use of various access tools became understandable. This is not possible through other means of data collection unless the investigator is a domain specialist.

2.3 Analysis

At first, sessions were identified in the data. A session is the unit of analysis. The data were cut into sessions based on shadowing notes. The field notes and log data on the sessions were merged. A total of 24 sessions, four sessions from each shadowee, were analyzed. Sessions were cut according to the work task performed during the shadowing session, including all discussion about the task at hand. An entire work task may consist of more than one session, interrupted by e.g. other work tasks. Session length range and average length of the sessions by researchers are shown in Table 1.

Table 1. Average session length and sessions length range (minutes)

Shadowee	Range	Average
R1	30-75	45
R2	30-100	66
R3	30-170	125
R4	30-100	61
R5	40-150	100
R6	30-140	80
Average	32-123	80

The various resources used for information access were classified into channels. Log data revealed the browser traffic for channels "Web search engines", "Web sites", "Literature databases", and, "Bio-Databases" (for short, WE, WS, LDB, BDB). Tools on the shadowees' computers (such as laboratory information management system and standard office tools) were classified into the channel called "PC". Other resources include printed sources such as books or one's handwritten notes, colleagues and other kinds of resources – the channel "Other" (O, non-digital). PC and Other channels were classified based on the shadowing field notes. The transitions between the channels were counted, but not the traffic between different databases/tools inside a given channel, because within-channel transitions could not be systematically identified in the logs and a sufficiently close observation would have severely interfered with the task process. Workflow charts (see Figs 2, 3, and 4) illustrate the sessions, transitions between, and queries in different channels.

Work tasks and work goals are based on the field notes. Shadowing was the only means by which such information could be collected. The search logs did not include any information about the context of the information access behavior.

Table 2. Task complexity based on a priori task performance knowledge

Task Complexity / Knowledge about	Complex				Semi-Complex			Routine
Resources needed	-	-	-	X	X	-	X	X
Task process	-	-	X	-	X	X	-	X
Type of outcome/goal	-	X	-	-	-	X	X	X

Sessions were classified into three complexity levels: complex, semi-complex, and routine (see Table 2). Complexity was assessed on the basis of the knowledge the shadowees had when beginning each task on the tools/services to be used on the task, the protocol to be used ('know how') and on the assumed outcome of the task. If all of these aspects were known, the task was classified as routine, if two of these were known, the task was semi-complex, and if one or none were known, the task was complex. Among the 24 sessions, 11 were routine, 9 semi-complex and 4 complex. Distribution of tasks among shadowees and their research groups and backgrounds are shown in Table 3.

Table 3. Position and task complexity distribution of shadowees

	Position	Group *	Tasks		
			Complex	Semi-complex	Routine
R1	PhD Student	1	0	0	4
R2	Post Doc	1	2	1	1
R3	PhD Student	1	2	1	1
R4	PhD Student	2	0	0	4
R5	Post Doc	2	0	4	0
R6	PhD Student	2	0	3	1
Sum			4	9	11

* Group 1 = Cancer Genetics, Group 2 = Bioinformatics.

Queries were classified according to the search goal and the type of the service (see Fig 1). The Web retrieval query category was based on Broder's [4] taxonomy, but it had to be extended. For Web queries, we added the factual query class. The Web retrieval query category was supplemented with categories for bio-databases and literature databases. Query goals were chosen based on the recorded user intentions on the field notes, and they were topical (T), factual (F), resource (R) and for web search engines navigational (N).

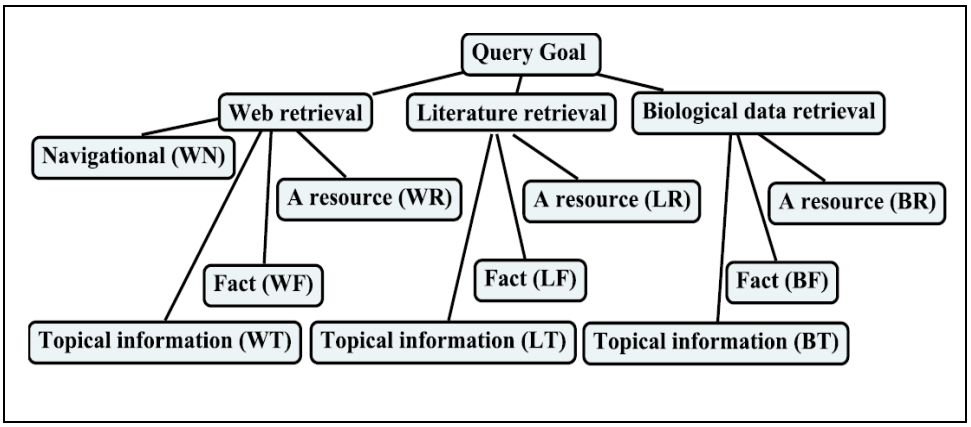

Figure 1. Query goals.

3. RESULTS

3.1 The Work Task and Session Levels

Work tasks were part of the shadowees' own research projects or parts of larger collaborative projects. The tasks shadowed in the cancer genetics group were about analyzing data, some bioinformatics tasks (e.g. "siRNA design"), and designing laboratory experiments. In the other group, bioinformatics group, the shadowed tasks were about updating a database, article writing, and analyzing data. The data analysis tasks differed between the two groups based on the nature of their research focus.

Two of the shadowees were PhDs and four PhD students in different phases of their doctoral projects. The work tasks we observed were parts of their longer-term commitments. Each work task consisted of one or more sessions, lasting from 20 minutes to 2 hours, and were sometimes interrupted by other tasks and working day limits. We focused on sessions involving IA, excluding all sessions performed in the wet-lab alone, or meetings.

Figures 2 – 4 represent three sample work tasks of different levels of complexity. The top lines indicate any queries employed, coded in the classification of Figure 1. The bottom lines approximate task duration. Each shaded and white section in the bar represents 15 minutes (based on the log data). Note that they are, graphically, of varying lengths because some time-slots involve many events while others may involve just examining a found resource.

The middle area is divided into broad horizontal channels (e.g. bio-databases). Each line within a channel represents the same unique resource throughout the session. The solid arrows represent the workflow, the dashed arrows represent data flows, and the dotted arrows represent transitions by links. The top lines represent the query types occurring in the sessions, these are further analyzed in Section 3.3. The middle areas represent the transitions from resources and channels to other. These are further analyzed in the next section.

In Figure 2 there is a workflow chart of a routine task, taking about 30 minutes, where the process alternates in eleven steps between tools on PC and three distinct biological resources. The queries are topical (BT) and factual (BF) in a bio-database. In the session the researcher is finding information about what is known about a target gene.

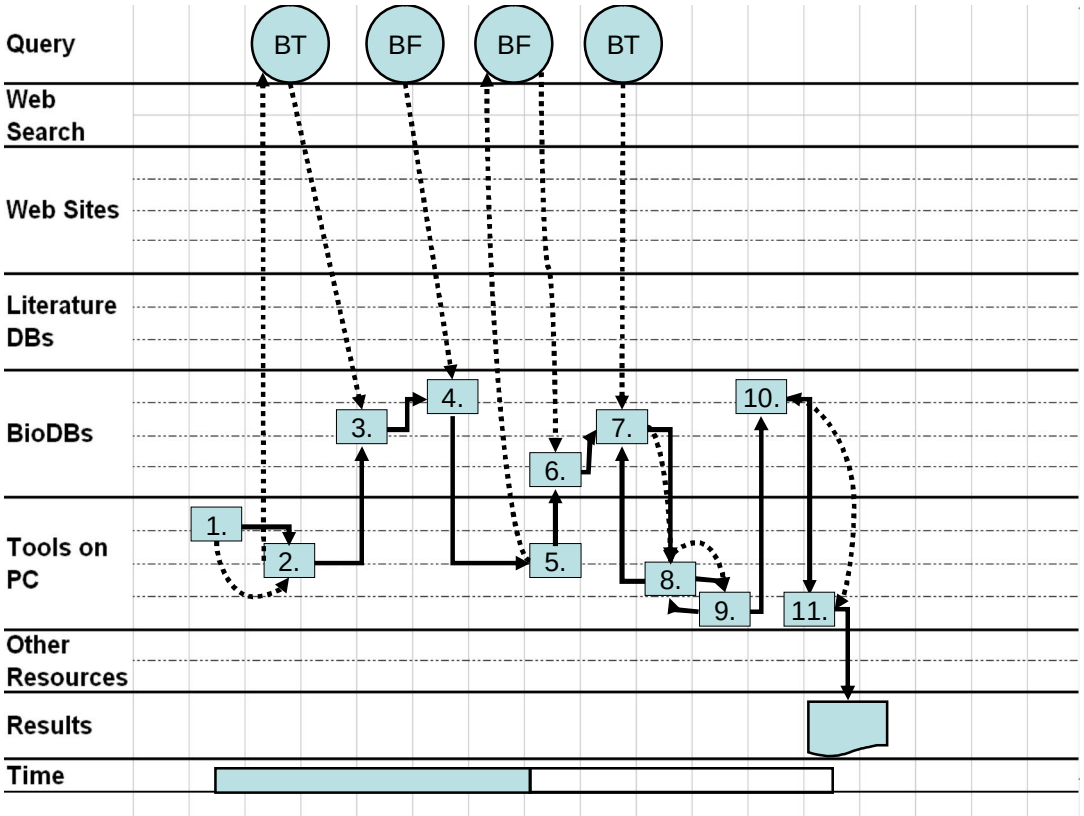

Figure 2. Workflow chart of a routine task session.

(1) She starts by producing a list of interesting targets originating from experiment laboratory data. (2) She makes the list and downloads it into a spreadsheet program. (3) She picks an item from the list, copies it to a bio-database, where she can retrieve published oncology microarray data. She wonders the results for being peculiar. (4) She pastes the same target into another gene-centered database, and thinks the results are wrong and that the gene name might be wrong. (5) She goes back to the spreadsheet and copies the sequence of the target. (6) She pastes it into a service in which a nucleotide sequence is compared with the contents of a nucleotide sequence database, in order to identify the sequence. (7) She goes back to the first bio-database, and copies the new gene name there. (8) She collects the information found on gene (in pictorial form) after a small-scale reformulation with a basic image processing tool and (9) collects all the information into a presentation program on her computer. She repeats the phases (7), (8), and (9). (10) She switches to the gene database she visited earlier in phase (4) and (11) she copies information to the spreadsheet. The outcome is an aggregate of information about a gene found on the two databases.

Figure 3 represents a semi-complex session consisting of 28 steps over duration of more than one hour. While the majority of action takes place between three bio-resources, this session spreads across all channels. The queries are navigational (WN) and topical (WT) Web queries and bio-resource queries (BR). The vertical column in the middle represents a break of fifteen minutes, when a colleague comes to talk about a different work task (teaching a course). The task is to revise a web service in order to write an article manuscript. (1) He opens an earlier started manuscript in a document preparation system, reads and writes some. He finds some information about what should be done to the database. (2) He searches the Web with a search engine in order to shift to a (3) nucleic database of homologous processed pseudogenes. He reads the description, the querying instructions and the information about the form of the result list (4). He then switches to a web dictionary and checks some words and (5) returns to manuscript writing. He repeats the phases (4) and (5). (6) He looks at some other paper on his computer and looks at the references on that paper. (7) He searches some files on PC and (8) opens reference management software and (9) starts to edit the reference list on the manuscript. (10) He goes to a bio-database (which is the subject of his manuscript), and reads the search instructions. (11) Then there is a pause of 15 minutes (talk about a course with a colleague).

After the pause (12) he logs on to a server with a client from his PC. (13) He checks again the instruction pages on the latest database, and (14) copies a figure from there on his PC. (15) He looks at some program code on the server and edits it. (16) He goes back to the database and looks at the results page. (17) Then he searches with web search engine by a four keyword query on a topical issue on user instructions of a program. He eyeballs through the result page, (18) but returns to the database and glances trough it. (19) Then he edits again the code, (18) goes back to database to test it, and repeats the two phases (18 and 19). (20) The he goes to a gene-centered database and searches with a gene name. (21) He copies the access number from the result to the database he has been working on. (22) He is then searching with web search engine for a topical information need about how to code some feature into the database. (23) He moves to a discussion forum where he finds the answer. (24) Then he returns to editing the code. (25) He again searches the web with web search engine with similar information need, but on different topic on programming language. He looks at the result list, but does not select anything from that. (26) He goes back to programming. (27) He crosschecks the earlier results from the bio-database, and ends the session. The result of the session is a revised database on the Web for researchers in MM.

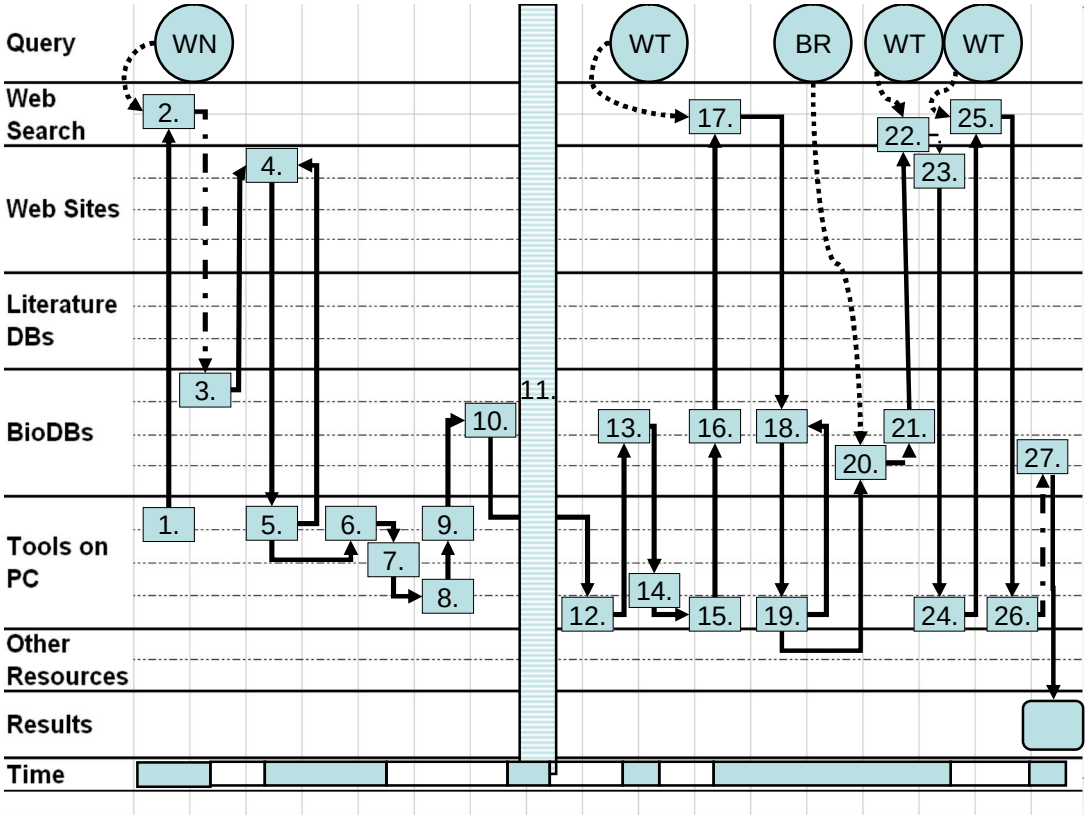

Figure 3. Workflow chart of a semi-complex session.

Figure 4 represents a complex session consisting of 23 steps over duration of about 2 hours. While the majority of action takes place between three PC tools and two bio-databases, search engines and websites are utilized as well. The queries are: two for resources in bio-databases (BR), two navigational for the Web (WN), and six for factual in bio-databases (BF). (Every query was not put in a separate circle because of the lack of space; two of them have two leaving arrows). The vertical column in the middle represents a lunch break of an hour (14). The task is a bioinformatics task (siRNA design). Because the shadowee does not have knowledge about the procedure, the tools and how to use them and she is not able to assess the quality of the outcome, this session is classified as complex.

(1) She starts the session by searching a gene-specific database with a gene name. (2) Then she follows a link to another bio-database, where she finds a sequence. She is unsure how to proceed, and (3) she is seeking some information from a paper. (4) She goes to a bioinformatics course Web page for instructions, but does not find anything useful. (5) She uses a Web search engine in order to move (6) to a certain bio-tool. She copies the sequence from the second service to that tool and waits for results. Then she repeats the first two actions in steps (7) and (8). (9) She uses Web engine to switch (10) to the bio-tool (same as in step 6). (11) Then she returns to the first database, and repeats the search. (12) She follows a link to another database and browses the dataset and finds something interesting. (13) She goes to a genome browser and searches with a gene name, (14) followed by a one-hour lunch break. When she continues, (15) she goes on with the genome browser. (16) She has left the browser window at box 12 open in the background, and returns to it. She browses the data and finds something, and (17) compares the results with some data on a spreadsheet on her PC. The targets were not same, so (18) she goes back to the first gene database and looks at the previously retrieved result list. She checks some result and (19)

follows a link to an article in a literature database. She eyeballs through the abstract but (20) returns to the datasets she was browsing earlier. She queries the datasets with gene name, (21) selects some datasets, copies and saves the list on her PC. (22) She looks at the gene database and (23) the gene browser before giving up and quitting. The outcome was unsuccessful in terms of carrying out the task.

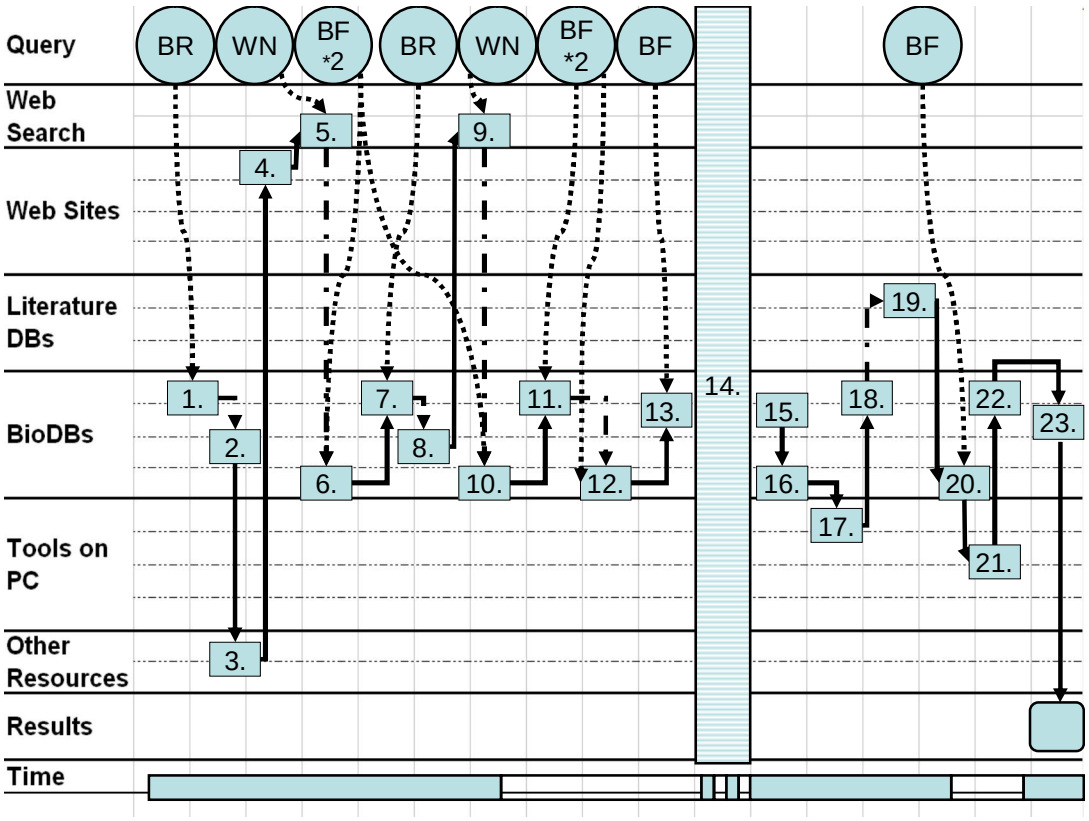

Figure 4. Workflow chart of a complex session.

While challenging to construct, the workflow charts are very illustrative regarding the work task processes, by presenting an overview of their duration, used channels and resources, and of their query types and they can give an easy and quick glance about the details on intricate task processes. The examples discussed above reveal multiple transitions between several channels of different types, and could have created several server side logs at search engines and bio-databases.

3.2 The Interaction Level

Information is transferred in the interaction processes between channels and between resources within the channels. We first discuss the content of interactions between channels and resources and then transition statistics between them.

Within a channel, typical interactions involve finding a sequence of a target gene to be used in another bio-database for information retrieval (see Fig 4). *Between channels* information is typically transferred between the PC and bio-databases. Information on the PC, derived from, e.g., lab experiments, is used to search bio-databases to gather information about the target genes that seem interesting. In the other direction, the information found on these targets is gathered on the PC for later use, or, the data found is further edited to be used in another bio-database. When acquiring literature, the task is usually a writing

task and literature is typically sought for referencing. In that case, a popular approach is to copy-paste sentences from a manuscript into a web search engine or into a literature database (PubMed). This strategy works quite well with web search engines but with PubMed it is not that useful, because the service only provides bibliographic data and possibly links to articles. From literature, usually the transferred information is hints about useful tools, data or other literature items. Between the channels Other and the rest, usually some colleague gives a hint about a resource and how to find it, or some paper or book gives some good query words for solving an information problem or resource need.

Transition tables (Tables 4 - 7) show the integrated use of various resources and channels in sessions. The statistics on the transitions are normalized so that each task has an equal weight for the overall statistic. These are shown as percentages of all transitions in a session. Shading is used to indicate the frequency of transitions, darker shading indicating increased frequency. Tables are read from left to right, for example, on the line WE in Table 4 the first possible transition is to WS, which means the move from web engine to web site and so on. The capital '**N**' stands for the number of sessions, and '**n**' denotes the gross number of transitions in the sessions.

The summary of transitions (Table 4) shows the overall trends of transitions. Most of the shifts between channels are concentrated between two channels, bio-databases and PC. There can be seen some transitions between the PC channel and all remaining channels, but transitions from WE channel to PC channel are quite rare, for understandable reasons. Most transitions from web engine are to the WS channel, and some to the bio-database channel. The bio-databases are in the so called deep web, so some of these transitions are done by navigational searches. One reason for this kind of behavior is that the tools to be used in complex tasks are not known.

Tables 5-7 detail the transitions at the different task complexity levels. The complex sessions' transitions (Table 5) are spread more evenly across the different channels than in simpler tasks. There is not much concentration between just two channels - in contrast to the overall picture - but different channels are entered from multiple directions. Mostly there were transitions from web engines to bio-databases (13.6 %) and from PC to bio-databases, and vice versa, from bio-databases to PC (12.5 %). It is interesting to note that 10.6 % of transitions are from bio-databases to web engines. These figures may indicate that the use of the bio-databases did not help in task performance, and web search was used to find some other bio-database for better results.

On the semi-complex level (Table 6), there is much more concentration on some resources. Resource use is concentrated between the PC and BDB, in both directions. There are more transitions from web search engines to websites and from websites to PC. Table 7 shows the routine sessions, where most of the interaction (ca 65 %) is centered between bio-databases and the PC.

Table 4. Summary of transitions in all sessions, % (N=24, n=754)

	WE	WS	LDB	BDB	PC	O	
WE	0.0	5.4	0.2	3.4	0.4	0.0	
WS	2.5	0.0	1.2	2.0	4.4	0.2	
LDB	0.0	2.0	0.0	1.4	3.4	0.6	
BDB	3.7	2.3	2.1	0.0	24.6	2.1	
PC	2.5	2.1	4.8	25.2	0.0	0.4	
O	0.0	0.4	0.0	0.2	2.6	0.0	
							100.2

Table 5. Transitions in complex sessions, % (N=4, n=76)

	WE	WS	LDB	BDB	PC	O	
WE	0.0	4.7	0.0	13.6	0.0	0.0	
WS	5.0	0.0	4.7	3.8	2.7	0.0	
LDB	0.0	5.4	0.0	3.6	0.0	0.0	
BDB	10.6	5.4	7.6	0.0	12.5	2.3	
PC	2.7	0.0	0.0	13.6	0.0	0.0	
O	0.0	2.3	0.0	0.0	0.0	0.0	
							100.0

Table 6. Transitions in semi-complex sessions, % (N=9, n=242)

	WE	WS	LDB	BDB	PC	O	
WE	0.0	9.5	0.5	2.8	1.1	0.0	
WS	4.5	0.0	0.9	0.5	9.5	0.5	
LDB	0.0	2.0	0.0	0.0	3.5	0.5	
BDB	2.4	1.0	1.0	0.0	20.9	0.7	
PC	5.4	4.0	4.8	21.3	0.0	0.7	
O	0.0	0.0	0.0	0.0	2.4	0.0	
							100.5

Table 7. Transitions in routine sessions, % (N=11, n=436)

	WE	WS	LDB	BDB	PC	O	
WE	0.0	2.3	0.0	0.1	0.0	0.0	
WS	0.0	0.0	0.1	2.7	0.8	0.0	
LDB	0.0	0.8	0.0	1.8	4.6	0.9	
BDB	2.3	2.3	1.0	0.0	32.0	3.3	
PC	0.1	1.2	6.6	32.6	0.0	0.3	
O	0.0	0.0	0.0	0.5	3.7	0.0	
							100.0

One may summarize the findings on task complexity and the integrated use of channels and tools as follows: The more complex task, the more integrated and varied use of different channels and tools. The simpler or more routine, the more switching between PC tools and bio-databases, but the overall integration is not extensive. Routine tasks are often served by domain (and task) specific channels and tools, which are readily available and obvious choices. This corroborates empirically the view presented in [10, p77]. In complex tasks, useful channels and tools are less obvious and therefore more ad hoc exploration with generic tools like search engines is needed.

Table 8 exhibits the between channel transitions by channel: each row indicates the percentage of the target channels when the current (row) channel is left. We see, e.g., that the researchers go from search engines to web sites and bio-databases, seldom elsewhere, and from web sites back to engines (for more or better information) and the PC (with satisfactory information).

Table 8. Transition distribution by channels, %

(N=24, n=754)

	WE	WS	LDB	BDB	PC	O	Sum %
WE	0.0	50.0	4.8	40.5	4.8	0.0	100.0
WS	18.3	0.0	13.3	13.3	53.3	1.7	100.0
LDB	0.0	25.0	0.0	6.3	65.0	3.8	100.0
BDB	6.2	2.9	5.0	0.0	76.9	9.1	100.0
PC	4.7	5.3	20.7	68.0	0.0	1.3	100.0
O	0.0	3.3	0.0	10.0	86.7	0.0	100.0

3.3 The Query and Navigation Levels

Table 9 shows the queries on different session complexity levels and as summarized. We employ the classification of Figure 1. An overall tendency is that most of the queries are done to bio-databases (61 %). Queries, however, are distributed differently at different complexity levels. Most of the complex session queries are of type BF, for factual biological data. At the semi-complex sessions, most are of type BT, which are topical. This may indicate, that the information needs in the complex sessions are so fuzzy, that topical requests are not possible. The topic is unknown. On routine level there are lots of resource requests to bio-databases (BR 41.7 % and LR 22.3 %). Routine tasks include simple data resource gathering for further analysis, and literature gathering for e.g. database updating. In our study, most of the web engine queries fell into type navigational – this may be due to the search bar on the web browser and the huge number of databases available: book marking with the browser is not handy enough.

When the web searches are analyzed separately, there were 41 queries, of which 22 (53.4 %) are navigational. One fourth of queries are factual queries (24.4 %) and one fifth topical queries (19.5%). In all, Web queries add up to 14.2 % of the total number 289 of queries classified thus representing a narrow band of IA. Moreover, Figures 3-4 suggest that significant IA in other channels may intervene web searches even in quite narrow time frames. As the search goals are likely to change due to the intervening IA, it must be difficult to make sense of the goals in technical sessions in typical log analysis studies, in which sessions are defined by queries from the same IP address within a narrow time frame.

Usually the topics of the researchers' literature requests are quite narrow due to their specialized field. They already are aware of useful papers for citing (known items), and those were classified as fact finding. These are prominent at the semi-complex session level (17.5 %).

Furthermore, we looked at a sample of sessions from the link navigation perspective. The links followed in order to switch between resources were counted in the logs of 17 sessions. There were 125 links, of which one half (50.4%) were discovered from three sessions. The distribution between sessions was therefore not even. These three sessions were in the routine session category. Link navigation was, consequently, most common in the routine sessions (N=6), which included 52 % of all link navigations. In complex task sessions (N=4), there were 25.6 % of the link navigations, and in semi-complex (N=5) 22.2 %. Most of the followed links appear to be between bio-databases, meaning that links are inside the Bio-Database channel, since the transitions in the routine session level is concentrated between Bio-Database and PC.

Table 9. Distribution of query types in sessions of different complexity

Query type	Complex, % (n=53)	Semi-Complex, % (n=97)	Routine, %(n=139)	Number	%
WT	7.5	4.1	0	8	2.8
WF	0	9.3	0.7	10	3.5
WN	18.9	8.2	2.9	22	7.6
WR	1.9	0	0	1	0.3
LT	0	0	8.6	12	4.2
LF	7.5	17.5	0	21	7.3
LR	0	8.2	22.3	39	13.5
BT	7.5	35.1	7.9	49	17.0
BF	37.7	16.5	15.8	58	20.1
BR	18.9	1.0	41.7	69	23.9
Sum	100 %	100 %	100 %	289	100%

4. DISCUSSION

We begin by discussing the answers to the research questions, followed by broader consequences and limitations of the study.

First, our study shows that tasks and sessions in MM considerably differ in IA at different levels of task complexity. In routine tasks the task goals are clear and the process is straightforward. The task outcome is predictable so assessing the task result and success is easy. In semi-complex tasks assessing the outcome requires more effort. There is more variation in the information needs. Further, in complex tasks the task process is vaguer, and assessing the outcome can be seriously difficult. There are frequent revisits to resources already seen and the sessions tend to last longer.

The differences in task complexity show also at the interaction level. Tables 5-7 indicate that the more complex task, the more integrated and varied use of different channels and tools. Routine tasks concentrate on changes between two channels. There are sequences of mechanical repetitive actions. Complex sessions spread across several channels alternating between them.

At the query/navigation level we observe similar behavior with regard to task complexity. The needs behind the queries on the routine level are not informational, but resource needs. On the semi-complex and complex level intentions behind querying are more often topical or factual, although the factual literature queries (LF) are prominent on the routine level. The reasons may lie in the domain features: researchers in MM tend to seek literature in routine gene screening tasks. On the routine level there is more link usage, which indicates that these workflows are predictable, and the services already provide good integration supporting routine task performance.

Methodologically, the presented type of analysis of real life IA in work tasks in MM is challenging and time consuming, but at same time rewarding. The methods we used, triangulation, initial data filtering,

workflow charts, transition tables, and query intention classification, facilitate answering the research questions discussed above.

4.1 Integrated Access and Information Needs

Our findings indicate active integration of many resources across different channels, in relatively short time-frames, when accessing information for work tasks. The task performer integrates these channels and resources in the task process. A part of this integration is semiautomatic in nature. For example, a biological fact database may contain, for some bio-molecular object, hard-wired links to literature in PubMed – by clicking the link, one gets immediate access to the PubMed record of relevant literature. One still needs to click. Full automation would mean automatic (implicit) collection and aggregation of relevant bits of information of different types from various sources for the task performer to view – a kind of multi-source aggregated search. The manual alternative is that there is no link; the task performer needs to copy the reference information and explicitly search for the known item in another channel (here PubMed). The bottom-line manual alternative is that the bio-database gives no reference at all and thence the task performer needs to make a topical search in the other channel (PubMed). We believe that real-life work task sessions are full of all classes of integration from the bottom-line type to full automation.

An information need in task-based IA is a function of task requirements and of (lack of) integration of information systems: a (automated) routine look-up may turn into a topical search if integration is lacking. In the bottom-line alternative, the only *handle* [10, p. 328, p. 355-56] to required information is factual/topical (e.g., a gene name), thus a topical search. In the other manual alternative, the handle is a reference, thus a known-item search. Work task goals determine *genuine information needs*, but the task process and the degree of integration between channels and resources greatly affect *instrumental information needs*, which may appear (in a log) in all possible disguises depending on the handles the task performer has available and the systems can employ. When the systems or the processes change, the instrumental needs are likely to change as well. This is an example of the dynamic interaction between task goals, performer, available information and information systems. Changing any component causes changes in the other and in their interaction [10, p. 77-78].

A significant share among the transitions between the PC and the other channels (bio-databases in particular) are (a) capturing some bits of information from the latter to tools on the PC and (b) transforming the captured information on the PC for input to another channel/resource. The reasons are (a) that sometimes no individual resource covers all items needed as input to a given resource, and (b) that the supplied data are incompatible. This is a natural consequence of distributed and autonomous production of the resources. Data incompatibility may be due to syntactic or semantic heterogeneity. The former can be solved by the increasing adherence to XML in data exchange. Semantic data heterogeneity means that, in two or more data sources, data with the same intended meaning are represented in varying ways and/or data with different intended meaning are represented in identical ways. This is more challenging to solve (see Section 4.3).

Our findings suggest that, in the interest of improving task performance, the resources for IA should not be designed in isolation, because their use is integrated and bad design results in increased manual integration efforts. If the resources were more directly integrated from the perspective of the observed tasks in a simple and usable way, the sessions might be significantly affected – for the better.

4.2 Log Analysis

Several recent studies analyze query logs seeking to capture users' information needs and intents, e.g., [19]. While this may be successful in the generic web domain and for popular queries, we argue that in

work task contexts the users' information needs are only captured by closely examining their work tasks and entire information environment. The information captured solely on the queries in a single channel, say search engines, tells quite little about the users' intentions. For example, a gene name abbreviation can be used in all the query classes we discussed: from WT (search engine/topical) to BR (bio-database/resource). Moreover, by logging any single channel queries (and clicks), i.e. server-side logs, one obtains only a narrow window to IA. For example, Search engine queries add up to only 14.2 % of the total queries in our data. There may be much intervening IA in other channels between two queries in any single channel (see Figs 2-4). Therefore it must be difficult to make sense of technically defined sessions – queries from the same IP address within a narrow time frame – in task-based IA.

The role of instrumental information needs and the consequent handles further increase the difficulty of mining user intentions in server-side logs.

4.3 Consequences, Limitations, and Further

Our findings suggest several consequences for information (retrieval) systems design from the work task perspective. When some resources and channels are frequently integrated in IA, it would make sense to strive away from manual integration, performed by the user, which is due to bad system-level integration and consequent instrumental needs. We see four possible strategies here:

- Implicit and automatic aggregation of likely relevant information from across resources. This is alike aggregated search in the Web, but now across task based resources.
- Hard-wired integration of resources through direct linking. This is already an ongoing activity in MM, e.g., bio-databases at NCBI often link directly to PubMed. Such activities should be continued.
- Provision or suggestion of task-relevant possible handles for access when neither of the former is available as proposed by Bates [3]
- Data harmonization at the levels of its representation, naming and structuring [20]. Here data in different resources can be, in part, made directly compatible by cross-domain standardization, and in part harmonized at access time.

Semantic data heterogeneity may be due to [20]:

- the representation of data (e.g., metric vs. other units; textual values in different languages)
- naming conventions of data (e.g., attribute names or other data labels for similar data differ)
- data structuring (e.g., varying formats for bibliographic information; alternation between data as values (instance level) and data as labels (schema level)).

The purpose of data harmonization is to remove such heterogeneous factors when data are extracted from separate resources, that is, the same information is represented in one uniform way. This supports manual resource integration, facilitates hard-wired linking and makes automatic aggregation of information possible.

IA is traditionally seen as an individual activity, as an individual using systems. Collaboration has been proposed as a more realistic viewpoint to IA (e.g. [8]). Here humans are seen to collaborate either directly (face-to-face) or indirectly (via systems and documents). Perhaps this view should be extended. As IA systems are autonomous and becoming increasingly knowledgeable and active, it seems increasingly important to view them as collaborating independent actors (even if designable) in task processes. This conception is similar to Sonnenwald's [24] work on information horizons, where information resources may be viewed as reflexively collaborating among each other.

There are some limitations in our study. We analyzed a limited number of sessions performed by a limited number of people within a specific domain and institution. The specific findings cannot be generalized to other contexts. However, these data are sufficient to show that channel/resource integration and interleaved access are essential, that the process or task perspective is required to make sense of logs at any server side, and that varied search goals are pursued through varied query types.

Overall, the findings help to understand real-life IA in a specific context, molecular medicine. While the specifics are bound to change from domain to domain, institution to institution, and researcher to researcher, our findings indicate active integration of many resources across different types of channels in task-based IA. It is highly unlikely that another sample of MM research tasks would suddenly show that searching would be limited to the Web or PubMed, or that task sessions would focus on a single channel and tool for longer time periods. The real world perspective brings forward important knowledge that could not have been learned in more controlled environments. This kind of methodology is laborious and slow, and uses relatively small sample of real life IA phenomena, but it is rich in detail and highly informative compared to some easier and quicker methods, such as log mining techniques. It may also inform what real life aspects, such as multiple systems, should be incorporated in laboratory experiments.

5. CONCLUSIONS

Task-based information access is a significant context for developing information retrieval systems. Molecular medicine is an information intensive domain with high profile research topics. There are numerous generic and domain-specific tools and databases available for online information access and this offers a fruitful soil for IR studies. Research in information retrieval aims at enabling access to information for supporting purposeful action, which is, in the present study, the work task performance. We present a methodology to task based approach of IR and provide results on three levels. Firstly, the work tasks are analyzed in a real work environment and at three complexity levels. Secondly, we show that interaction between different information channels increases proportionally to the complexity increase. Thirdly, we show that, similarly, the queries are more concentrating on resource level in routine tasks, but the prominence of factual and topical queries increases in complex tasks. In task-based information access, interaction logging at any single channel (like a search engine) may give a distorted picture of the searcher's needs and intentions. Therefore, the contribution to system development is that it should not be done in isolation as there is considerable interaction between systems in real world use. Significant benefits may be achieved by taking this into account in system design. These can be achieved through (a) implicit and automatic aggregation of likely relevant information from across resources, (b) hard-wired integration of resources through direct linking, (c) provision of task-relevant possible handles for access, and (d) harmonization of semantic heterogeneity of similar data across resources.

6. ACKNOWLEDGEMENT

This research was supported by the Academy of Finland Grants 124131 and 133021.

7. REFERENCES

[1] Baeza-Yates, R., Calderón-Benavides, L., and González-Caro, C. 2006. The intention behind web queries. In String Processing and Information Retrieval, 98–109.

[2] Bartlett, J.C. & Toms, E.G. 2005. Developing a protocol for bioinformatics analysis: An integrated information behavior and task analysis approach. Journal of the American Society for Information Science and Technology, 56(5), 469–482.

[3] Bates, M.J. 1990. Where should the person stop and the information search inter-face start? Information Processing & Management, 26(5): 575–591.

[4] Broder, A. 2002. A taxonomy of web search. SIGIR Forum, 36(2):3–10.

[5] Byström, K. and Järvelin, K. 1995. Task complexity affects information seeking and use. Information Processing & Management, 31(2):191–213.

[6] Cochrane, G. R. and Galperin, M. Y. 2010. The 2010 nucleic acids research database issue and online database collection: a community of data resources. Nucleic Acids Research, 38(suppl_1):D1–4.

[7] Downey, D., Dumais, S., Liebling, D., and Horvitz, E. 2008. Understanding the relationship between searchers' queries and information goals. In CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management, pages 449–458, New York, NY, USA. ACM.

[8] Hansen, P. & Järvelin, K. 2005. Collaborative information retrieval in an information-intensive domain. Information Processing & Management 41(5): 1101–1119.

[9] He, D., & Göker, A. 2000. Detecting session boundaries from Web user logs. In Proceedings of the BCS/IRSG 22nd Annual Colloquium on Information Retrieval Research, pp. 57–66. Cambridge, UK.

[10] Ingwersen, P. & Järvelin, K. 2005. The turn: integration of information seeking and retrieval in context. Heidelberg: Springer.

[11] Jansen, B. J., Spink, A., and Taksa, I. 2009. Handbook of research on web log analysis. Information Science Reference Hershey, PA: IGI Global.

[12] Jansen, M. B., A. Spink, J. Bateman, and T. Saracevic. 1998. Real life information retrieval: a study of user queries on the web. SIGIR Forum, 32(1):5–17.

[13] Kellar, M., Watters, C., and Shepherd, M. 2007. A field study characterizing web-based information-seeking tasks. Journal of the American Society for Information Science and Technology, 58(7):999–1018.

[14] Kelly, D. and Belkin, N. J. 2004. Display time as implicit feedback: understanding task effects. In SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. AMC, New York, NY, USA, 377–384.

[15] Kumpulainen, S., Järvelin, K., Serola, S., Doherty, A., Byrne, D., Smeaton, A. F & Jones, G. F. J. 2009. Data collection methods for analyzing task-based information access in molecular medicine. MobiHealthInf 2009 - 1st International Workshop on Mobilizing Health Information to Support Healthcare-related Knowledge Work, Porto, Portugal, January 16, 2009.

[16] Lin, J. and Wilbur, W. J. 2009. Modeling actions of PubMed users with n-gram language models. Information Retrieval., 12(4):487–503.

[17] MacMullen, W. J. & Denn, S. O. 2005. Information problems in molecular biology and bioinformatics. Journal of the American Society for Information Science and Technology, 56(5), 447–456.

[18] McDonald, S. 2005. Studying actions in context: a qualitative shadowing method for organizational research. Qualitative Research, 5(4):455–473.

[19] Murray, G. C. and Teevan, J. 2007. Query log analysis: social and technological challenges. SIGIR Forum, 41(2):112–120.

[20] Niemi, T. & Näppilä, T. & Järvelin, K. 2009. A relational data harmonization approach to XML data. Journal of Information Science. 35: 571–601.

[21] Roos, A., Kumpulainen, S., Järvelin, K. and Hedlund, T. 2008. Information environment of the researchers in molecular medicine. Information Research, 13(3) paper 353.

[22] Rose, D. E. and Levinson, D. 2004. Understanding user goals in web search. In WWW '04: Proceedings of the 13th international conference on World Wide Web. ACM, New York, NY, USA, 13–19.

[23] Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. 1999. Analysis of a very large web search engine query log. SIGIR Forum, 33(1):6–12.

[24] Sonnenwald, D.H. 1999. Evolving perspectives of human information behavior: Contexts, situations, social networks and information horizons. In T. D. Wilson & D. K. Allen (Eds.), Exploring the Contexts of Information Behavior: Proceedings of the Second International Conference in Information Needs (pp. 176-190). London: Taylor Graham**.**

[25] Stevens, R., Goble, C., Baker, P. & Brass, A. 2001. A classification of tasks in bioinformatics. Bioinformatics, 17(2), 180–188.

[26] Teevan, J., Dumais, S. T., and Liebling, D. J. 2008. To personalize or not to personalize: modeling queries with variation in user intent. In: SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, New York, NY, USA, 163–170.

[27] Toms, E. G., 2002. Information interaction: Providing a framework for information architecture. Journal of the American Society for Information Science and Technology 53 (10), 855–862.

[28] Vakkari, P. Task-based information searching. 2003. Annual Review of Information Science and Technology (ARIST), 37, 413–64.

[29] Xie, H. I. 2000. Shifts of interactive intentions and information-seeking strategies in interactive information retrieval. Journal of the American Society for Information Science, 51(9):841–857.