Evaluating federated search tools: usability and retrievability framework

Khaled A. Mohamed
Department of Library and Information Science, Fayoum University, Fayoum, Egypt, and
Ahmed Hassan
School of Engineering, Ain Shams University, Cairo, Egypt

Received 4 December 2013
Revised 7 July 2014
30 September 2014
Accepted 9 December 2014

Abstract

Purpose – This study aims to explore a framework for evaluating and comparing two federated search tools (FSTs) using two different retrieval protocols: XML gateways and Z39.50. FSTs are meta-information retrieval systems developed to facilitate the searching of multiple resources through a single search box. FSTs allow searching of heterogeneous platforms, such as bibliographic and full-text databases, online public access catalogues, web search engines and open-access resources.

Design/methodology/approach – The proposed framework consists of three phases: the usability testing, retrievability performance assessment and overall comparison. The think-aloud protocol was implemented for usability testing and FSTs retrieval consistency, and precision tests were carried out to assess the retrievability performance for 20 real user queries.

Findings – Participants were directed to assign weights for the interface usability and system retrievability importance as indicators for FST evaluation. Results indicated that FSTs retrievability performance was of more importance than the interface usability. Participants assigned an average weight of 62 per cent for the system retrievability and 38 per cent for interface usability. In terms of the usability test, there was no significant difference between the two FSTs, while minor differences were found regarding retrieval consistency and precision at 11-point cut-off recall. The overall evaluation showed that the FST based on the XML gateway rated slightly higher than the FST based on the Z39.50 protocol.

Research limitations/implications – This empirical study faced several limitations. First, the lack of participants' familiarity with usability testing created the need for a deep awareness and rigorous supervision. Second, the difficulties of empirically assessing participants' perspectives and future attitudes called for mixing between a formal task and the think-aloud protocol for participants in a real environment. This has been a challenge that faced the collection of the usability data including user behaviour, expectations and other empirical data. Third, the differences between the two FSTs in terms of number of connectors and advanced search techniques required setting rigorous procedures for testing FSTs retrieval consistency and precision.

Practical implications – This paper has practical implications in two dimensions. First, its results could be utilized by FST developers to enhance their product's performance. Second, the framework could be used by librarians to evaluate FSTs performance and capabilities. The framework enables them to compare between library systems in general and FSTs in particular. In addition to these practical implications, the authors encourage researchers to use and enhance the proposed framework.

Social implications – Librarians can use the proposed framework to empirically select an FST, involving users in the selection procedures of these information retrieval systems, so that it accords with users' perspectives and attitudes and serves the community better.

The Electronic Library
Vol. 33 No. 6, 2015
pp. 1079-1099

0264-0473
DOI 10.1108/EL-12-2013-0211

Originality/value – The proposed framework could be considered a benchmark for FST evaluation.

Keywords Information retrieval, Usability testing, Federated search, Meta search, Think-aloud protocol

Paper type Research paper

Introduction

In the era of digital libraries, bibliographic and full-text databases have become common spaces for information resources searching and discovering, providing functionality and ease of use that are superior to printed products (Burke, 2001). The World Wide Web has introduced many advantages for databases and library searching tools as well. One of the main advantages of this environment is that it integrates many information resources and makes them available for searching from one location. Some of these resources are available on the surface web for public users and are able to be reached without any authentication requirements, while others, especially bibliographic databases, require user authentication to enable access to the deep web, which hide, most of the time, behind web scripts (Liu *et al.*, 2012).

Each one of these databases has a unique interface and searching capabilities. Searching different databases requires a good understanding, on the users' part, of the coverage and searching capabilities of each database, along with a perception of the area of interest and how to fully use the database's capabilities. Additionally, when switching databases, there is the difficulty of moving from one interface to the other. These difficulties have made the searching process tedious and time-consuming, rather than easy and time-saving. These challenges often push library users to complain about the complexity of database searching and compare it with web searching, such as World Wide Web search engines (Google, for example). The simplicity of web search engines (SEs) drive users to ask for a single search box that is able to aggregate and discover everything (Boyd *et al.*, 2006; Burke, 2001). The fact that SEs have succeeded in searching the surface web, but cannot easily handle the deep web resources, motivates researchers and information retrieval (IR) developers to look for new solutions for searching and discovering the deep web (Mohamed, 2006).

Deep and dynamic web resource searching requires special platforms different from those used for searching the surface web, to facilitate access to information resources hidden behind database interfaces. These unique methods of web searching lead to new platforms different from traditional surface web searching tools, such as directories, SEs and meta-search tools, because of their limitations in crawler technology and authentication requirements. On the other hand, discovering the deep web and searching multiple platforms requires extensive authentications and customization mechanisms from system administrators, as well as users. Although a variety of terms are used to refer to the concept of deep and multiple platform web searching, such as cross-database searching, meta-searching, aggregate search, library portal, and others, the term *federated search* is commonly used to refer to that technique by researches and practitioners (Tchangalova and Stilwell, 2012). One of the practical and applicable solutions for the problem described earlier has been the development of federated search tools (FSTs) (Avrahami *et al.*, 2005; Jacso, 2003).

Using FSTs as the main platform of resource discovery requires the following:

- generating a query from a unified interface and broadcasting it to a group of disparate databases (a predefined set of databases and/or SEs) with the appropriate syntax, which requires some form of query translation;
- combining the results aggregated from different sources and/or databases using data fusion and combination techniques (Dwork *et al.*, 2001; Kumar *et al.*, 2008);
- presenting the results in a succinct and unified format with minimal or no duplication (often called de-duplicating or de-duping); and
- providing a means, either automatically or by the portal, to sort the merged result set (Boyd *et al.*, 2006; Tennant, 2001).

All of these procedures have brought as many challenges as promises to web searching, especially in terms of adapting these systems to user needs and the severe impact that this new mode of searching has had on users' search behaviours (George, 2008).

Fusion and aggregation of information are major problems for all kinds of IR systems, from text and multimedia processing to decision-making (Yuwono and Lee, 1996). Nevertheless, there are two general approaches to this scheme, depending on the problem to be dealt with (Tsikrika and Lalmas, 2001). The first approach corresponds to the aggregation of preferences given by several individuals of a group or the aggregation of criteria to satisfy specific needs to make a decision. The second approach corresponds to the fusion of evidence provided by several sources. In many cases, the available information is imperfect; therefore, several methodologies and/or theories have been used to manage this imperfect information. Among the most important ones are probability theory, evidence theory, fuzzy set theory and possibility theory (Bouchon-Meunier, 1998). Many of these theories have been used to develop IR systems to allow developers to explore different models, including Boolean, vector space, probabilistic and language models. There are many tools available for federated searching utilizing these techniques to facilitate cross-searching domains worldwide. Some of them are turnkey systems, such as MuseGlobal, Summon, Endeavor, Ex Libris-MetaLib, WebFeat, DQM2, Deep Query Manager and Fretwell-Downing, while others are open source items, such as dbWiz search and OpenSiteSearch, among others. Some of the turnkey systems are developed to be standalone systems and utilize portal technologies and XML gateways, such as MuseGlobal and WebFeat, while others are developed as modules and parts of library management systems using Z39.50 protocols to build their connectors, such as MetaLib from Ex Libris and ENCompass from Endeavor.

There are two major retrieval protocols used by FSTs to facilitate multiple database searching: XML gateways and Z39.50 protocol (Abercrombie, 2008). The availability of these different protocols motivated researchers to evaluate these new tools and their baseline technology to help enhance the products and to support librarians in the selection of the appropriate tools. This study develops a framework for comparing two FSTs that use these two different retrieval protocols. The first FST was developed by MuseGlobal (www.museglobal.com) and based on portal technology. This FST utilizes XML gateways to create connectors, using search solver and link solver components, which allow for OpenURLs to be automatically updated. The second FST, developed as part of the Egyptian universities library management system, is based on an ordinary

web-based application and Z39.50 protocol to connect to the native database programmatically. Therefore, it requires database and electronic resource connectors and OpenURLs to be manually updated. The implication of this study would allow for FSTs to enhance and provide a framework for librarians and portal administrators to evaluate and compare different FSTs. The framework of the comparison matches the development scenario of FSTs, as they face two serious challenges. The first challenge is to make the search process intuitive, simple and easy. This process is tested and evaluated in the literature through usability techniques by users or experts. The second challenge regards retrieving appropriate resources, which are mainly defined by the source databases and tested by the well-known IR performance measurements. This study will develop a framework for evaluation and use it to evaluate FSTs in terms of usability and IR performance to provide a solid evaluation benchmark.

Literature review

This section presents the related literature for the current study. The review will cover FST studies and their evaluation techniques. Federated search systems have been evaluated and tested using different techniques, including user evaluation and IR performance assessment. The topic of federated search has seen great interest since the beginning of the new millennium. A number of these studies have been performed to scientifically measure the impact of federated search, like Fryer (2004), Lu and Callan (2005) and Si and Callan (2005).

A number of studies have focused on the usability of FST as a major approach for evaluation. Usability testing is one of the well-established and widely used techniques for evaluating user acceptance to library system tools, such as online public access catalogue (OPAC), online databases and library portals. Accordingly, it is a user-centred approach utilized for evaluating users' perspectives, attitudes and levels of satisfaction. Usability provides the investigators with rich quantitative and qualitative data to support their research practices. Usability testing involves gathering information about user behaviour while interacting with information systems. Studies frequently include five common indicators to evaluate a system from the users' perspectives: easy to learn, efficient to use, easy to remember, few errors and satisfaction (Mohamed and Hassan, 2008; Nielsen, 1992).

George (2008) conducted a usability study using think-aloud protocols to evaluate the MetaLib federated search system, which was developed by Ex Libris to be used as a library portal. A demographic questionnaire was distributed to a selected sample of eight volunteers, diverse with respect to affiliation, discipline, gender, language and computer experience. This study showed that participants faced difficulties in the following processes: system login, primary and secondary navigations, confusing terminology and inconsistency with site design and use expectation. Another important study evaluated the usability of one of the federated search interfaces used by the University of Maryland and its affiliated institutions (Wrubel and Schmidt, 2007). It investigated students' perceptions of the search system usefulness and the extent to which students could effectively complete search tasks using the federated search. Students perceived federated search as a useful tool, but they had low rates of success in completing some tasks (Wrubel and Schmidt, 2007). Cervone (2005) carried out a usability test for evaluating the library portal of Northwestern University in Evanston, IL. This study focused on the usability testing of open URL resolvers and FSTs.

Randall (2006) investigated the usability of the Endeavor Information System by gathering insights on the critical requirements of libraries and information professionals about the federated SEs capabilities which is a part of their integrated library system. The results of the study, conducted in conjunction with market research and consumer focus groups, provided insight into the relevancy of the Endeavor federated search technology and user needs and behaviours.

Chen (2006) described the features and capabilities of library federated SEs. He compared MetaLib and WebFeat as research tools by highlighting their strengths and weaknesses against Google and Google Scholar. MetaLib and WebFeat had fundamental differences and could not compete with Google in terms of speed, simplicity, ease of use and convenience, nor were they truly one-stop shopping. However, their strengths lay in the contents they searched, as well as in the objective way they retrieved and displayed results.

There are a few studies focused on evaluating federated SEs as IR systems. Avrahami *et al.* (2005) carried out a series of experiments to develop a prototype federated search system for the USA government's state web portal and addressed the issues in adopting research solutions for this operational environment. The series of experiments identified how well and important previous research results carried out for this purpose, including parameters settings and heuristics, applied in the state's environment. The study concluded with a set of lessons learned from this technology transfer effort, including observation about SE quality in the real world. Lampert and Dabbour (2007) carried out three assessment projects: two of them focused on the reaction of librarians to meta-search technologies from a reference and information literacy perspective and a third project using a survey that captured students' experiences, understanding and satisfaction with meta search at California State University at Northridge, CA. The authors presented an investigation of users' understanding of MetaLib's Combined Search (MCS) which is the federated search system implemented in Washington Research Library Consortium. Data show that students considered MCS primarily as a tool for locating full text, while librarians viewed it as a secondary search tool with disappointing performance. In discussing MCS's operation, students focused largely on its full-text retrieval capability and search efficiency, whereas librarians paid more attention to search strategies and retrieval quality.

In a recent study, Buck and Nichols (2012) explored a participatory design strategy to investigate user and librarian views on what a discovery system should look like and what functionalities it should have. The findings revealed that librarians think important features for these tools include navigational, searching and filtering capabilities. The authors also discussed the phases of the participatory design process. Jaffe and Mukherjee (2013) designed a system and method for dynamic context-sensitive federated search of multiple information repositories based on groupings of search results which included a plurality of labelled groups with a plurality of search results in each group. A sub-grouping of search results was generated for each labelled group of search results in the main grouping of search results using the second set of attributes.

Georgas (2013) investigated undergraduate student preferences and perceptions when using both Google and an FST. Students were asked to evaluate each search tool in terms of their preference and the perceived relevance of the retrieved results when using each search tool. Students were also asked to self-assess their searching skills. The

findings show that students believed that they possessed strong searching skills, were able to find relevant sources using both search tools and actually preferred the FST over Google for doing research. Thus, despite federated searching's limitations, students saw the need for it and indicated that libraries should continue to use federated search to provide access to their resources (especially if a discovery search tool is not available) and, accordingly, librarians should focus on teaching students how to use both federated search and Google more effectively.

Some other recent studies focused on the aggregated search mechanism and its impact on enhancing the retrievability and results presentation. For example, Kopliku *et al.* (2014) proposed an analysis framework for aggregated search and an overview of existing work about related domains, such as federated search, natural language generation and question answering. Bron *et al.* (2013) investigated whether user preference for source presentation changes during a multi-session search task when using aggregated search interfaces. Arguello *et al.* (2013) examined what factors affected aggregated search coherence and search behaviour while using media (images and video) or a specific type of search task (news and shopping) vertical SEs.

The related literature proves the importance of the topic and the need for a solid framework for evaluation and comparison. Therefore, this study proposes a framework relevant for evaluating and comparing two or more FSTs. The evaluation technique utilizes the same methodologies and indicators used by IR researchers, including usability testing and IR performance assessment with some adaptions to the FST's nature and capabilities. For example, retrieval consistency and 11-point cut-off recall have been used to evaluate the retrievability of FSTs. The simple usability test has been exploited using the think-aloud protocol, as the most relevant engineering technique to evaluate user acceptance in terms of perception and future attitude.

Research objective, design and methodology

Selecting the appropriate FST to use for searching multiple native databases and open-access resources requires succinct evaluation of the available tools based on simple and applicable methodology. Thus, this study has set up a framework for evaluating and comparing between two widely used FSTs in Egypt in terms of systems usability and retrievability performance. The first FST, known as Muse Portal (www.eul.edu.e.g.), developed by Muse Global, is based on portal technology using XML gateways for building connectors to allow searching multiple native databases. It contains two major integrated components, a search resolver and a link resolver. The second FST is a significant part of an Egyptian integrated library management system, known as Future Library System, which was deployed in Egyptian universities to enhance the automation of academic libraries and building theses repositories. It utilized the Z39.50 protocol for building the search connectors. The two FSTs are heavily used in Egyptian universities, where some universities are using the first one and others are applying the second one as the single access point for the Egyptian universities' electronic resources, including international databases, OPACs and electronic theses. Therefore, it is necessary to develop a framework for comparing and evaluating the important distinctions between the two FSTs in terms of user acceptance and system retrievability. This part would focus on identifying the research questions, methods of testing and data collection procedures.

Research questions

To set up the framework and test the system usability and retrievability, two simple research questions were formulated:

RQ1. What are the major evaluation criteria of FSTs in terms of usability and retrievability performance?

RQ2. What are the major differences between the two FSTs in terms of usability and retrievability performance?

To answer these two questions, a hypothetical framework for the evaluation procedure has been designed and tested in a real environment. The researchers assume that an FST based on portal technology using XML gateways would perform better than an FST based on library management system using the Z39.50 protocol in terms of usability and retrievability. The contribution of this paper is significant for several reasons. First, it proposes a new combined technique that uses usability testing and IR performance assessment as major indicators for FST evaluation. Second, it highlights users' preferences in terms of weighting the importance of each indicator. Finally, it compares an FST based on portal technology and XML gateways with another based on a library management system and the Z39.50 protocol.

Research design

Due to the large number of tasks and complex IR evaluation techniques that are used in this study, in addition to the tendency to fully investigate all the usability and retrievability problems, the investigators used a reasonable sample of 20 participants who have a solid background in IR. The participants are graduate students in the department of library and information sciences in Fayoum University, Egypt.

The participants have succinctly examined the two FSTs' usability and retrievability under the supervision of the investigators during a special topic in IR postgraduate course. Participants were recruited for a three-hour session, carried out on two different days: the first day for an awareness and usability test, while the second day was devoted to the retrievability assessment procedures and evaluation techniques.

The design of the study includes a preliminary investigation and three research phases: usability testing, retrievability assessment and overall evaluation. In the preliminary phase, the evaluation criteria were designed and discussed with the participants. In this phase, participants were directed to assign weights for the evaluation criteria, including interface usability and system retrievability as the main indicators for evaluating FSTs. A short questionnaire of three questions was designed and completed by the participants to indicate the importance of each indicator (usability, retrievability and any other indicators from the perspective of the participants) and then participants were directed to answer the following questions:

Q1. Have you ever used FSTs and are you aware of their techniques of searching?

Q2. Assign a value out of 100 per cent for the importance of the interface usability and system retrievability as indicators for evaluating FSTs.

Q3. From your point of view, are there any other important indicators for evaluating FSTs?

All participants reported that they had positive experiences using FSTs to find information resources. In all, 15 participants stated that they have used an FST at least once to find information for research purposes. Participants stated that they do not have enough information about FSTs techniques. Out of all, 18 participants indicated that FST retrievability is more important than interface usability, because they can discover the hidden features and learn how to use FSTs by receiving appropriate training, but cannot improve the system retrievability. Two participants indicated that they had no preference. Participants assigned an average weight of 62 per cent for FTS retrievability and 38 per cent for usability testing. Participants included some features of usability, including searching and browsing capabilities as other indicators, and the final analysis proves that all their comments were available in other features of the usability evaluation sheet. All the participants indicated that retrieval consistency and precision were the most important indicators for evaluating FST retrievability. Three participants noted the importance of results presentation and format of metadata as indicators for evaluation, but, due to the huge number of tasks, they finally consented to use retrieval consistency and precision for evaluating FST retrievability.

Participants then attended the awareness session that included detailed demonstration about the major concepts of meta and federated searching and how users can fully and deeply utilize the examined tools. After this preliminary phase, the design of the evaluation procedures was carried out and divided into three major phases: usability testing, system retrievability and overall evaluation.

Usability testing

Usability testing involves including real users as participants in the test. This evaluation investigates the most appropriate and important usability indicators that could be utilized by librarians to evaluate FSTs. The usability test primarily focused on the design of the interface, intuitiveness and other usability testing parameters, including accessibility, availability, accuracy, ease of use and learnability.

The usability phase was divided into two sessions, one for participants' awareness and learning and the other for usability task implementation. In the first phase, participants were given an awareness session for 45 minutes about how to use the two FSTs and directed to report on a topic of interest to be used for searching. Participants were advised to select topics in the area of library and information sciences to be able to evaluate the relevancy of the retrieved results.

The think-aloud protocol was used to gather participants' evaluation of the two FSTs, including the most common and critical features. These features included system navigational, searching and browsing capabilities. The think-aloud activity provided a model for users through asking them to speak out loud during task completion to verbalize what they are doing and thinking. Nielsen (1992) reported on the importance of this protocol by saying "thinking aloud may be the single most valuable usability engineering method". Participants were thus directed to use the two FSTs and speak aloud about what they were thinking, recognizing, perceiving and expecting to find, in order to investigate their perceptions and future attitudes.

Combining a formal task and think-aloud protocol for participants in a real environment allows for collecting usability data about user behaviours, expectations and other empirical data. A predefined short data collection form was designed to collect empirical data during the think-aloud protocol which will be discussed further in the

results and discussion section. Approaches to usability testing may vary, but the think-aloud protocol mixed with a structured open-ended questionnaire, where participants use FSTs and describe their experience out loud, responding to the required tasks and reporting their evaluation in a structured format, is a new technique demonstrated in this study.

The think-aloud protocol has been used to enhance the results of the analysis. All responses were compared with the recorded sessions to verify that the responses agreed with the recorded sessions and to fill in the incomplete responses. No recorded sessions were found to be incompliant with the completed responses. This result shows the capabilities of the participants and their consistency. Thus, this mix of techniques provides a balance between efficiency and quality of data collection. In the think-aloud session, participants were instructed to use the FST's interfaces to test what they have learned about the major components of FSTs in the awareness session, related to what they perceived and expected. Participants were also asked about their future attitudes towards the two tools.

The usability test included four groups of questions about the interface capabilities. Participants were directed to assign a value out of three for each question and to speak aloud when they decided which value she/he would assign to the evaluated parameter during the task and to explain their reasons behind the assigned score. The scoring system was based on a four-value satisfaction Likert scale: (3) satisfied, (2) partially satisfied, (1) not satisfied and (0) not available. The total value for each question is three which was multiplied by the number of participants (for example, 3×20 participants $=$ 60). The total usability value for each FST was calculated and normalized out of the 38 per cent preliminary indicator in the overall evaluation. Therefore, for example, if all the participants assigned a score of 2 (partially satisfied) for a specific item, then it would gather 40 points ($2 \times 20 = 40$ out of 60); if all the participants responded positively and select satisfied (a score of 3) for the specific item, then it would generate 60 points ($3 \times 20 = 60$ out of 60). These are the type of scores shown in the final evaluation sheets.

Retrieval performance assessment

FST retrieval performance for 20 real user queries was tested using retrieval consistency and precision as major evaluation indicators. Retrieval consistency of the two FSTs was compared using the total number of the retrieved items from the same set of databases and, accordingly, precision was assessed in terms of precision at 11-point cut-off recall values. This part of the study was conducted after the usability test in a separate session, as participants became fully aware of the features and capabilities of the two FSTs. Participants were directed to select queries and structure them as a title search for the sake of simplicity and accuracy of relevancy judgments. Therefore, the search terms were chosen to describe concepts and keywords only appearing in the title field, as a search for a generic topic would not be easy to judge the relevancy and would complicate the process. The concept selection was controlled according to the following conditions:

- Queries should represent real user needs for evaluation purposes.
- Queries should be simple (maximum of two or three terms that represent topics relevant to a title field search) to retrieve results that are reasonable to be analysed and evaluated.

In this phase, each participant was directed to search for the required query using the advanced search interfaces and submit the query to the title field. Participants assessed the retrievability performance of the FSTs using four specific and unified databases and reported the results in an Excel spreadsheet, including the number of items retrieved and the relevancy of the first ten items retrieved from each FST. To calculate the retrieval consistency, participants searched the native interfaces of the selected databases using the same query structure and field. This would ensure that, if the same query runs in the two FSTs and the native interfaces of the selected databases, it would probably retrieve the same list of items with different rankings. This concept is important to ensure that the same documents and results obtained by searching the native database are also obtained by using the FST. Precision of the retrieved results was calculated according to a three-point relevancy scale (0 – irrelevant, 0.5 – partially relevant and 1.0 – relevant) (Voorhees, 2000). Retrieval consistency of the two FSTs was compared with the native database results, and precision at 11-point cut-off recall was also assigned a value for assessing the total score of the retrievability. The total value of the retrievability indicators was calculated and interpolated to the average of the retrieval consistency and precision. This means that items were given an equal score of 31 per cent for each, out of the assigned 62 per cent total value that has been allocated for the retrievability performance indicator in the preliminary phase.

Overall evaluation

The overall evaluation value was calculated for each FST by summing up the value of the usability to the value of the retrievability to assign a grand total score for each FST. These baseline values for each indicator were taken into consideration the score of each phase assigned by the participants (62 per cent for retrievability and 38 per cent for usability).

Results and discussion

The following section presents and discusses the results of each phase, including: usability testing, retrievability assessment and overall evaluation.

Usability testing

The think-aloud protocol was used to explore participants' perceptions and future attitudes of the two FSTs. The difference between perception and future attitude is that: perception is how people look at something, while attitude is the way they act towards something in the future. The authors selected perception and future attitude to consider the cases where perception is negative and the participant was not comfortable with the interface capabilities and functionalities, although the participant still had the willingness to continue using the system in the future if appropriate training is provided. The awareness session was audio recorded with participants divided into two groups of ten participants in each. Participants were directed to use the two FSTs to expend 15 minutes for each and to record their responses for each item, while executing a predefined set of tasks in the investigation session. Participants were instructed to speak aloud to effectively record their responses, indicating their impressions, perceptions and expected future attitudes for the three major tasks, as displayed in Tables I-IV, including navigation, searching, browsing and other comments. The audio records involved capturing the verbal feedback of the participant's evaluation for each part of the procedures. A predefined think-aloud protocol evaluation sheet was designed

that included all the expected responses. The usability evaluation sheet included two parts: one for perception and another for future attitudes. Each part (perceptions and attitudes) was assigned a score out of 60 for each item according to the previously explained Likert scale.

The participants were directed to record their usability responses on the evaluation sheet, according to their responses in the audio recorded session in the first part of this phase. The evaluation sheet was used to gather the responses about the participants'

Major category	Task list items	Perception FST1	Perception FST2	Future attitude FST1	Future attitude FST2
Navigational items	Total score	60	60	60	60
1	Open FST (for public on campus)	45	45	43	37
2	URL findability	43	43	32	38
3	Time to download the home page	37	37	35	30
4	Navigation overall impression	40	40	38	39
Total score of category	$4 \times 60 = 240$	165	153	148	144
Grand total	$8 \times 60 = 480$				
Total FST1			313 out of 480		
Total FST2			297 out of 480		

Table I. Participant evaluation of navigational capabilities of FSTs

Major category	Item #	Task list items	Perception FST1	Perception FST2	Future attitude FST1	Future attitude FST2
Searching		Total score	60	60	60	60
	1	Easy to find simple search box	36	35	37	39
	2	Advanced search	35	36	32	34
	3	Level of simplicity	33	33	33	34
	4	Flexibility	28	30	37	36
	5	Accountability	30	29	33	38
	6	Available in eye catch zone	38	37	36	35
	7	Way to support query formulation	35	36	31	32
	8	# of items retrieved for test query	30	25	33	31
	9	Relevance of the first two items	35	31	38	33
	10	Searchable databases	40	44	39	36
	11	Results statistics	45	45	31	38
	12	Easy to manage	33	36	36	37
	13	Results description	44	40	42	36
	14	Response time	42	39	39	34
	15	Searching overall impression	40	35	38	32
Total score of category		$15 \times 60 = 900$	538	531	535	525
Grand total score		$30 \times 60 = 1{,}800$				
Total FST1				1,079 out of 1,800		
Total FST2				1,056 out of 1,800		

Note: The total value for searching is 1,800, representing 15 questions multiplied by 60 multiplied by 2 representing perceptions and attitudes

Table II. Evaluation of search capabilities of the FSTs

Major category	Item #	Task list items	Perception FST1	Perception FST2	Future attitude FST1	Future attitude FST2
Browsing		Total score	60	60	60	60
	1	Subject browsing	45	45	35	36
	2	A-Z list	46	44	37	32
	3	Database browsing	43	33	31	34
	4	Searchability	39	35	41	31
	5	Easy to use	44	45	36	40
	6	Browsing overall impression	44	42	35	42
Total score of category		$6 \times 60 = 360$	261	244	217	215
Grand total score		$12 \times 60 = 720$				
Total FST1				478 out of 720		
Total FST2				459 out of 720		

Table III.
Evaluation of browsing capabilities

Comments	Task list	Comments FST1	Comments FST2
Positive comments	It is fast in searching and link resolving	8	4
	More effective than database searching	8	3
	There is no way to directly show the full text	5	9
	Better than Google Scholar	5	2
	It allows searching within a subject category	8	3
	It would be my future starting search	14	12
	Easy to use	14	6
	It allows discovering all that I might need	5	2
	Very productive and efficient	6	2
	We should teach it to everybody	14	12
	Flexible	8	5
	Very important	6	3
	Easy to browse	9	4
	Retrieved results are valid	9	9
	Results description is poor	−2	−5
	Very complex	−2	−9
Negative comments	Huge number of results retrieved	−6	−6
	It is not easy to narrow or broaden search	−5	−6
	Expected more complex search parameters	−7	−5
	There is no way to modify search	−2	−6
Total		99	48

Table IV.
Participants' comments on the FSTs

Note: Negative comments receive a negative score, while positive comments receive a positive score

perceptions at the time of the experiment and their future attitudes which reflect their willingness to use each tool in the future. The sheet was used to facilitate gathering and collecting empirical data and to compare it with their recorded responses. Their audio comments showed their willingness to use the tool in the future, according to their perception satisfaction level. These comments were compared to their corresponding

response on the sheet. Four groups of usability features were evaluated. These are discussed and analysed below.

Navigation

The grand total score of the navigational capabilities is the sum of the eight elements presented in Table I. Four represent the participants' perceptions and four represent future attitudes. The total value for each item is 60, which represents 20 participants multiplied by three that refers to the highest value in the Likert scale, and means that the grand total score of the navigational capabilities equal 8 items multiplied by 60 = 480.

Table I shows that participants have reported better value for FST1 than that of FST2 in terms of the navigational capabilities. In general, FST1 has gathered a total value of 313 out of the 480, representing 65 per cent, while FST2 has gathered a total value of 297 out of the 480, representing 62 per cent of the total score.

Searching

The most frequent positive behaviour, reported by participants while searching the two FSTs, is the statistical display of the total number of items retrieved by each native database. The least frequent positive behaviour is the relevancy of the first two items, as it substantially demonstrated the capability of the evaluated FSTs to rank the retrieved results and to display the most relevant items on the top of the retrieved list. In general, there is no significant difference between the two FSTs in terms of searching capabilities, as the first tool gathered a score of 1,079 out of 1,800, representing 60 per cent, and the second tool collected a score of 1,056, representing 59 per cent, meaning that user perceptions and future attitudes towards both of them would not dramatically change. Table II shows the results of the searching capabilities from the users' perspectives.

Browsing

In terms of browsing capabilities, where information resources are grouped under subject categories and then divided into subcategories, FST1 collected over 66 per cent of the reported score and FST2 gathered 64 per cent of the total score, meaning that there is a slightly significant difference between the two FSTs. Table III shows detailed analysis for each FST and the total score of each category beside the grand total score for this item.

In general, participants reported that FST1 is slightly better than FST2, as it corresponds better to their perceptions and would prefer to use it in the future. Figure 1 shows the result of the total values for each tool in terms of perceptions and future attitudes.

Figure 1 shows that FST1 is slightly better than FST2 in terms of user perceptions and future attitudes. It is also clear that the total difference between the two FSTs, including the grand total of perceptions and future attitudes together, is 58, which indicates no significant difference between the two FSTs.

Other comments

The most important comments, collected and analysed from the audio recorded responses are described in Table IV. Results show that 70 per cent of the participants would prefer to start with FST1 in the future, while 60 per cent might use FST2 in the future, and some of the participants commented that they had no preference.

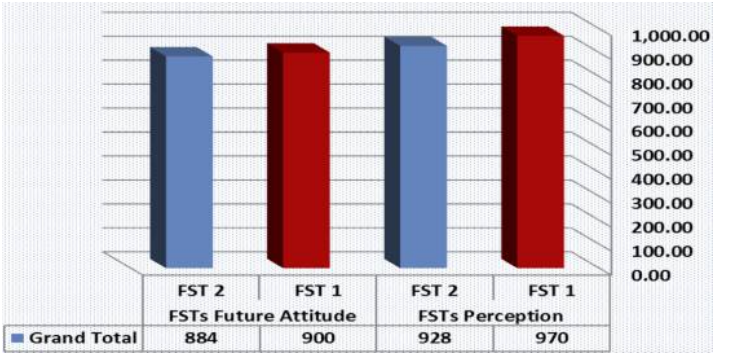


Figure 1.
Total score of perceptions and future attitudes

Additionally, 70 per cent found that both FSTs provide an easy way to find their resources and discover new resources.

Table IV shows that, in general, participants report more positive comments and feedback about FST1 than FST2.

Usability test: overall analysis

Figure 2 shows that the overall grand total score of the usability test for the two FSTs and summarizes the overall collected scores for each FST in Tables I-IV. It is clear that FST1 generates higher value in general than FST2. The grand total values reported in this figure are used to calculate the 38 per cent participant estimated weight of the usability test in the third phase of this study.

Table V shows the results of two independent sample *t*-test analysis for user perceptions and future attitudes which were carried out to statistically examine if there is any significant difference between the two FSTs. The two tests proved no significant difference exists between the two FSTs in terms of perception and future attitude, as the *p*-value is higher than 0.05 and equal to 0.876 for the perception and 0.782 for the future attitude.

The final analysis shows that a portal-based federated search solution is more usable and able to retrieve accurate and consistent results than the library management federated-based search solution, as it provides more accepted results in terms of retrievability, as well as more usable navigational, searching and browsing options. Although the analysis shows a preference towards FST1, users still perceive that FST2 is a good alternative, in case FST1 is not available or down as they reported in their comments in Table IV.

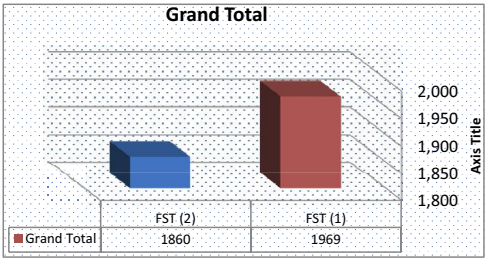


Figure 2.
Grand total value for each FST collected in the usability test

Table V.
Independent sample *t*-test for user perception and attitude

					Levene's test for equality of variances					
	N	Mean	SD	FST Perception	*F*	Significance	*t*	df	Significance (two-tailed)	Mean difference
FFST1	225	37.60	5.65685	Equal variances assumed	0.025	0.876	−0.769	48	0.446	−1.20000
FFST2	225	338.80	5.36967	Equal variances not assumed			−0.769	447.870	0.446	−1.20000
				FST attitude						
FFST1	225	335.92	33.41	Equal variances assumed	0.078	0.782	0.605	448	0.54	0.56
FFST2	225	335.36	33.12	Equal variances not assumed			0.605	447.61	0.54	0.56

FST retrieval performance

The retrieval performance is tested for 20 real user queries representing real information needs. FST performance is evaluated using the standard measurements in IR, including retrieval consistency and precision. Retrieval consistency of FSTs was calculated by comparing the retrieved results from each FST with the native databases performance for 20 real user queries. Each query was submitted to six different interfaces: two FSTs and four native databases. The search process was refined to submit the queries for the selected databases title search field regardless of whether the query is submitted indirectly through the FSTs or directly through the native database. The total number of the retrieved items was compared by calculating the deviation of each FST from the four native databases. Table AI shows the results of this step. It is clear from Table AI that FST1 is much more effective in terms of retrieval consistency, as it retrieved almost the same number of results as the native databases in most of the cases, while FST2 deviated from the native databases results in 18 queries. It is also clear that there are two queries which retrieved zero results. The Spearman correlation coefficient test is carried out to explore the relationship between each FST results for the 18 queries and the total number of items retrieved from the four native databases. The analysis shows that FST1 is more consistent in terms of retrieving results coinciding with the total number of results retrieved by the native databases with a correlation coefficient equal to 0.98 compared to 0.92 for FST2 (Table AI). This part shows that retrieval consistency could be easily used and tested to calculate the efficiency of the IR performance of FSTs. The correlation score is used in the final calculation step to report the overall evaluation.

Precision at 11-point cut-off recall (P11)

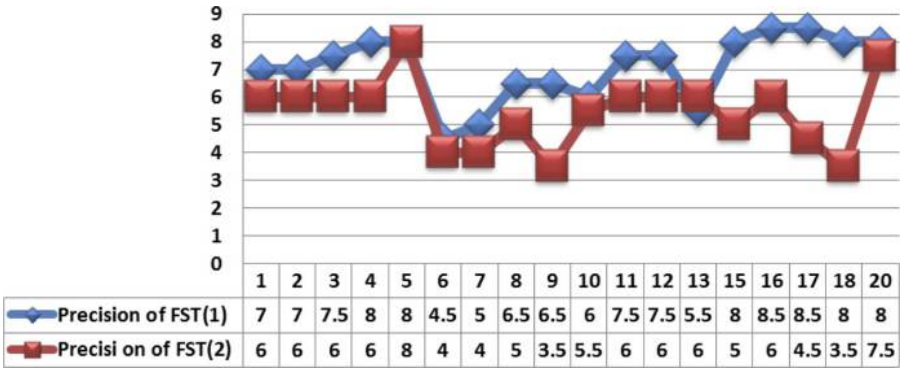
Precision at 11-point cut-off recall is computed using recall level at the standard 11 points. A common method is used to compute the 11-point average precision by considering the average precision over the standard recall points (0, 10, 20, 30 … 100 per cent). To calculate the precision for these standard recall points, precision and recall for each relevant document in the result set is calculated and interpolated. These standard levels allow for measuring the performance in the different areas of the retrieved results distribution. For example, if the system retrieved only four relevant documents out of ten at points 20, 30, 50, and 70, then at recall point 0.30, precision is $2/3 = 0.667$ because among the top three documents, only two documents are relevant. At recall point 0.60, precision is $3/6 = 0.50$ because among the first six documents, three documents are relevant. At recall point 0.90, precision is $4/9 = 0.444$ and so on:

$$\mathbf{P11} = \frac{\Sigma\ \mathbf{precision\ relevant, Q}}{\mathbf{N}}$$

Where $N = 20$ queries

Each participant evaluates the first ten items retrieved for her/his query in terms of a relevancy scale of 3 points (0 – irrelevant, 0.5 – partially relevant and 1 – relevant). Figure 3 shows the results of the precision analysis for each query.

Figure 3 indicates that FST1 retrieves more precise results than FST2 and that it is more effective in terms of precision values than FST2 in 16 cases out of 18. Both FSTs have similar precision scores for query number five, while FST2 was by chance more effective in query number 13.



	1	2	3	4	5	6	7	8	9	10	11	12	13	15	16	17	18	20
Precision of FST(1)	7	7	7.5	8	8	4.5	5	6.5	6.5	6	7.5	7.5	5.5	8	8.5	8.5	8	8
Precisi on of FST(2)	6	6	6	6	8	4	4	5	3.5	5.5	6	6	6	5	6	4.5	3.5	7.5

Figure 3.
Average precision for FST1 and FST2 for each query

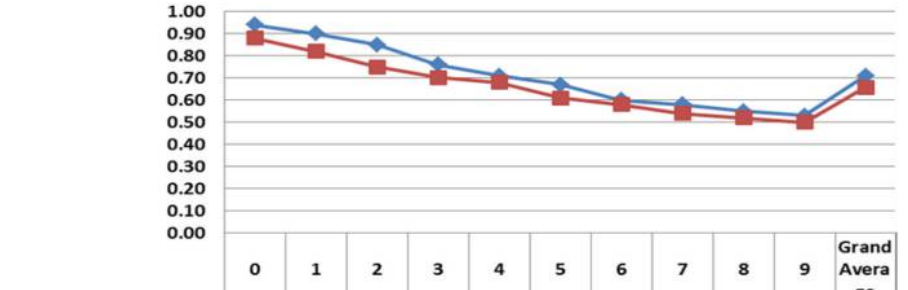
Figure 4 demonstrates the results of the 11-point cut-off recall. Results also show that FST1 is much more effective in terms of precision in all positions, and on average as well, which means that in all positions FST1 is much more consistent with user perception than FST2.

Overall evaluation

The overall values of the first two phases are interpolated according to the participants' initial weights at the preliminary phase. Table VI shows the interpolated weights for each phase.

The value of the overall evaluation is calculated for each FST by reporting the usability percentage then interpolating it to 0.38 and calculating the average retrievability score from the values of the retrieval consistency correlation and precision at 11-point cut-off recall average and then interpolating it to 0.62. It is clear that FST1 has a higher value than FST2 in terms of usability, retrievability and, hence, for the overall evaluation. It is also clear that usability and retrievability could be easily interpolated to compare IR systems in general with FST in particular.

The results of this phase are consistent with the general conclusion of the first two phases of the study. Thus, the overall conclusion is that the FST based on portal technology and XML gateways leverage the functionalities of usability and retrievability due to user



	0	1	2	3	4	5	6	7	8	9	Grand Average
Average for FST(1)	0.94	0.90	0.85	0.76	0.71	0.67	0.60	0.58	0.55	0.53	0.71
Average for FST(2)	0.88	0.82	0.75	0.70	0.68	0.61	0.58	0.54	0.52	0.50	0.66

Figure 4.
Precision at 11-point cut-off recall

acceptance for interface capabilities and retrieval consistency and accuracy of the results, as well as some other sophisticated features, such as data fusion, combination technique and deduplication, which need to be investigated in future research.

Implications and limitations

The paper develops a framework for comparing two FSTs using two different retrieval protocols: XML gateways and Z39.50. This study is significant for system developers and IR agents for effective design and evaluation of FSTs. Therefore, it has practical implications in two dimensions. First, the results could be utilized by FST developers to enhance their product's performance. Second, the framework could be used by librarians to evaluate FST performance and capabilities. The framework enables them to compare library systems in general with FSTs in particular. In addition to the practical implication, the authors encourage researchers to use and enhance the proposed framework.

On the other hand, this empirical study has several limitations. First, the lack of participants' familiarity with the usability testing creates the need for a deep awareness and rigorous supervision. Second, the difficulties of empirically assessing participants' perspectives and future attitudes called for mixing between a formal task-based and think-aloud protocol for participants in a real environment. This has been a challenge that affects the collection of usability data including user behaviour, expectation and other empirical data. Third, the differences between the two FSTs in terms of number of connectors and advanced search techniques require setting rigorous procedures for testing FST retrieval consistency and precision.

Conclusion

This paper proposes a framework for comparative evaluation of two FSTs based on usability testing and retrievability performance. The recommended framework for evaluation contains three phases: usability testing, retrievability performance assessment and overall comparison. Usability testing is usually conducted through task-based testing with users or expert testing. This study utilizes think-aloud tests. The retrievability performance evaluation is based on solid IR measurements, including retrieval consistency and precision. The overall evaluation and comparison combined the two approaches to assess the effectiveness and efficiency of FST from a user and system perspective. The proposed model designed in this paper could be exploited in FST evaluation, as it includes the most important indicators for evaluation as indicated by the participants reflecting their perspectives and future attitudes. Participants

Table VI.
Overall evaluation results

Evaluation approach	Score	FST1	FST2	Total
Usability	Grand usability score	1969	1860	3829
	Interpolated usability %	52%	48%	100%
	Interpolation usability to 38%	20%	18%	
Retrievability	Retrieval consistency correlation value	98	92	
	Precision at 11-point cut-off recall precision average	71	66	
	Average retrievability	84.5	79	
	Interpolation retrievability to 62%	52.39%	48.98%	
Grand total usability and retrievability		72.39	66.98	

assigned a value of 62 per cent as a weight for the system retrievability, as their preferred indicator, and a value of 38 per cent for usability testing to reflect their level of importance in the evaluation framework. The final results indicate that the FST that is based on portal technology, including search, link resolver and XML gateways, is much more effective than the FST based on library system technology that used the Z39.50 protocol for database connectivity. Companies developing FSTs and library management systems could utilize the final results of this study to enhance their final products, and librarians could use the proposed framework of evaluation to compare among different FST products.

References

Abercrombie, S.E. (2008), "Evaluation of federated searching options for the school library", *School Library Media Research*, Vol. 11, available at: www.ala.org/aasl/sites/ala.org.aasl/files/content/aaslpubsandjournals/slr/vol11/SLMR_EvaluationFederated_V11.pdf (accessed 15 January 2013).

Arguello, J., Capra, R. and Wu, W.C. (2013), "Factors affecting aggregated search coherence and search behaviour", *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 1989-1998.

Avrahami, T.T., Yau, L., Si, L. and Callan, J. (2005), "The FedLemur project: federated search in the real world", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 3, pp. 347-358.

Bouchon-Meunier, B. (1998), "Aggregation and fusion of imperfect information", *Studies in Fuzziness and Soft Computing*, Physica-Verlag Heidelberg, New York, NY.

Boyd, J., Hampton, M., Morrison, P., Pugh, P. and Cervone, F. (2006), "The one-box challenge: providing a federated search that benefits the research process", *Serials Review*, Vol. 32 No. 4, pp. 247-254.

Bron, M., Van Gorp, J., Nack, F., Baltussen, L.B. and de Rijke, M. (2013), "Aggregated search interface preferences in multi-session search tasks", *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 123-132.

Buck, S. and Nichols, J. (2012), "Beyond the search box", *Reference & User Services Quarterly*, Vol. 51 No. 3, pp. 235-245.

Burke, L. (2001), "The future role of librarians in the virtual library environment", *Australian Library Journal*, Vol. 51 No. 1, pp. 31-45.

Cervone, F. (2005), "What we've learned from doing usability testing on OpenURL resolvers and federated search engines", *Computers in Libraries*, Vol. 25 No. 9, pp. 10-14.

Chen, X. (2006), "MetaLib, WebFeat, and Google – the strengths and weaknesses of federated search engines compared with Google", *Online Information Review*, Vol. 30 No. 4, pp. 413-427.

Dwork, C., Kumar, R., Naor, M. and Sivakumar, D. (2001), "Rank aggregation methods for the web", *Proceedings of the 10th International Conference on the World Wide Web*, *ACM*, pp. 613-622.

Fryer, D. (2004), "Federated search engines: federated searching aggregates multiple channels of information into a single searchable point", *Online*, Vol. 28 No. 2, p. 16.

Georgas, H. (2013), "Google vs. the library: student preferences and perceptions when doing research using Google and a federated search tool", *Portal-Libraries and the Academy*, Vol. 13 No. 2, pp. 165-185.

George, C.A. (2008), "Lessons learned: usability testing a federated search product", *The Electronic Library*, Vol. 26 No. 1, pp. 5-20.

Jacso, P. (2003), "Savvy searching", *Online Information Review*, Vol. 28 No. 6, pp. 454-460.

Jaffe, H. and Mukherjee, R. (2013), "System and method for dynamic context-sensitive federated search of multiple information repositories", *US Patent Application* 10/743, 196, available at: www.google.com/patents/US20050149496 (accessed 12 June 2013).

Kopliku, A., Pinel-Sauvagnat, K. and Boughanem, M. (2014), "Aggregated search: a new information retrieval paradigm", *ACM Computing Surveys (CSUR)*, Vol. 46 No. 3, p. 41.

Kumar, S., Sanaman, G. and Rai, N. (2008), "Federated search: new option for libraries in the digital era", *International CALIBER*, pp. 267-285.

Lampert, L. and Dabbour, K. (2007), "Librarian perspectives on teaching metasearch and federated search technologies", *Internet Reference Services Quarterly*, Vol. 12 Nos 3/4, pp. 253-278.

Liu, T., Wang, F. and Agrawal, G. (2012), "Stratified sampling for data mining on the deep web", *Frontiers of Computer Science*, Vol. 6 No. 2, pp. 179-196.

Lu, J. and Callan, J. (2005), "Federated search of text-based digital libraries in hierarchical peer-to-peer networks", *Advances in Information Retrieval*, Springer, Berlin Heidelberg, pp. 52-66.

Mohamed, K. (2006), *Merging Multiple Search Results Approach for Meta-Search Engines*, *PhD thesis*, University of Pittsburgh, Pittsburgh, PA.

Mohamed, K. and Hassan, A. (2008), "Web usage mining analysis of federated search tools for Egyptian scholars", *Program: Electronic Library and Information Systems*, Vol. 42 No. 4, pp. 418-435.

Nielsen, J. (1992), "The usability engineering life cycle", *Computer*, Vol. 25 No. 3, pp. 12-22.

Randall, S. (2006), "Federated searching and usability testing: building the perfect beast", *Serials Review*, Vol. 32 No. 3, pp. 181-182.

Si, L. and Callan, J. (2005), "Modeling search engine effectiveness for federated search", *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *ACM*, pp. 83-90.

Tchangalova, N. and Stilwell, F. (2012), "Search engines and beyond: a toolkit for finding free online resources for science, technology, and engineering", *Science and Technology Librarianship*, available at: www.istl.org/12-spring/internet1.html (accessed 7 July 2013).

Tennant, R. (2001), *Digital Libraries: Cross-Database Search: One-Stop Shopping*, available at: http://libraryjournal.reviewsnews.com/index.asp?layout=articlePrint&articleID=CA170458 (accessed 7 July 2013).

Tsikrika, T. and Lalmas, M. (2001), "Merging techniques for performing data fusion on the web", *Proceedings of the Tenth International Conference on Information and Knowledge Management*, *ACM*, pp. 127-134.

Voorhees, M. (2000), "Variations in relevance judgments and the measurement of retrieval effectiveness", *Information Processing & Management*, Vol. 36 No. 5, pp. 697-716.

Wrubel, L. and Schmidt, K. (2007), "Usability testing of a metasearch interface: a case study", *College & Research Libraries*, Vol. 68 No. 4, pp. 292-311.

Yuwono, B. and Lee, D.L. (1996), "WISE: a World Wide Web resource database system", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8 No. 4, pp. 548-554.

Appendix

#	Queries	FST1	FST2	Science Direct	Emerald	Sage	ProQuest	Total of the four databases
1	Digital libraries management	22	52	1	11	0	0	12
2	Semantic web technologies	90	141	25	4	2	7	38
3	Library management system	49	208	4	16	4	25	49
4	Information retrieval system	290	303	131	8	22	120	281
5	System analysis and design	610	614	93	5	0	610	708
6	Libraries web sites design	27	17	0	0	0	0	0
7	Academic libraries automation	10	17	1	2	0	6	9
8	Public libraries organization	22	27	0	0	0	3	3
9	Cost effectiveness of information	15	67	0	1	0	4	5
10	Electronic government	114	168	37	16	9	51	113
11	Information-seeking behaviour	193	209	74	15	20	119	228
12	Federated search	18	37	6	9	1	3	19
13	Digital information management	13	19	0	0	0	0	0
14	Machine readable catalogue	0	0	0	0	0	0	0
15	Dublin core	17	60	1	15	2	0	18
16	Database management system	110	116	108	4	18	119	249
17	Semantic search	42	49	10	5	4	27	46
18	Cross search	31	114	0	1	0	30	31
19	Webometrics analysis	0	0	0	0	0	0	0
20	Human computer interaction	360	491	210	17	59	143	429
Total		033	809					2,238
Recall		0.8	25.5					
Correlation		0.98	0.92					

Table AI. Appendix 1. Total number of items retrieved for each query from the different searching tools

Corresponding author
Khaled A. Mohamed can be contacted at: kam00@fayoum.edu.eg