Evaluating Federated Search Tools: A Comparative Study

Khaled A. Mohamed
Library and Information Science Dep.
El-Fayoum University
Email: kam00@fayoum.edu.eg

Ahmed Hassan
School of Engineering
Ain Shams University
Email: ahassan@ictp.edu.eg

Abstract:

Federated search tools (FSTs) are Meta information retrieval systems, developed mainly to facilitate multiple resources searching through a single search box. They allow searching heterogeneous platforms, which include bibliographic and full text databases, OPACs, web search engines, open access resources...etc, using Meta searching mechanisms. Meta and federated searching tools would facilitate discovering and reaching the deep web through openURLs and search resolver. This study would focus on exploring a framework for evaluating and comparing different federated search tools. The proposed framework consists of three phases: the usability testing, retrievability performance assessment, and overall comparison. Usability testing is usually conducted through users' task or expert testing. Think a loud protocol has been examined for usability testing and FSTs recall and precision have been used to assess the retrievability performance for 20 real user queries. Participants have been directed to give weights for the interface usability and system retrievability importance as indicators for FSTs evaluation. They decided that FSTs retrievability is much more important than interface usability as they can discover the hidden features by training and cannot improve system retrievability without system improvements. They give an average weight of 62 % for the system retrievability and 38 % for interface usability. The think a loud test found no significant difference between the two FSTs while information retrieval (IR) performance measurements found minor differences in terms of recall and precision. The overall evaluations shows that FST which has been developed based on portal technology is better than FST which has been based on library system technology.

الملخص العربى:

أدوات البحث الفيدرالى هى أحد أنظمة استرجاع المعلومات التى تنتمى لفئة ما وراء النظم. وقد تم تطوير هذه الاداوت لكى تقوم بعمليات البحث فى أكثر مصدر معلومات واحد فى نفس الوقت من خلال صندوق بحث موحد. وتتيح هذه الادوات التعامل مع منصات البحث غير المتناغمة والتى تشمل قواعد البيانات الببليوجرافية، وقواعد بيانات النصوص الكاملة، والفهارس المتاحة على الخط المباشر وفهارس الويب، ومحركات البحث، ومصادر الويب المفتوحة ...الخ من خلال استخدام أليات ما وراء البحث. وتساعد أدوات ما وراء البحث وأدوات البحث الفيدرالى على اكتشاف المصادر والوصول إلى المواد المتاحة فى الويب العميق من خلال استخدام معرف المصادر الموحد المفتوح OpenURL ومحلل البحث search resolver. وتسعى هذه الدراسة إلى وضع إطار لتقييم والمقارنة بين أداوت البحث الفيدرالى. ويشتمل هذا الاطار على ثلاث مراحل: مرحلة اختبارات القابلية للاستخدام Usability، أختبارات قياس أداء الاسترجاعية Retrievability، التقييم العام Overall Evaluation. وقد تم الاعتماد على بروتوكول التفكير بصوت عال A Loud Thinking Protocol فى إجراء اختبارات قابلية الاستخدام، واستخدام قياسات الاستدعاء والتحقيق لقياس كفاءة أداء الاسترجاع وذلك لـ ٢٠ مستفيد واستفسار حقيقى. وقد تم توجيه المشاركون فى الدراسة إلى إعطاء وزن نسبى لكفاءة واجهة التعامل وكفاءة الاسترجاع كمؤشرات أساسية لتقييم كفاءة أدوات البحث الفيدرالى. وقد قرر المشاركون أن القدرة على الاسترجاع (مقياس الاسترجاعية) هو أكثر أهمية من قابلية الاستخدام حيث حددوا نسبة ٦٢ % كوزن نسبى لقيمة الاسترجاعية فى مقابل ٣٨ % وزن نسبى لقيمة القابلية للاستخدام. وأثبت بروتوكول التفكير بصوت عال لتقييم القابلية للاستخدام أنه لا توجد فروق جوهرية بين النظامين الذين تم تقييمهما، فى حين أثبتت مقاييس أداء استرجاع المعلومات أنه توجد فروق طفيفة فى كل من الاستدعاء والتحقيق. وقد أوضح التقييم العام قابلية الاطار المقترح للاستخدام والتطبيق الفعلى، كما أثبت أن أدوات البحث الفيدرالى التى تم تطويرها بالاعتماد على امكانيات بوابات البحث والتى تستخدم حلول البحث والربط Link and Search Solver أكثر فعالية وكفاءة من أدوات البحث الفيدرالى التى تعتمد على تكنولوجيا نظم المكتبات والبحث باستخدام بروتوكول Z39.50.

Introduction:

In the era of digital libraries, bibliographic and full text databases have become important resources for education and research, providing functionality and ease of use 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 allows huge amount of information resources to be integrated and become available for searching from one hyper space. Some of these resources available for public user and could be reached without the need for authentications as they are available in the surface web, while bibliographic databases need user authentication to enable digging in the databases, because most of the time they hide behind scripts which are known as deep web (Liu, Tantan; Wang, Fan; Agrawal, Gagan, APR 2012).

Each one of these databases has its own interface and searching capabilities. Searching these different databases requires deep understanding, from the users' side, of coverage of each database, perceiving the area of interest, and how users could fully employ the database capabilities, beside the difficulties of moving from one interface to another. These make the searching process tedious and time consuming rather than time saving. All of these requirements pushed library users to always complain about complexity of databases searching comparing it with web searching, such as World Wide Web search engines (ex. Google). The simplicity of such tools motivates library users to ask for a single search box that able to explore everything (Boyd, Hampton, Morrison, & Cervone, 2006; Burke, 2001). The fact that search engines has successed in searching the surface web, but couldn't easily handle the deep web resources, motivates researchers and IR developers to look for new solutions for searching and discovering the deep web (Mohamed, 2004).

Deep and dynamic web resources need special efforts for providing access to hidden resources using special techniques different than those used for searching the surface web. These different methods of information access lead to different tools such as web directories, search engines, and Meta searching, which provide access to the surface web, without any authentication, while discovering the deep web requires extensive authentications and customizations efforts from the system

developers and users as well. The large volume of documents housed in abstracting and full text databases, which known as deep web, is not searchable through these tradtional Internet searching tools because of the limitations in crawler technology and authentication requirements. One of the practical and applicable solutions for this problem has been discovered in developing a Meta searching tool known as federated search tools -FSTs ((Jacso, 2004, Avrahami, T. T., Yau, L., Si, L., & Callan, J, 2006). Although there are many terms used in the specialized literature referring to federated searching such as: federated search tools (FSTs), cross-database searching, meta-searching, or library portal; the first term is the most known one among researches and practitioners as well (Tchangalova & Stilwell, 2012).

Using FSTs as the main platform of resource discovering require the following: (1) generating a query from a unified single interface and broadcasting it to a group of disparate databases (predefined set of databases and/or search engines (SE)) with the appropriate syntax, which require some form of query translation, (2) combining the results collected from different sources and / or databases using data fusion and combination technique (Dwork, Kumar, Naor, & Sivakumar, 2001, Kumar, Sanaman, Rai, 2008) (3) presenting them in a succinct and unified format with minimal or no duplication (often called de-duplicating or de-duping), and (4) providing a means, performed either automatically or by the portal, to sort the merged result set (Tennant, 2001, Boyd, Hampton, Morrison, & Cervone, 2006, Zhang & Dong, 2002). All these complicated 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, this new mode of searching have, on users' search behaviors (George. 2008).

Fusion and aggregation of information are major problems for all kinds of knowledge base and information retrieval systems, from text

and multimedia processing to decision making. (Yuwono & Lee, 1996). Nevertheless, there are two general approaches to this scheme, depending on the problem to be dealt with (Smeaton & Crimmins, 1997). The first one corresponds to the aggregation of preferences given by several individuals of a group or the aggregation of criteria to satisfy specific needs in order 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. Several methodologies and/or theories are useful to manage such imperfect information. Among the most important ones are probability theory, evidence theory, fuzzy set theory and possibility theory (Bernadette, 1998). Many of these theories have been used to develop IR systems to allow developers to explore the different IR models including the Boolean, vector space, probabilistic, and language models. There are many tools available for federated searching utilized these techniques to facilitate cross searching domain around the world. Some of them are turnkey or ready made systems such as MuseGlobal, Summon, Endeavor, Ex Libris- MetaLib, WebFeat, DQM2, Deep Query Manager, and Fretwell-Downing, others are open sources such as dbWiz search, OpenSiteSearch, etc. Some of the turnkey systems are mainly developed to be standalone system and mainly developed depending on portal technologies such as MuseGlobal and web feat which based on web portals technologies, while others are developed as modules and parts of library management systems using OPAC searching mechanisms. The availablity of these different tools motivated researchers to evaluate these new tools and their baseline technology to support librarians in the selection of the appropriate tools.

This study corresponds to this motivation in order to help librarians to evaluate and compare among these different FSTs. The framework of the comparison gets along with the development scenario of FSTs,

which faces two serious challenges. The first one is making 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 one is retrieving the appropriate resources, which mainly defined by the source databases and tested by the well-known IR performance measurements. This study would try to develop a framework for evaluation then use it to evaluate FSTs in terms of usability and IR performance in order to investigate a solid evaluation benchmark.

I. Related Literature:

This part would present the related literature to the current study. The review would cover FSTs studies and the evaluation techniques used to discover these tools capabilities.

Federated search has been evaluated and tested using different techniques including user evaluation and information retrieval performance assessment. Usability testing is one of the well-established and accepted techniques for evaluating the usability of different library system tools, such as OPAC, online databases and library portal. As a user centered approach for evaluating user's perspective, attitudes, level of satisfaction, usability have been employed to investigate user interfaces intuitiveness and capabilities. It provides the investigators with rich quantitative and qualitative data to support their research practices. Usability test means gathering information about user behaviour while interacting with information systems. It usually uses five commons indicators to evaluate any system from users perspective such as: easy to learn, efficient to use, easy to remember, few errors, satisfaction (Mohamed & Hassan, 2008, Nielsen, 1993; Soken et. El ,1993; Shneiderman, 1998):

George (2008) conducted a usability testing using think a loud protocols for evaluating MetaLib federated search system which has been developed by Ex Libris Ltd to be used as a library portal. A demographic

questionnaire has been distributed to selected sample of eight volunteer, diverse with respect to affiliation discipline, gender, language and computer experience. Participants have been guided by mediators to use the Metalib interface to complete real world tasks while verbalizing their thoughts about the FST. Their study shows that participants face difficulties in the following processes: system login, primary and secondary navigations, confusing terminology, and inconsistency with site design and use expectation.

Another important study has evaluated the usability of one of the federated search interface used by the University of Maryland and its affiliated institutions. It has investigated student's perceptions of the search usefulness and to what extent student could effectively complete search tasks using federated search. Student perceive federated search to be useful tool but they had low rates of success in completing some tasks (Wrubel, L, Schmidt, K, 2007).

Cervone (2005) has carried out a usability test for evaluating the library portal of Northern University in Evanston, Illinois. This study has focused on what have been learned from usability testing of open URL resolvers and federated search tools. The test has investigated Northern University in Evanston, Illinois portal users' perception and attitudes to make sure that federated search tool work from user perspective.

Randal (2006) investigated the usability of endeavors information system by gathering insights of the critical requirements of libraries and information professionals about the federated search engines capabilities' which is part of their integrated system. The results of the study, conducted in conjunction with market research and consumer focus group, to investigate the relevancy of endeavors federated search technology to their needs and behaviors. The usability testing has been implemented in two universities libraries'. Twenty participants have guided to use the FST. Endeavors usability testing has proven the value

and power of federated searching and there are always possibilities for libraries to make user experience more satisfying and fruitful.

Xiaotian (2006) describes the features and capabilities of library federated search engines. He compared between MetaLib and WebFeat as research tools by highlighting their strengths and weaknesses against Google and Google Scholar. He reported that MetaLib and WebFeat have fundamental differences and they cannot compete with Google in speed, simplicity, ease of use, and convenience, nor can they be truly one-stop shopping. Their strengths lie in the contents they search as well as in the objective way they retrieve and display results.

There are few numbers of studies that have focused on evaluating federated search engines as information retrieval system. Avrhami et al (2005) has carried out a series of experiences to develop a prototype federated search system for US government's fed state web portal, and the issues addressed in adopting research solutions to this operational environment. A series of experiments identified how well previous research results, parameters settings and heuristics applied in the fed states environment. The study concluded with a set of learned lessons from this technology transfer effort including observation about search engines quality in the real world.

Lampert and Dabbour (2007) have carried out three assessment projects: two of them focused on libraries' reaction to Meta search technologies from a reference and information literacy perspective and the third one is a user survey that attempts to capture students experiences, understudying, and satisfaction with meta search at California state university Northridge. They have presented an investigation of users' understanding of MetaLib Combined Search (MCS), as federated search system implemented in the Washington Research Library Consortium. Through a survey instrument, libraries, and students reported their experience, usage, and opinions of the

system. Upon responding to process-related questions about a search simulation included in the questionnaire, participants described and illustrated their understanding of MCS operation. Data shows that student 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 operation, students focused largely on full text retrieval capability and search efficiency whereas librarians paid more attention to search strategies and retrieval quality. Both students and libraries indicated that lack of background information about MCS operation was problematic.

In a recent study Buck & Nichols (2012) have explored a participatory design strategy to investigate users views on what a discovery system should look like and function. They have asked groups of participants to draw their idea of what a discovery system should look like. The findings reveals what librarians think are important features for these tools. The authors also discuss the use of the participatory design process.

In another recent study Georgas (2013) has investigated undergraduate student preferences and perceptions when using both Google and a federated search tool. 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 believe that: they possess strong searching skills, are able to find relevant sources using both search tools, but actually prefer federated search tool to Google for doing research. Thus, despite federated searching's limitations, students see the need for it, libraries should continue 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.

This study would present a framework for evaluating and comparing between two FSTs. The evaluation technique would depend on using usability testing and IR retrieval performance assessment. The study would combine the lessons learned from related literature and other evaluation study implemented by the researchers in real the world environment systems similar to this study.

II. Research Design and Methodology:

Selecting the most appropriate FST tool to access the multiple platform search from a single interface requires succinct evaluation for the different features of the available tools based on solid evidences. Thus, this study would try to evaluate and compare between two widely used FSTs in Egypt in terms of systems usability and retrievability performances. The first system is based on portal technology and it consists of two main components which are integrated together in one library portal known as OVID Portal1. The two components are search resolver developed by MuseGlobal and Link Resolver developed by OVID Technology, the second system is based on library system technology as it is integrated with a library management system developed in El- Mansoura and Zagazig Universities in Egypt. This system is deployed in the Egyptian universities libraries as the backbone of the Egyptian universities libraries union catalog. The two FSTs are heavily used in the Egyptian universities because some universities are using the first one as the primary FST and others are using the second one as the single access point for all Egyptian universities resources including OPACs and digital theses2. Therefore, it is important to compare and evaluate the features, and capabilities of the two FSTs. The major purpose of this study is to identify the main differences between those two tools and

1 www.eul.edu.eg

2 www.eulc.edu.eg

proposing a framework including the procedures of evaluation, which could facilitate such comparisons in the future. This part would focus on identifying research questions, methods of testing, and data collection procedures.

Research Questions:

Two simple and concise questions have been examined in order to evaluate the two FSTs. These questions are:

1- What are the major differences between the two FSTs in terms of usability indicators?

2- What are the major differences between the two FSTs in terms of retrievability performance measurements?

In order to answer these two questions statistically a hypothetical framework for the evaluation procedure has designed and tested in a real environment. The researchers assume that FST which is based on portal technology would perform higher than the FST based on library management system in terms of usability and retrievability.

Research Design:

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

The participants have investigated the FSTs usability and retrievability under the supervision of the investigators during a post graduate course which is a special topic in libraries and information science. Participants were recruited for three hours lap session, carried out in two different days, the first one used for awareness and usability and the second one discussed the retrievability testing procedures and evaluation technique.

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

1. Assign a value out of 100 % for the importance of the interface usability and System Retrievability as indicators for evaluating FSTs?
2. From your point of view, are there any other important indicators for evaluating FSTs?
3. Have you ever used FST and are you aware of their techniques of searching?

Eighteen Participants decided that FSTs retrievability is much more important than interface usability as they can discover the hidden features by training but cannot improve the system retrievability, while two participants decided that they are equal. Participants indicated an average weight of 62 % for the system retrievability and 38 % for interface usability. Participants have indicated some features of the usability testing including searching and browsing capabilities as other indicators for the evaluation, so they did not add any new valid or important indicator for evaluation.

Fifteen participants have stated that they have used FSTs tool at least once to find information for research purposes. Participants stated that they do not have enough information about FSTs techniques which they are using to fetch and search for information resources.

participants then attended the awareness session which includes detailed demonstration about the major concepts of Meta searching and how user can fully and deeply utilize the examined tools.

After this preliminary phase, the design of the evaluation procedures has been taken place and divided into three major phases, including usability testing, system retrievability and overall evaluation.

1- Usability Testing:

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

The usability phase divided into two sessions, one for awareness and learning and the other is for executing the required usability task, then evaluating the two different tools. In this phase, participants have been given an awareness session for 45 minutes about how to use the two FSTs and asked to report a topic of interest to be used for searching in the area of library and information sciences. Participants are then directed to try to use the two tools and speak loud about what they are thinking, perceiving, expecting and recognizing.

Think a loud protocol has been exploited to gather participants' evaluation for the two FSTs to investigate the interface capabilities including navigational, searching, and browsing capabilities. This activity provides a mental model for users through asking them to speak out loud during task completion and verbalizing what they are doing and think out. Nielsen (1993) has reported the importance of think loud protocol by saying "thinking loud may be the single most valuable usability engineering method".

This study has mixed between setting a formal task and thinking a loud protocol for participants on a real environment. This

allows collecting usability data about user behavior, expectation and other empirical data. A predefined short data collection form has been designed to collect the empirical data during the think loud protocol as it is clear in the results and discussion section. Approaches to usability testing may vary, but the "think loud" protocol mixed with structured open ended questionnaire, where participants use FSTs and describe their experience out loud, responding to the required task and reporting their evaluation in a structures format. This technique is offering a balance between efficiency and quality of data collection. In the think a loud session participants were instructed to use the FSTs interfaces to test what they have learned about the major components of FST in the awareness session, what they perceive and expect? And finally, what are their future attitudes towards the two tools?

The usability test includes four groups of questions about the interface capabilities (see tables 1-4). Participants have been directed to assign a value out of three for each question and speak loudly when they decide which value she/he would give to the evaluated parameter during the task. The total value for each question is 60 considering a score of 3 multiplied by 20 participants. Then the total usability value for each FST has been calculated and normalized out of the 38 % preliminary indicator in the overall evaluation.

2- Retrieval performance Assessment:

The performance of the FSTs has been tested as information retrieval system in terms of recall and precision. A set of 20 queries has been examined to measure the performance of the two tools. Recall of the two FSTs has been compared using the total number of the retrieved items form the same set of databases, and accordingly precision has been measured in terms of precision at 11 point cut of recall values. This part of the study has been executed after the usability test in a separate session as participants have become fully aware about the features of the two FSTs. Participants have been directed to select

queries and structured them in the format of title search according to the following conditions:

1-They should represent real user needs for evaluation purposes.

2-Quereies should be simple (two or three terms in maximum representing topics) to retrieve results to be analyzed and evaluated.

In this phase, each participant has been directed to search for the required query using the advanced search interfaces and submit the query to the title field. Participants have assessed the retrievability performance of the FSTs using four specific and unified databases (see table 1), and reported the results of their query in an excel sheet including the number of items appeared and the relevancy of the first 10 item retrieved from each tool. In order to calculate the recall, participants have 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 would probably retrieve the same list of items with different ranking. Recall and precision of the retrieved results have been calculated according to 3 points relevancy scale (0 – Irrelevant, 0.5 Partially Relevant, 1.0 Relevant). Recall of the two federated search tools has been compared with the original and native results retrieved from the source databases. The total value of the retrievability indicators has been calculated as the recall and precision have been given equal score of 50 % for each, out of the assigned 62 % total value which has been allocated for retrievability importance in the preliminary phase.

3- Overall Evaluation:

The overall value of the evaluation has been calculated for each FST by summing up the value of the usability to the value of the retrievability in order to allocate a total value for each FST. These

baseline values for each indicator consider the value of the output of the two phases in terms of 62 % for the retrievability and 38 % for usability.

VI. Results and Discussions

The following parts would show and discuss the results of the three phases of this study including: the usability testing, retrievability assessment and overall results.

Phase I: Usability Testing

The think a loud protocol has been investigated in this phase to explore participants' perceptions and future attitudes toward the two FST's. The awareness session has been audio recorded then participants have divided into two groups, 10 participants for each. They have been directed to structurally use the two FSTs, 15 minutes for each which have been audio recorded, to execute a predefined set of tasks in the investigation session. Participants have been instructed to speak up loudly to record their responses indicating their impressions, perceptions, and expected future attitudes for three major categories of the FSTs interfaces. The audio records involved capturing the verbal feedback and their evaluation for each part of the procedures. The required task contains three major categories, as appeared in tables 1-4, including Navigation, Searching, Browsing, and other comments. A Predefined loud protocol evaluation sheet has been structured including all the expected responses. The usability evaluation sheet includes two parts, one for perception and the other one is for future attitudes. Each part (perceptions and attitudes) has been given a score out of 60 for each element. Participants have directed to record their responses using a likert scale of three to indicate their Perceptions and attitude towards each element of the FSTs. The likert scale includes the following responses:

1- Negative 2- Partially Positive 3- Positive

Then participants have been directed to report their usability responses in the evaluation sheet according to their response in the first part of this phase. There are four groups of usability features have been evaluated, and the following section would discuss and analyze the participants' responses for these groups.

1– Navigation:

Table (1) shows that participants have reported a higher value for FST (1) than FST (2) in terms of navigational capabilities. In general FST (1) has collected a total value of 313 which represents 65 % of the total score3 of the navigation features and capabilities, while FST(2) has gathered a total value of 297 representing 62 % of the total score.

Table (1) Participants evaluation to FSTs Navigational capabilities

Major Category	Task List	FSTs Perception		FSTs Future Attitude	
		FST 1	FST 2	FST 1	FST 2
Navigation	Total Score	60	60	60	60
1	Open FST	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		165	153	148	144
Total FST(1)		313 out of 480			
Total FST (2)				297 out of 480	

2– Searching:

[3] Total Score represents 4 Indicators for perceptions and 4 for attitudes each one is out of 60 which means 8 * 60 = 480

The most frequent positive behavior participants have reported while searching the two FST is the statistical analysis of the results, which shows the number of items retrieved from each database. The least frequent positive behavior was the relevancy of the first two items, as they are presenting the capabilities of the FSTs to rank the retrieved results and display the most relevant items at 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 value of 1079 out of 1800 representing 60 % and the second tool collected a value of 1056 representing 59% which means that user perceptions and future attitudes toward both of them would not dramatically change. Table (2) shows the results of searching capabilities from user perspectives.

Table (2) Participants evaluation to FSTs Searching capabilities

Major Category	Task List	FSTs Perception		FSTs Future Attitude	
		FST 1	FST 2	FST 1	FST 2
Searching[4]	Total Score	60	60	60	60
	Easy to find simple search Box	36	35	37	39
	Advanced Search	35	36	32	34
	Level of Simplicity	33	33	33	34
	Flexibility	28	30	37	36
	Accountability	30	29	33	38
	Available in eye catch zone	38	37	36	35
	Way to support query formulation	35	36	31	32
	No. of Items retrieved for my test query	30	25	33	31

[4] Searching total value is 1800 representing 15 question multiplied by 60 multiplied by 2 representing perceptions and attitudes

Major Category	Task List	FSTs Perception		FSTs Future Attitude	
		FST 1	FST 2	FST 1	FST 2
	Relevance of the first two items	35	31	38	33
	Searchable Databases	40	44	39	36
	Results Statistics	45	45	31	38
	Easy to manage	33	36	36	37
	Results Description	44	40	42	36
	Response time	42	39	39	34
	Searching Overall Impression	40	35	38	32
Total Score of Category		538	531	535	525
Total FST(1)		1079 out of 1800			
Total FST(2)		1056 out of 1800			

3- Browsing:

In terms of browsing, where information resources are grouped by subject categories and then divided into subcategories. FST(1) has collected over 66 % of the reported score and FST(2) gathered a 64 % of the total score which means that there is a slightly significant difference between the two tools in terms of browsing capabilities perception and future attitude. The following table shows detailed analysis of both FST according to the participants' responses.

Table (3) Participants evaluation to FSTs Browsing capabilities

Major Category	Task List	FSTs Perception		FSTs Future Attitude	
		FST 1	FST 2	FST 1	FST 2
Browsing	Total Score	60	60	60	60
	Subject Browsing	45	45	35	36
	A-Z list	46	44	37	32

Major Category	Task List	FSTs Perception		FSTs Future Attitude	
		FST 1	FST 2	FST 1	FST 2
	Databases Browsing	43	33	31	34
	Searchability	39	35	41	31
	Easy to use	44	45	36	40
	Browsing Overall Impression	44	42	35	42
Total Score of Category		261	244	217	215
Total FST(1)		478			
Total FST(2)					459

In general, participants have reported that FST (1) is slightly better than FST (2) 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.

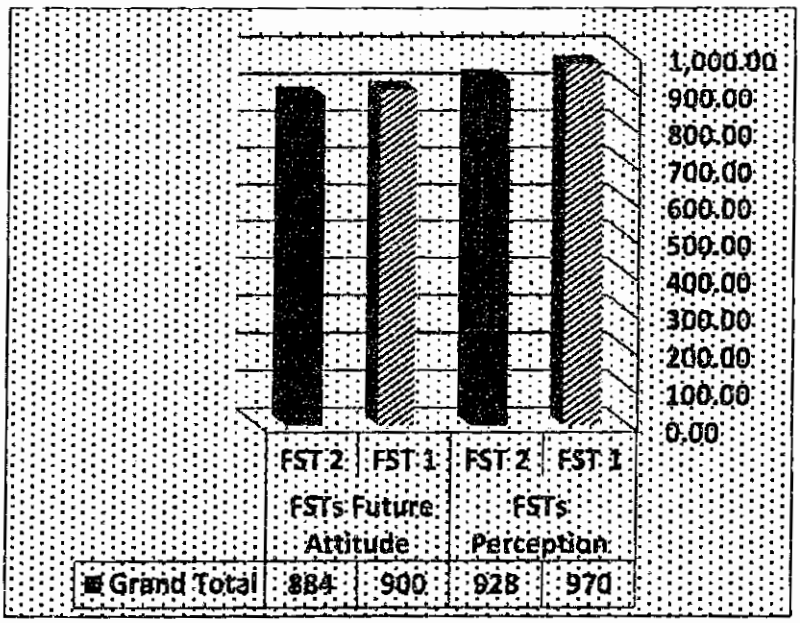


Figure (1) Total score of perceptions and future attitudes

Figure (1) shows that FST (1) is slightly better than FST (2), from the participants' perspective, in terms of user perception and future attitudes. It is also clear that the total differences between the two FSTs, including the grand total of perception and future attitudes together, is 58, which means no significance difference between the two FST.

1- Other Comments:

The most important participant's comments have been collected and analyzed in the following table, which shows that 70 % of the participants would prefer to start with FST(1) in the future, 60 % might use FST(2) in the future and some of the participants commented that they have no preferences . Also 70 % found that both FSTs are providing easy way to find their resources and discover new resources.

Table (4) Participants Comments on FSTs

Comments	Task List	FSTs Comments	
		FST (1)	FST (2)
Positive Comments	It is Fast in searching and link resolving	8	4
	More effective than databases searching	8	3
	There is no way to direct show the full text	5	9
	Better than Google Scholar	5	2
	It allow searching within subject category	8	3
	It would be my future starting search	14	12
	Easy to use	14	6
	It allow discovering all what I might need	5	2
	Very Productive and efficient	6	2
	We should teach it to every body	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

Comments	Task List	FSTs Comments	
		FST (1)	FST (2)
Negative[5] Comments	Very Complex	-2	-9
	huge number of results retrieved	-6	-6
	it is not easy to narrow or boarded search	-5	-6
	Expected more complex search parameters	-7	-5
	There is no way to modify search	2 -	-6
Total		99	48

This table shows that in general participants have reported more positive comments and feedback about FST(1) than FST(2).

5 – Usability Test: Overall Analysis –

Figure (2) shows the overall grand total analysis of the usability test, which present the grand total of all the usability values. It is clear that FST (1) has achieved a higher values in general than FST (2). T test analysis also proved that there is a significance difference between the two FST as $\mu = 0.512204$, although the analysis shows that FST (1) is better than FST (2), user still see that FST(2) is a good alternative for FST (1) in case it is not available or down. The grand total values which have reported in this figure would be used to calculate the 38 % of the usability test in third phase of this study.

[5] Negative comments get a negative score while positive comments get a positive score

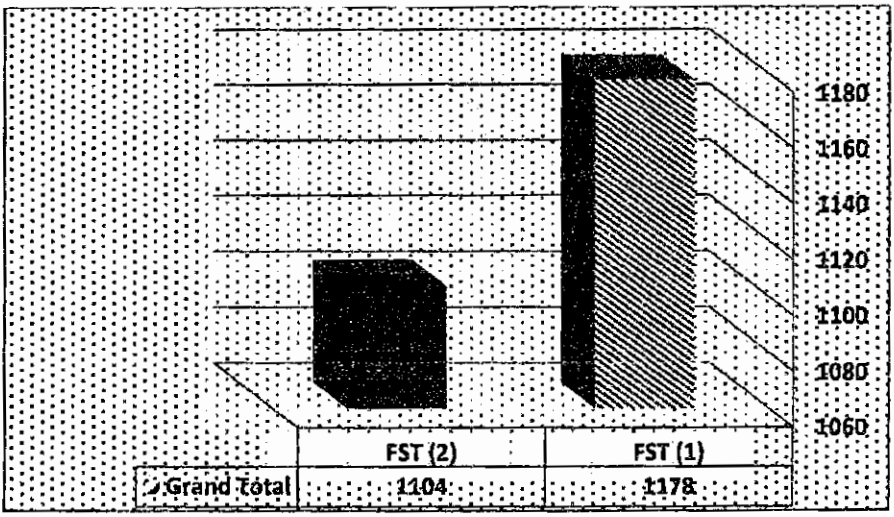


Figure 2 – The Grand Total of the Usability Test

Phase II: Retrieval Performance:

The retrieval performance has been tested for 20 real user queries representing real information needs. FSTs performance has been evaluated using the standard measurements in IR including recall and precision. The following part would explain the assessment procedures and results for each step.

• Federated Search Tools Recall:

Recall of FSTs has been calculated by comparing the retrieved results from each FST with the native databases performance for the 20 queries. Each query has been submitted to 6 different interfaces (two federated search tools and 4 native databases interfaces). The search process has been refined to submit the queries to the selected databases titles search field regardless whether the query is submitted indirectly through the FSTs or directly through the native interface of the database. The total number of the retrieved items has been compared by calculating the deviation of each FST from the native interfaces for the 4 original databases. Appendix (1) shows the results of this step:

It is clear that FST (1) is much more effective in terms of recall calculation as it has been more consistent with the original databases in

most of the cases, while FST(2) has deviated from the original databases results in 18 queries, which retrieved results, as there are two queries have retrieved zero results. Spearman correlation coefficient has also been calculated to explore the relationship between each FST results for the 18 queries and the total results retrieved from the 4 databases. The analysis shows that FST(1) is more consistent in terms of retrieving results coincide with the total number of results retrieved by the 4 databases with a correlation coefficient equal 0.98 compared to 0.92 for FST(2), which means that FST(1) is more consistent with the original databases in terms of recall value than FST(2). This part shows that recall could be easily used and tested to calculate the efficiency of FSTs IR performance by calculating recall consistency and correlation with original and native databases. The correlation score would be used in the final calculation step to report the overall evaluation.

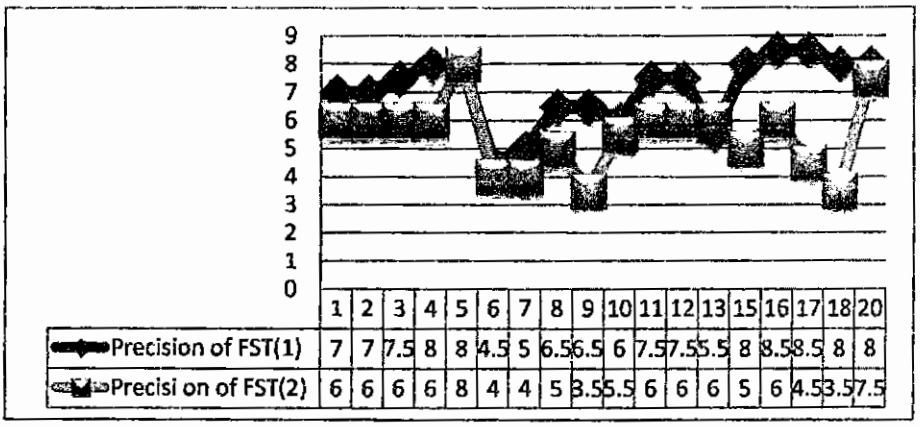
- **Precision at 11 point Cut off Recall (P11):**

Precision at 11 point cut off recall has also been computed using the recall level at the standard 11 points. A common method is used to compute the 11-point average precision by averaging the precision over the standard recall points (0%,10 %, 20%, 30%, etc.). To get the precision for these standard recall points, precision and recall for each relevant document in the result set has been calculated and interpolated. These standard levels allow measuring the performance in the different areas of the retrieved results distribution. For example if the system retrieved only 4 relevant documents out of 10 at points 2, 3, 5, and 7, then at recall point 0.30 precision is 2/3 = .667 since among the top three documents only two documents are relevant. At recall point 0.60 precision is 3/6 = 0.50 since among the first 6 documents three documents are relevant. At recall point 0.90 precision is 4/9 = 0.444 and so on.

$$P11 = \frac{\sum \text{precision relevant, Q}}{N}$$

Where N = 20 queries

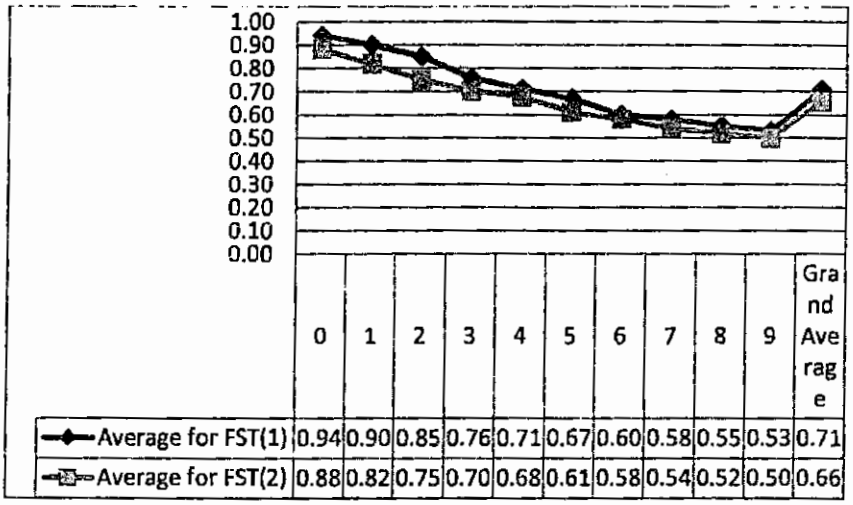
Each participant has evaluated the first 10 items retrieved for his query in terms of a relevancy scale of 3 points (0- Irrelevant, 0.5, Partially Relevant, and 1- Relevant). The following figure shows the results of the precision analysis for each query.



	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 FST (1) and FST (2) for each query

This figure shows that FST(1) has retrieved more precise results than FST (2), and it is more effective in terms of precision values than FST(2) in 16 cases out of 18 and coincide in one case (number five) while FST (2) was by chance more effective in query number 13 because the different is not large. Note that two queries have been removed because they have retrieved zero results.



This figure also shows that FST (1) is much more effective in terms of precision in all positions and in average, which means that in all position FST (2) is much more consistent with user perception than FST (2).

Phase III: Overall Evaluation:

The overall values of the first two phases have been interpolated according to the baseline weights which have been assigned by the participants in the beginning of the study. The next table shows the interpolated weights for each phase:

Table no. (5) The overall evaluation results

Evaluation Approach	Score	FST (1)	FST (2)	Total
Usability	Grand Usability Score	1178	1104	2282
	Interpolated Usability %	52%	48%	100%
	Interpolation Usability to 38 %	20 %	18 %	
Retrievability	Recall Correlation value	98	92	
	Precision at 11 cut off recall precision	71	66	

Evaluation Approach	Score	FST (1)	FST (2)	Total
	Average			
	Average Retrievability	84.5	79	
	Interpolation Retrievability to 62%	52.39 %	48.98 %	
Grand Total Usability and Retrievability		72.39	66.98	

The value of the overall evaluation has been calculated for each FST by reporting the usability percentage then interpolating it to .38 and calculating the average retrievability score from the values of recall correlation and precision at 11 point cut off recall average, then interpolating it to .62. It is clear that FST (1) has gathered a higher value than FST(2) in all the cases. It is also clear that usability and retrievability could be easily interpolated to compare between information retrieval systems in general and FST in particular.

Conclusion:

This paper presents a framework for comparative evaluation of two FSTs based on usability testing and retrievability performance. The proposed framework for evaluation consists of three phases including: usability testing, retrievability performance assessment, and overall comparison. Usability testing is usually conducted through users' task or expert testing. This study conducted a loud thinking user task protocol for usability testing. The retrievability performance evaluation is based on solid IR measurements including recall and precision. The overall evaluation and comparison combined the two approaches together to reach to an overall conclusion about the effectiveness and efficiency of FST from user and system perspective. The proposed model designed in this paper could be exploited in FSTs evaluation as it includes the most important indicators for evaluation as decided by the participants reflecting their perspectives

and attitudes. Participants decided that system retrievability is much more important than usability testing, although both of them are important. They assign a weight of 38 % reflecting the importance of usability and 62 % reflecting for the system retrievability importance. The final results show that the FST which is based on portal technology including search and link resolver is much more effective than FST which is based on library system technology.

References:

Avrahami, T. T., Yau, L., Si, L., & Callan, J. (2005). The FedLemur project: Federated search in the real world. Journal of the American Society for Information Science and Technology, 57(3), 347-358. doi: 10.1002/asi.20283

Bouchon, B. (1998). *Aggregation and fusion of imperfect information* (Vol. 12): Springer.

Boyd, J., Hampton, M., Morrison, P., Pugh, P., Cervone, F., & Scherlen, A. (2006). The one-box challenge: Providing a federated search that benefits the research process. *Serials Review, 32*(4), 247-254. doi: 10.1016/j.serrev.2006.08.005

Burke, L. (2001). The Future Role of Librarians in the Virtual Library Environment. *Australian Library Journal, 51*(1), 31-45.

Buck, S., & Nichols, J. (2012). Beyond the search box: Using participatory design to elicit librarians' preferences for unified discovery search results pages. doi: http://hdl.handle.net/1957/35951

Cervone, F. (2005). What we've learned from doing usability testing on OpenURL resolvers and federated search engines. *Computers in Libraries, 25*(9), 10-14.

Chen, X. (2006). MetaLib, WebFeat, and Google - The strengths and weaknesses of federated search engines compared with Google. *Online Information Review, 30*(4), 413-427. doi: 10.1108/14684520610686300

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). *Rank aggregation methods for the web.* Paper presented at the Proceedings of the 10th international conference on World Wide Web.

George, C. A. (2008). Lessons learned: usability testing a federated search product. *Electronic Library, 26*(1), 5-20. doi:

10.1108/02640470810851707

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, 13(2), 165-185.

Jacso, P. (2003). Savvy searching. *Online Information Review, 28*(6), 454-460.

Jakob, N. (1993). Usability engineering. *AP PROFESSIONAL-BOSTON, Capitolo, 2*, 191-194.

Kumar, S., Sanaman, G., & Rai, N. (2008). Federated Search: New Option for Libraries in the Digital Era. 267-285.

Lampert, L. D., & Dabbour, K. S. (2007). Librarian perspectives on teaching Metasearch and federated search technologies. *Internet Reference Services Quarterly, 12*(3-4), 253-278.

Liu, T., Wang, F., & Agrawal, G. (2012). Stratified sampling for data mining on the deep web. *Frontiers of Computer Science, 6*(2), 179-196. doi: 10.1007/s11704-012-2859-3

Mohamed, K. A., & Hassan, A. (2008). Web usage mining analysis of federated search tools for Egyptian scholars. *Program: electronic library and information systems, 42*(4), 418-435.

Mohamed, K. A. E.-F. (2006). *Merging Multiple Search Results Approach For Meta-Search Engines.* University of Pittsburgh.

Randall, S. (2006). Federated searching and usability testing: Building the perfect beast. *Serials Review, 32*(3), 181-182. doi: 10.1016/j.serrev.2006.06.003

Shneiderman, B. (1998). *Designing the user interface: strategies for effective human-computer-interaction*: Addison Wesley Longman.

Smeaton, A. F., & Crimmins, F. (1997). *Using a data fusion agent for searching the www.* Paper presented at the Poster presented at the WWW6 conference.

Soken, N., Reinhart, B, Vora, P & Metz, S. (1993). *Methods for Evaluating Usability Section 5B*: Honeywell.

Tchangalova, N., Stilwell, F. (2012). *Search Engines and Beyond: A Toolkit for Finding Free Online Resources for Science, Technology and Engineering. Science and Technology Librarianship.* Retrieved from http://www.istl.org/12-spring/internet1.html website: doi:10.5062/F4D21VHZ

Tennant, R. (2001). Digital Libraries: Cross-Database Search: One-Stop Shopping. Retrieved from

http://libraryjournal.reviewsnews.com/index.asp?layout=articlePrint&articleID=CA170458 website:

Wrubel, L., & Schmidt, K. (2007). Usability testing of a metasearch interface: A case study. *College & Research Libraries, 68*(4), 292-311.

Yuwono, B., & Lee, D. L. (1996). WISE: A World Wide Web resource database system. *Ieee Transactions on Knowledge and Data Engineering, 8*(4), 548-554.

Appendix (1) Total No. of items retrieved for each query from the different searching tools

No	Queries	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
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	165	37	16	9	51	113
11	Information Seeking Behavior	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	214	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		2033	2809					2238
Recall		90.8	125.5					
Correlation		0.98	0.92					