IntegraL: An Effective Link-based Federated Search Infrastructure

Shuyuan Mary Ho, Min Song, Michael Bieber, Eric Koppel, Vahid Hamidullah, Pawel Bokota

New Jersey Institute of Technology

College of Computing Sciences, Department of Information Systems

University Heights, Newark, NJ 07102

(973) 596-3000

[smho, song, bieber, erk7, vh22, pmb9] @njit.edu

http://smho.mysite.syr.edu http://web.njit.edu/~bieber http://web.njit.edu/~song

ABSTRACT

This research provides a new means for making digital library services interoperable. IntegraL facilitates a virtual restructuring of public web spaces and services, bringing authenticated digital libraries into broad "federated" digital library spaces constructed from numerous interrelationships. Elements of users' search interests reside within a rich context of meta-information that helps users understand and work with them. This provides a ripe environment for organizations and individual people to develop small, specialized collections and services, which automatically become part of the federated space and accessible to those they can benefit. IntegraL extends the boundaries of how we think about and interact with digital libraries.

General Terms

Systems, Design, Experiments

Keywords

Digital library infrastructure, Federated search

1. INTRODUCTION

IntegraL is a scalable, "lightweight" search infrastructure that brings a plethora of relevant resources directly to library users. IntegraL virtually integrates collections and services, including search services of libraries nationwide. It helps users to effectively search structured content information based on identified name entities across heterogeneous digital libraries. Users not only interact with the libraries and search engines just as before, but also see extra link anchors. Upon selecting one link anchor, IntegraL automatically generates a list of links to relevant documents, services and metadata (Song & Bieber 2008). IntegraL further provides recommendation features for search users (Im & Hars 2007). IntegraL allows the library systems to act as *information requesters* (a customized set of links embedded in display screens that widens search horizontally) and *information providers* (link anchors leading to a systems' documents and services vertically). IntegraL provides federated search across all relevant resources and helps users locate information more effectively. It also increases the accessibility and effective usage of library resources.

This paper contains six major sections describing this study. In the following section, we review why an infrastructure is necessary for integrating multiple digital libraries. In the third section, we describe the IntegraL system architecture and IntegraL's innovative way of expanding the use of multiple digital libraries, databases, and search engines at the user's preference. We present our research questions and hypotheses in the fourth section. In the fifth section, we describe the design and execution of a user study for this virtual integration infrastructure. In the sixth section, we discuss the results of our hypotheses testing, and conclude our study in the seventh section.

2. RELATED WORK

Rao (2004) described the progression of information search from the 60's to the 90's. Users' information search has been drastically enhanced from simple query-in, result-out in the 60's, to information digest, indexing, extraction, categorization, visualization, and further to federated research. While users' search capability has been empowered, the design and development of digital libraries have become more sophisticated. Information retrieval will be more based on open and flexible infrastructures (Kazai & Doucet 2008; Rao 2004). However, this scalable archival infrastructure facilitates the collaboration among heterogeneous digital libraries. Being able to accurately retrieve documents from distributed uncooperative digital libraries becomes critical with foreseen and unforeseen problems. They include issues with archival preservation of digital content, indexing in each collection, representable query phrase, merging and transforming retrieved data, effective use of metadata for search, robust retrieval algorithms, seamlessness interactions between user and the data, integration between services and tools, and inevitably privacy and security considerations when accessing data for sensitive purpose.

Merging results from different databases and search engines requires acquisition of database resource description, selecting from collection, and merging results into a single rank list (Si & Callan 2002). In merging high volume data streams in a web-based infrastructure, Mazzucco and Ananthanarayan (2002) uses data mining to processing streams of data, extract patterns and anomalies.

Federated search (or, distributed information retrieval) has been discussed extensively in the research community. Open Archival Initiative (OAI) is a framework that provides search services over aggregated metadata among federated digital libraries for both service providers and data providers (Lagoze & Van de Sompel

2001). Maly and Zubair, et al. (2005) researched a Grid-based architecture for parallel harvesting among large amounts of computing resources to be shared across organizational boundaries. This federated digital library architecture indexes and harvests hundreds of metadata from data providers. This type of architecture requires extensive load balance and may still suffer insufficient service performance if low bandwidth of the data providers and high volume of the harvest nodes are encountered. On the other hand, a user modeling approach to full-text federated search focuses on collecting user behavior by analyzing a user's long-term persistent interests based on user's past queries (Lu & Callan 2006). While this approach may enhance the robustness of the search, redundant documents from the static search collections may lead to unnecessary processing costs in federated search in an uncooperative environment. Shokouhi and Zobel (2007) studied how different queries used can reduce the overlap in search results from dynamic collections. Shokouhi, Baillie, et al. (2007) further studies accurate retrieval in updating dynamic search collection for federated search. Different retrieval process algorithms that enhance recall and precision are studied in uncooperative distributed search environments (Callan & Connel 1999; Callan, Lu et al. 1995; Paltoglou, Salampasis et al. 2007; Paltoglou, Salampasis et al. 2008; Si & Callan 2003; Si & Callan 2005).

Not only are the synchronous operations of the architecture among multiple harvesting nodes important to federated search, the ability to harvest item-level metadata would also enhance the performance of federated search. The advent of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) helps the sharing and harvesting of item-level descriptive metadata for selected digital resources (Arms, Dushay et al. 2003; Hagedorn 2003; Simon & Bird 2003). Foulonneau and Cole, et al. (2005) states that this granular collection-level descriptive metadata provides attributes to the retrieved documents, which would bring more relevant documents in response to a query.

Wu and Li, et al. (2006) suggested noun phrase used as key-phrase in automatic text extraction; these key-phrases can be used as document metadata for web searching (Li, Wu et al. 2004; Wu, Li et al. 2006). Bot and Wu, et al. (2005) used these extracted key-phrases as topical oriented categories from the retrieved documents to correspond to different semantic aspects of the query. Highlight, as one example of metadata search engine, is composed of document acquisition, document pre-classification and automatic concept hierarchy generation. Document acquisition retrieves documents in response to a query. The document pre-classification module classifies documents into pre-defined categories. Then, based on the extracted key-phrases, the hierarchy module automatically generates individual concept hierarchies within each active category (Bot, Wu et al. 2005).

3. SYSTEM ARCHITECTURE

IntegraL offers a scalable infrastructure that links among heterogeneous digital libraries through the development of a web-based proxy server, sitting on top of Tomcat, an open source servlet container as the outer dotted line represented in Figure 1. When a user makes a request to access a digital library, he or she is authenticated by a single sign-on (SSO) mechanism. We adopt an open source authentication mechanism, Shibboleth, in order to allow seamless browsing across digital libraries that have also adopted Shibboleth. Once the user is authenticated, this SSO mechanism allows the user to surf among various subscribed digital libraries without going through repetitious logins at each stage. The user's credential information is stored in hash files on the proxy.

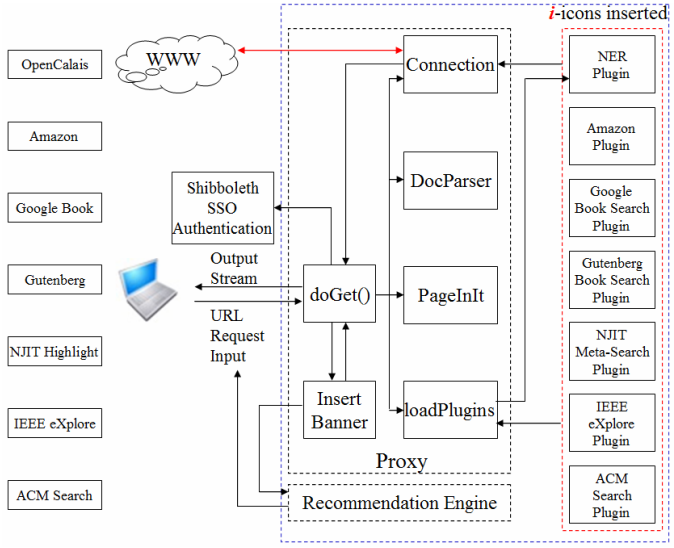

Figure 1: System Architecture

All users' requests are handled by the proxy. When a user browses any other subscribed digital library, their request to access the other digital library is taken care of by the proxy. If a user requests to use other search engines, the pages of their request are returned to the user untouched. When a user makes a URL request, an IntegraL banner is inserted on top of the digital library. All HTML pages are converted into DOM Documents; all relative URL paths are converted to absolute URL paths. What we innovatively create is the addition of i-icons (Figure 2). The i-icons are inserted whenever a name entity is recognized by OpenCalais, a service that annotates data with rich semantic metadata. The name entity recognition module, or NER plugin, receives the document requested by the user from the proxy, analyzes the lexical meaning of recognized categories, such as person, location, etc., and then creates rich semantic metadata that serves as recommended search for user's further references.

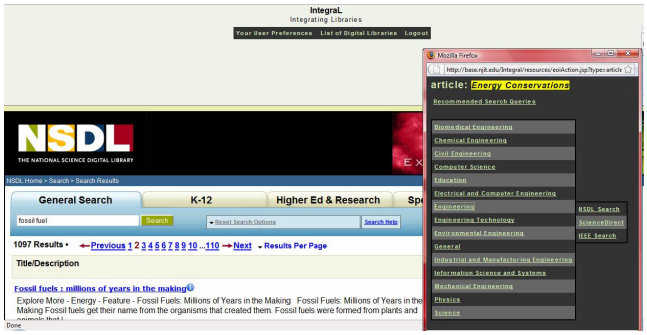

Figure 2: IntegraL links with NSDL

Illustrated on the right-hand side of the Figure 1, plugins are developed using XPath, which parse the HTML documents and insert i-icons wherever elements of interests have been located. The development of plugins is based on templates. This provides flexible expansion of integrating search engines and digital

libraries. IntegraL's web-based administrative interface empowers users with ad hoc configuration.

The strengths of IntegraL lie in its capability to include a federated search engine and recommendation engine that empowers users' deep search. The elements of interests (EOIs) allow ease of integration among heterogeneous databases and library resources. The function of name entity recognition plugins are dynamically plugged in to recognize more named entities: (1) EOIs (elements of interests) are statically defined by the layout-based plugins in the proxy. The plugin templates allow ease of integration among heterogeneous databases. (2) The NER (name entity recognition) plugins are dynamically plugged in to recognize more named entities. (3) Its capability to include the federated search engine and recommendation engine that empowers users' deep search.

4. RESEARCH QUESTIONS

In order to understand how effectively IntegraL can assist users to conduct advanced, deep search through this virtual integration of libraries, we plan to answer the following research questions.

1. Can enhanced access to information through virtual integration help users find information more effectively and to perform tasks involving library resources more effectively?
2. Can enhanced access to information through virtual integration increase the accessibility and utilization of library resources?

4.1 Hypotheses Testing

We form three research hypotheses in order to answer our research questions.

H1: Users perform better objectively when they use IntegraL than without IntegraL.

H2: Users perceive more effective searching subjectively when they use IntegraL than without IntegraL

H3: Users utilize and access more various virtually integrated databases with IntegraL than without IntegraL.

5. METHOD

5.1 Experiment Design

A *within-subject* experiment using a step-by-step approach that helps users learn how to search scholarly articles.

5.2 Data Collection

We conducted a pilot usability study as well as a main experiment during Fall 2009. The total of 139 participants from freshmen in a Physics class, above 18 years of age, participated in the search experiment in October 2009. 7-point Likert scale survey instruments were used. All data were collected in three stages: before, during, and after the experiment. Four different ways were designed to collect our data.

Four different were designed to log our data.

1. Participants' demographics were surveyed.
2. Participants' clickstreams were logged.
3. Participants' objective performance that measured the quality of their search, including relevancy of reference and citation, the use of scholarly database, participants' own judgment of the relevancy of the articles found, and participants' own reasons for ranking the reference lists. The performance measures were rated by 4 graders with Cronbach's alpha value of 0.965.
4. Participants' perceived satisfaction and effectiveness of their search were collected in the post-task survey.

5.3 User Task

All participants were given two tasks to complete. The two tasks were designed with similar search steps. Each participant took about 20-25 minutes to work on each task. The amount of time for the completion of these two tasks was totaled as 45-50 minutes maximum. We scheduled a five minute orientation session before the experiment started; this helped participants get situated easily.

5.4 Rotation of Conditions and Tasks

There are altogether 8 combinations of two tasks (T1, T2) rotated within one treated condition with IntegraL, and one baseline condition for all participants. In order to evenly distribute task assignments with different combinations of systems conditions, the system conditions are controlled centrally at the proxy server and the rotation of the task assignments are combined and controlled at the handouts. In other words, there are eight combined system conditions and there are four sets of task assignment combinations. Table 1 illustrates how these combinations of task assignments and system conditions should be randomized. With this rotation design of user's tasks and system conditions, the threats to validity occurring within subjects can be eliminated.

Table 1: Rotation of Conditions and Tasks

Tasks Rotation			Systems Conditions	
T1	T2	×	IntegraL	Baseline
			Baseline	IntegraL
T2	T1		IntegraL	Baseline
			Baseline	IntegraL
2 sets of tasks rotation			4 sets of system conditions	
2 (tasks) * 4 (conditions) = 8 combined tasks & conditions				

5.5 Data Analysis

The data we used to test the hypotheses were from the experiment in Fall 2009. Data were aggregated and cleaned based on the criteria whether all four aspects of data were completely collected. Data were synchronized and reorganized based on a unique identifier. A complete and useful dataset was reduced to 65 out of the 139 dataset. This small sample size contained confounding effects and affected our hypotheses testing results. Descriptive statistics were conducted to understand general participants' demographics, user's perceptions, and objective measures. We used open source R statistics tool and SPSS to run our analysis. Hypotheses were tested. The results showed that null hypotheses were rejected, and our research hypotheses were supported.

5.6 Threats to Validity

This experiment does contain some threats to validity. However, we were able to justify in our research design how to reduce those foreseeable threats.

1. The design of systematically rotating tasks and conditions eliminates threats to internal validity.

2. The grading rubric provides initial guiding principles of judging user search performance. This inter-rater reliability eliminates threats to criterion validity regarding a user's objective search quality and search performance.
3. Participants usually have high skills in experiencing multiple search engines. This pre-existing search engine experience creates inequality judgments and participants' bias when users come to use a new different ways of search. This threat was taken care of in the design of how we answer our research question. We choose to use objective outcome measures on users' quality of search rather than using objective clickstream and the amount of time spent on completing one task, because the clickstream and the amount of time spent on using IntegraL depict users exploration of this virtual integration of library services, rather than completing a task.

6. DISCUSSIONS

6.1 Existing Library Experience

On a 7-point Likert scale, with 1 as least experienced and 7 as most experienced, most of 65 participants were very skillful with the mean of 5.94 in using the search engines (Figure 3). However their experience of using a digital library such as NSDL, ACM, IEEE and Science Direct, was very slim with the mean 3.45 (Figure 4).

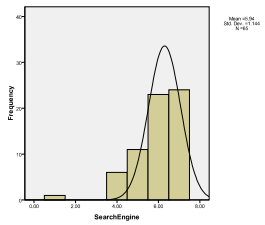

Figure 3: Existing Search Engine Experience

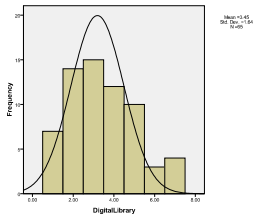

Figure 4: Existing Digital Library Experience

6.2 Outcome Measures of User's Performance

Users generally perform about the same level of search results, however if we compare user objective performance in quality search (Figure 5), users who used IntegraL virtual services would perform higher quality search results than without.

The positive skew distribution presented in Figure 6 depicts that IntegraL helps to enhance search productivity. Users find IntegraL useful in locating relevant information. IntegraL also helps users to do tasks more effectively.

Participants' performance were measured by 4 graders with Mean of bivariate correlations R value, 0.839 (p=0). Because of the *within-subject* experimental design, each participant worked on one task in a treatment condition (with experience of IntegraL), and the other task in the baseline condition (without experience of IntegraL). When comparing the outcome performance measures between the treatment condition and baseline condition, we discovered that participants who used IntegraL system are likely to obtain higher performance measures than those participants who did not use IntegraL system, $t(128) = 1.409$, p=0.002. The t value of 1.409 is not more extreme than the cutoff t of 2.364. Our findings could not reject null hypotheses and this is probably caused by our small sample size of only 65 participants in dataset.

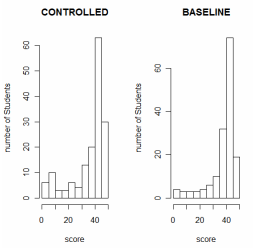

Figure 5: Comparison of User's Objective Performance

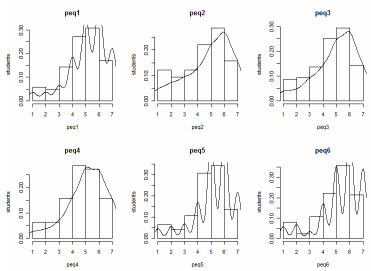

Figure 6: Perceived Effectiveness of Search

However, we conducted a one-way ANOVA estimating population variance in the outcome performance measures from variation within each sample. The $F(1, 128)=10.097$, p=0.002 is more extreme than the cutoff F of 6.64, meaning we could reject the null hypothesis; the research hypothesis that users perform better objectively when they use IntegraL than without IntegraL is supported.

H1: *Users perform better objectively when they use IntegraL than without IntegraL.*

6.3 Users' Attitude and Perceived Acceptance

Users have neutral attitude toward IntegraL, which is possibly due to many good search engines available to users, and that users tend to be used to existing ways of search (Figure 7).

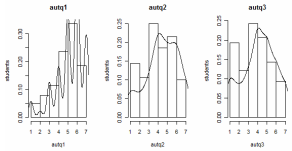

Figure 7: Attitude

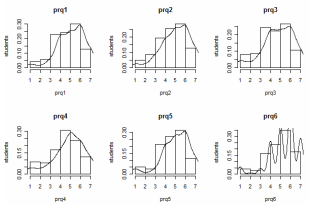

Figure 8: Perceive Performance of IntegraL over other search engines

Users are confident of IntegraL's ability to provide satisfactory search results as illustrated in Figure 8. When comparing IntegraL with other regular search engines as designed in our baseline condition, most users perceive IntegraL to provide better results than those currently available on the market.

We conducted a t-test for a single sample of all independent variables on user perception after they experienced IntegraL system. The results prove all t values are more extreme than the cutoff t of 2.654. Therefore, reject the null hypothesis; the research hypothesis is supported.

We then compared user's subjective perception from before the task to after the task. We conducted a t-test for paired samples, and derived $t(64)=1.769$, p=0.082. The t value 1.769 is more extreme than the cutoff t of 1.669 (two-tailed t-test). Therefore, we reject the null hypothesis; the research hypothesis is supported.

H2: *Users perceive effective searching subjectively when they use IntegraL than without IntegraL.*

We conducted a principal factor analysis to reduce construct dimensions and identified 5 factors of which the Eigenvalues are above 1 (Table 2).

Table 2: Principal Factorial Analysis

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	14.363	59.848	59.848	14.363	59.848	59.848	4.807	20.029	20.029
2	1.780	7.419	67.266	1.780	7.419	67.266	4.796	19.982	40.011
3	1.331	5.547	72.813	1.331	5.547	72.813	4.670	19.459	59.470
4	1.122	4.675	77.488	1.122	4.675	77.488	3.325	13.856	73.326
5	1.026	4.274	81.762	1.026	4.274	81.762	2.025	8.436	81.762

Extraction Method: Principal Component Analysis.

Then, we used Varimax with Kaiser Normalization to run a factor analysis of a correlation matrix and identified the survey response items to be closely correlated. All the items were positively correlated with statistical significance.

6.4 Users' Effective Utilization of Virtual Integration

In order to find out whether users access more various virtually integrated resources, we compared the number of clicks for baseline and treatment conditions, and the amount of time spent for baseline and treatment conditions. We conducted a two-tailed t-test for independent means for two groups on the number of clicks, $t(128)=-1.703$, $p=0.091$. The t value -1.703 is more extreme than the cutoff t of -1.645. Therefore, we reject the null hypothesis; the research hypothesis is supported. We also ran a two-tailed t-test for independent means for two groups on the amount of time spent, $t(128)=-4.855$, p=0.000. The t value -4.855 is more extreme than the cutoff t of -2.576. Therefore, we reject the null hypothesis; the research hypothesis is supported.

We conducted a one-way ANOVA estimating population variance in the number of clicks from variation within each sample. The $F(1, 128)=0.046$, $p=0.831$ is not more extreme than the cutoff F, we could not reject the null hypothesis; the research hypothesis that users access more various virtually integrated databases than without IntegraL is not supported. This finding showed that if we use only the number of clicks to measure the amount of accessibility to other library resources through IntegraL virtual integration is still insufficient. We may need to combine hypothesis 3 with hypothesis 2 (which involves more participants' subjective views.) But, we conducted a one-way ANOVA estimating population variance in the amount of time spent from variation within each sample. The $F(1, 128)=6.568$, $p=0.012$ is more extreme than the cutoff F, we could reject the null hypothesis; the research hypothesis is supported that users tend to utilize more various virtually integrated databases than without IntegraL.

H3: *Users utilize and access more various virtually integrated databases with IntegraL than without IntegraL.*

Moreover, participants of this experiment gave us feedback about their experience of this infrastructure that provides the virtual integration among digital resources. Below are some quotes from the participants.

> "*I really enjoyed using IntegraL because other search engines I have used never gave me relevant references I was looking for. It takes longer to search for your topic using Google, Ask.com, etc whereas IntegraL makes it easier, faster, and it really beneficial.*"
>
> "*I liked using this method because it is easier to find similar topics.*"
>
> "*IntegraL was easier to use than that of Google or yahoo. I was happy with the I-icon and the availability of other resources at one time.*"
>
> "*The Integral system is very useful. It compiles many different search engines in one site so that you do not have to go about looking separately for them. They also give articles from credited sources so that you know they are professional journal entries.*"
>
> "*IntegraL was a very helpful system, it definitely cut my searching time in half and it allowed me to find precise articles pertaining to my article.*"

7. CONCLUSION

Without the complication of complete system integration, IntegraL adopts a light-weight approach that links multiple heterogeneous digital libraries and search engines. It allows interoperability among different search results, search engines and digital libraries. The system itself is mostly built on open-source software, which can be reliable, auditable and cost-effective. This approach provides recommendations to users for further deep search based on identified elements of interests among wide ranges of virtual resources.

Our study showed that participants *perceive more effective searching subjectively when they use IntegraL than without IntegraL* (hypothesis 2), this answered our first research question that enhanced access to information through IntegraL virtual integration would help users find information more effectively. Our study also showed that participants *perform better objectively when they use IntegraL than without IntegraL* (hypothesis 1). This also demonstrates that IntegraL helps users to perform tasks involving library resources more effectively. We also answered our second research question that enhanced access to information through IntegraL virtual integration has helped the users to *utilize more various virtually integrated databases than without IntegraL* (hypothesis 3) and users *perceive IntegraL to be effective in searching activities* (hypothesis 2).

8. ACKNOWLEDGEMENTS

Partial support for this research was provided by the National Science Digital Library (NSDL) under grant LG-02-04-0002, by the National Science Foundation under grants DUE-0434581 and DUE-0434998, and by the New Jersey Institute of Technology. The authors thank Xiangmin Zhang for his advice on the

experimental design, and many graduate assistants who worked on data collection and analysis of this project.

9. REFERENCES

[1] Arms, W. Y., Dushay, N., D., F. and Lagoze, C. (2003) A case study in metadata harvesting: the NSDL., Library High Tech, 21, 228-237.

[2] Bot, R. S., Wu, Y. B., Chen, X. and Li, Q. (2005) Generating Better Concept Hierarchies Using Automatic Document Classification, CIKM'05, Bremen, Germany, pp. 281-282.

[3] Callan, J. P. and Connel, M. (1999) Query-based sampling of text databases, ACM Transaction of Information Systems, 19, 97-130.

[4] Callan, J. P., Lu, Z. and Croft, W. B. (1995) Searching distributed collections with inference networks., SIGIR'95, New York, NY, USA, pp. 21-28.

[5] Foulonneau, M., Cole, T. W., Habing, T. G. and Shreeves, S. L. (2005) Using collection descriptions to enhance an aggregation of harvested item-level metadata, JCDL'05, Denver, Colorado, pp. 32-41.

[6] Hagedorn, K. (2003) OAIster: a "no dead ends" OAI service provider, Library High Tech, 21, 170-181.

[7] Im, I. and Hars, A. (2007) Does a one-size recommendation system fit all? the effectiveness of collaborative filtering based recommendation systems across different domains and search modes, ACM Transactions on Information Systems (TOIS), 26.

[8] Kazai, G. and Doucet, A. (2008) Overview of the INEX 2007 Book Search track: BookSearch '07.

[9] Lagoze, C. and Van de Sompel, H. (2001) The Open Archives Initiative: Building a low-barrier interoperability framework, Proceedings of teh First ACM/IEEE Joint Conferent on Digital Libraries, Roanoke, VA.

[10] Li, Q., Wu, Y. B., Bot, R. S. and Chen, X. (2004) Incorporating Document Keyphrases in Search Results, The 10th Americas Conference on Information Systems, New York, New York, pp. 1-8.

[11] Lu, J. and Callan, J. P. (2006) User Modeling for Full-Text Federated Search in Peer-to-Peer Networks, SIGIR'06, Seattle, Washington, pp. 332-339.

[12] Maly, K., Zubair, M., Chilukamarri, V. and Kothari, P. (2005) GRID Based Federated Digital Library, CF'05, Ischia, Itala, pp. 97-105.

[13] Mazzucco, M., Ananthanarayan, A., Grossman, R. L., Levera, J. and Rao, G. B. (2002) Merging Multiple Data Streams on Common Keys over High Performance Networks, Proceedings of the 2002 ACM/IEEE Conference on Supercomputing, Baltimore, Maryland, pp. 1-12.

[14] Paltoglou, G., Salampasis, M. and Satratzemi, M. (2007) Hybrid results merging, CIKM'07, New York, NY, USA, pp. 321-330.

[15] Paltoglou, G., Salampasis, M. and Satratzemi, M. (2008) Integral Based Source Selection for Uncooperative Distributed Information Retrieval Environments, LSDS-IR'08, Napa Valley, California, pp. 67-74.

[16] Rao, R. (2004) From IR to Search, and Beyond, Queue: Open Source Grows Up, 2, 66-73.

[17] Shokouhi, M., Baillie, M. and Azzopardi, L. (2007) Updating Collection Representations For Federated Search, SIGIR'07, Amsterdam, The Netherlands, pp. 511-518.

[18] Shokouhi, M. and Zobel, J. (2007) Federated Text Retrieval From Uncooperative Overlapped Collections, SIGIR'07, Amsterdam, The Netherlands, pp. 495-502.

[19] Si, L. and Callan, J. P. (2002) Using Sampled Data adn Regression to Merge Search Engine Results, SIGIR'02, Tempere, Finland, pp. 19-26.

[20] Si, L. and Callan, J. P. (2003) A semisupervised learning method to merge search engine results, ACM Transaction of Information Systems, 21, 457-491.

[21] Si, L. and Callan, J. P. (2005) Modeling Search Engine Effectiveness for Federated Search, SIGIR'05, Salvador, Brazil.

[22] Simon, G. and Bird, S. (2003) Building an Open Language Archives Community on the OAI Foundation, Library High Tech, 21, 210-218.

[23] Song, M. and Bieber, M. (2008) IntegraL: Lightweight Link-Based Integration of Heterogeneous Digital Library collections and Services in the Deep Web, 10th IEEE Conference on E-Commerce Technology and the 5th IEEE Conference on Enterprise computing, E-Commerce and E-Services pp. 369-375.

[24] Wu, Y. B., Li, Q., Bot, R. S. and Chen, X. (2006) Finding Nuggets in Documents: A Machine Learning Approach, Journal of the American Society For Information Science and Technology, 57, 740-752.