

Combined Searching of Web and OAI Digital Library Resources*

Aaron Krowne and Martin Halbert
Emory University, Library Systems
Atlanta, GA 30322
{akrowne,mhalbert}@emory.edu

ABSTRACT

In this paper, we describe an experiment in combined searching of web pages and digital library resources, exposed via an Open Archives metadata provider and web gateway service. We utilize only free/open source software components for our investigation, in order to demonstrate feasibility of deployment for all institutions.

Categories and Subject Descriptors

H.3.7 [Information Systems]: Information Storage and Retrieval—*Digital Libraries*

General Terms

Experimentation, Human Factors

Keywords

digital libraries, Open Archives, OAI, DP9, crawlers, search engines

1. INTRODUCTION

Digital libraries are increasingly aiming to integrate valuable web resources, identified by human expert or automatic means, together with “native” DL records. Currently there are means to acquire web resources (lists of URLs, manually constructed or crawled) and DL records (harvested via the Open Archives Protocol for Metadata Harvesting), but these two remain separate both in representation and service functionality.

The MetaScholar Initiative [3] found that users do not understand why DLs must keep web and native resources separate, and do not find this method of organization particularly easy to use. There is a need, then, for DL services to more completely and meaningfully integrate resources from these two worlds.

*We would like to thank the Andrew W. Mellon foundation for supporting this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL'04, June 7–11, 2004, Tucson, Arizona, USA.
Copyright 2004 ACM 1-58113-832-6/04/0006 ...\$5.00.

In this article we discuss an experiment with such a web and native DL integrated search system, built using “off-the-shelf” open source software. We selected searching as it is a major type of DL service, and the results of MetaScholar were in the context of DL searching. For this experiment we take the approach of unifying both DL and web resources via a web crawl, then building a search system on top of this union. In future work we plan to explore the “opposite” approach; unifying all resources into an Open Archives framework as the basis for the search system.

2. EXPERIMENTAL SYSTEM

The general strategy for constructing our experimental combined search system (with software used) was:

1. Expose digital library records via an OAI provider (ARC).
2. Expose Open Archives version of DL records to web crawls via a gatewaying service (DP9).
3. Build an index via crawling the union set of {WWW resource URLs, the gatewaying service entry point URL} (Swish-e).
4. Build a search service over the union index (Swish-e).

DP9 is a gateway service which is part of the ARC digital library software suite from ODU [2]. It allows any type of web agent to view the resources of an OAI repository through standard HTTP requests. Swish-e is a free, open-source search engine which can build its index via web crawling [4].

DP9 is a gateway service which is part of the ARC digital library software suite from ODU [2]. It allows any type of web agent to view the resources of an OAI repository through standard HTTP requests. If the entry point to a DP9 gateway is accessible from the home page of the digital library, its native resources can be indexed by web search engines. In our system, we used DP9 to a similar end, albeit with our own search engine.

Swish-e is a free, open-source search engine which can build its index via web crawling [4]. After examining many alternatives [?], we selected Swish-e, due to its ability to support searching by META tags and other reasons. Because DP-9 exposes metadata fields as META tags, Swish-e's META tag support means that native DL metadata need not be lost when building a combined web search system via crawling.

Our test setting was the AmericanSouth.org digital library. The WWW resources of our experiment were the Web Links from AmericanSouth, which were identified by subject scholars as valuable resources.

3. ANALYTIC METHODOLOGY

We probed the quality of our system by repeated focused searches for *specific* web or DL resources, using subject scholars in a real-world setting. We analyzed the results by looking at the metrics of *rank of the desired item* in the results list and the *number of queries* issued to fulfill the information need. We did not use precision and recall, as these metrics are meant for *sets* of documents. With the time limitations of our study, we found it more tractable for users to search for a *specific document* and look for this document in the results list, then average scores over many instances¹. Due to this, our experiment more closely resembles the “real world” situation.

We built five alternative search engine indexes, and compared the same searches issued on all of them to determine the quality of our combined system. As we only had a single user session, we used simulations to approximate the effect of having users repeat their searches in all of the search systems. To do this, we recorded their queries and the positions of results, and constructed a system to “play them back” on any search system and record the new results.

The combined search systems were **DL** for “digital library”, containing only the OAI-exposed digital library records; **SW** for “shallow web”, containing only top-level pages of the web links; **DW** for “deep web”, containing all pages up to two links from the weblinks top-level pages, inclusive; **DL+SW**, containing both DL and SW items, and **DL+DW**, containing both DL and DW items. In addition to looking at the objective quality of a combined system like DL+SW, this setup enables us to measure any negative effect on results due solely to combining items from the two different domains.

3.1 EXPERIMENT

Our subject scholars were stationed at a computer with two web browser windows, one opened to the AmericanSouth.org web site and one opened to our test Swish-e search engine. They were given instructions to pick and memorize any three AmericanSouth.org “Research Collection” items (our native DL resources) and any three AmericanSouth “Web Links” items, recording the URL or identifier of each. A brief reminder sufficient to get the “gist” of each resource was recorded/memorized.

After the selection phase, the searchers switched to the browser window containing the combined search system and attempted to find each resource, issuing up to three queries to do so. They were told to search as they normally would, recording the result list position of the resource when found². At the end, they were also given a short 3-question survey

¹With this approach, note that precision beyond the first few items doesn’t matter, as users do not normally look beyond the first page of results (or even the first few items) before issuing another query. In addition, recall will always be $1/n$ when the desired item is found within the first n results.

²If the target item was not found on the first page of results, it was considered not found for that query. We think this is both more stringent and more realistic than most IR studies.

Table 1: Baseline results for uncombined search systems.

statistic	DL	SW	DW
% items found	100	100	53
avg. # of queries needed	1.3	1.0	1.0
% of found in 1st or 2nd pos	70	100	25
avg. pos of found items	2.5	1.1	5.3
% found on 1st query	77	100	53

Table 2: Results for combined search systems.

statistic	DL+SW			DL+DW		
	all	dl	web	all	dl	web
% items found	100	100	100	60	69	53
avg. # of queries needed	1.1	1.3	1.0	1.1	1.3	1.0
% of found in 1st or 2nd pos	86	70	100	35	44	25
avg. pos of found items	1.6	1.3	2.1	4.2	3.4	5.0
% found on 1st query	89	77	100	50	46	53

asking their perceptions about the quality of the combined search system.

The results below are based on this recorded information, and focus on whether the combined search system successfully gives satisfactory results in a real-world sense.

4. RESULTS

Our five scholars each performed six search tasks, so there were thirty search tasks total. Two (both research collection tasks) had to be thrown out because the subject did not follow instructions. This leaves a sample set of 28 search tasks. There were 36 queries issued in attempt to fulfill these search tasks.

The sizes of the five search systems were as follows: DL+SW, 29,695; DL+DW, 35,420; DL, 29,611; SW, 84; DW, 5809. Thus, the reader should note that the shallow-web (SW) collection is very small compared to all of the other collections.

Table 1 shows some key data for the “uncombined” systems (systems which only had records from either the web or DL domain). This serves as a baseline for comparison with Table 2, which has the same statistics for the search systems that combined the digital library collection with either the shallow or deep web crawl.

Note that the stats in Table 2 are also separated out into “dl” and “web”, showing just how the digital library or web content subsets performed. These can be compared directly with the corresponding baselines.

Comparing DL+SW to DL and SW, we can see that this combined search system seems to be nearly lossless. All items were found in both the combined and uncombined systems. In fact the only change is that results positions in the combined system were slightly different, with DL dropping from 1.3 to 2.5, and web items rising from 2.1 to 1.1 (surprisingly, given the relative sizes of the two collections, this is the opposite of what one would expect).

The results for DL+DW appear to be not as good, with

the largest effect being the addition of the ~6,000 deep web resources hurting the results for searches within the ~30,000 DL resources. Search results for the web pages themselves did not change much. The survey results were also quite good, but are omitted for space reasons.

5. DISCUSSION

The above results seem to indicate that combining top-level web items, i.e. “home pages”, with DL resources in a combined search system works well. This is doable currently with freely available open source software. However, we uncovered problems in combining large web crawls with DL resources, even when the crawls were limited to two links deep.

A few qualifications are in order. First, one must remember that the results shown above for DL+DW, DL, SW, and DW are all *simulations*. These are based on “replays” of user queries which were actually only issued for DL+SW. During simulation, a failed query leads to issuing the next recorded query, which had been issued by the user for the same search task. However, sometimes such a query is not available, and even if it is, it is no longer being constructed in response to the current situation. Thus, our simulations provide what we believe is a very conservative *lower bound* on results.

Secondly, our “deep web” crawl was limited only by two criteria: (1) distance from top-level page ≤ 2 , and (2) page is at or below the same level as top-level page in URL hierarchy. These are the limitations of what we could achieve with Swish-e without considerable modification. Optimally, it would be more fitting to use a focused crawler [1] to protect from off-topic pages on the same web site, and even allow traversing the web in general, with good precision. We plan to explore this in the future.

Finally, we feel that deep-web results could have been much better had there been a way to “tweak” the search engine weightings for home/seed pages, particularly considering that these are the only types of pages our users searched for. This knowledge would be easy to pass to the search engine, but it is not supported in the search engine core (of Swish-e or any other search engine we know of). This is indicative of a general search object granularity/attribute-weighting problem that we plan to explore in the future.

6. CONCLUSIONS

We believe the results above show that a functional and useful combined web and DL resource search service can be built currently with free, Open Source software. The results are particularly good when chiefly “home pages” make up the web portion of the collection, and users mainly search for DL records and home pages. However, we think more work is needed to improve search results for combined search systems built on general collections.

REFERENCES

- [1] Chakrabarti, Soumen, et al., Focused crawling: a new approach to topic-specific Web resource discovery. Computer Networks vol. 31 nos. 11-16 (Amsterdam, Netherlands: 1999).
- [2] Liu, Xiaoming, Maly, Kurt, Zubair, Mohammed, Nelson, Michael L., *DP9: an OAI gateway service for*

web crawlers, JCDL 2002: 283-284 (Portland, Oregon, USA, June 14-18)

- [3] MetaScholar Initiative Homepage, <http://metascholar.org/>.
- [4] Swish-Enhanced (Swish-e) Homepage, <http://www.swish-e.org/>.