Augmenting OAI-PMH Repository Holdings Using Search Engine APIs

Martin Klein, Michael L. Nelson
Department of Computer Science
Old Dominion University
Norfolk, VA 23529
{mklein,mln}@cs.odu.edu

Juliet Z. Pao
NASA Langley Research Center
Hampton, VA 23681
juliet.z.pao@nasa.gov

ABSTRACT

In this poster, we give the preliminary results of our project to acquire Atmospheric Science Data Center (ASDC) project-related web resources, not with focused crawling, but by using the search engine (SE) APIs directly. We aggregate the results and create archive-ready complex objects.

Categories and Subject Descriptors: H.3.7 Information Storage and Retrieval[Digital Libraries] **General Terms:** Design, Management **Keywords:** Digital Libraries, Repository Enhancement, Search Engine API

1. INTRODUCTION

The NASA ASDC (http://eosweb.larc.nasa.gov/) maintains an Earth Science OAI-PMH repository [1] that contains earth science data consisting of 42 science projects with over 1700 data sets and 2M data granules in a combination of almost 2 petabytes of online and nearline storage. While the ASDC has the scientific data in its repository, it contains very little descriptive metadata for the associated projects. Most of the descriptive metadata that are available exists only in semi-structured or unstructured HTML pages and not in the ASDC repository itself. Consequently the ASDC has a great interest in ingesting structured data from external resources on the Internet into their repository.

2. CONCEPT

Instead of gathering pages with focused crawling [2], for each ASDC project we use the title and the project acronym as queries to the APIs of the three major SEs (Google, Yahoo! and MSN). For simplicity we just use the top ten results from each SE. In order to merge the three result sets (30 URLs total) we calculate the weight for each URL (W_{url}) where B stands for the amount of top results we are using (in our case 10) and $R(s)$ represents the rank of the returned URL in the result set of the SE where it was retrieved (s):

$$W_{url} = \sum_{s \in SE} (B - R(s)) + 1$$

This results in a complete list of retrieved results, ranked by their relative weight. The URLs are categorized, dedup'ed and eventually crawled. These are design issues and will not be addressed here. We create a complex object (encoded in MPEG-21 DIDL [3] and ready for inclusion into the ASDC's OAI-PMH repository) containing the resource identified by the URL and the appropriate metadata (e.g., timestamp of the crawl and our computed weights). For redundancy and preservation reasons we include the data object itself, both by value and by reference into the DIDL object.

3. CONCLUSION AND FUTURE WORK

We are aware that a project's title is generally insufficient to best describe the project, but the results of our experiments are indeed promising. For example, for the International Satellite Cloud Climatology Project (ISCCP) we retrieved 23 unique URLs (the top 5 results appeared in more than one result set) while using the top ten only. Using the top 25 results for the same project, we retrieved 64 unique URLs (overlap of 11) with the maximum weight of 50. Out of this result set, we can extract 39 unique top level domains (16 .gov, 8 .edu, 5 .org, 3 .com and 7 cc) which implies a great variety in the results.

Part of the ongoing work is to also query Digital Libraries in order to discover related publications. Search Engines do return non-HTML documents (PDF, DOC, etc) so parsing through these file formats is also a promising approach to discover relevant data. Given the specificity of the project names and acronyms, the top ten results have high precision. Our next step is to develop a semi-automated method for determining how deep in the results list to go to increase recall while maintaining high precision (based on ASDC administrator feedback).

4. REFERENCES

[1] C. Chu, W. E. Baskin, J. Z. Pao, and M. L. Nelson. OAI-PMH Architecture for the NASA Langley Research Center Atmospheric Science Data Center. In *Proceedings of the 10th ECDL*, pages 524–527, 2006.

[2] D. Bergmark. Collection Synthesis. In *Proceedings of the 2nd ACM/IEEE-CS JCDL*, pages 253–262, 2002.

[3] J. Bekaert, P. Hochstenbach, and H. Van de Sompel. Using MPEG-21 DIDL to represent complex digital objects in the Los Alamos National Laboratory digital library. *D-Lib Magazine*, 9(11), 2003.

JCDL'07, June 17–22, 2007, Vancouver, British Columbia, Canada.
ACM 978-1-59593-644-8/07/0006.