

Arc—An OAI Service Provider for Cross—Archive Searching

Xiaoming Liu, Kurt Maly, Mohammad Zubair, and Michael L. Nelson

Old Dominion University
Department of Computer Science
Norfolk, VA 23592 USA
+1 757 683 4017

{liu_x, maly, zubair, nelso_m}@cs.odu.edu

ABSTRACT

The usefulness of the many on—line journals and scientific digital libraries that exist today is limited by the lack of a service that can federate them through a unified interface. The Open Archive Initiative (OAI) is one major effort to address technical interoperability among distributed archives. The objective of OAI is to develop a framework to facilitate the discovery of content in distributed archives. In this paper, we describe our experience and lessons learned in building *Arc*, the first federated searching service based on the OAI protocol. *Arc* harvests metadata from several OAI compliant archives, normalizes them, and stores them in a search service based on a relational database (MySQL or Oracle). At present we have over 165K metadata records from 16 data providers from various domains.

Categories and Subject Descriptors

H.3.7 [Digital Libraries]: *collection, dissemination, standards.*

General Terms

Design, Experimentation, Standardization, Languages

Keywords

Digital Library, Open Archive Initiative

1. INTRODUCTION

A number of free on—line journals and scientific digital libraries exist today, however, there is a lack of a federated service that provides a unified interface to all these libraries. The Open Archive Initiative (OAI) [1] is one major effort to address technical interoperability among distributed archives. The objective of OAI is to develop a framework to facilitate the discovery of content in distributed archives. The OAI framework supports data providers (archives) and service providers. The service provider develops value—added services that are based on the information collected from data providers.

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, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL '01, June 24–28, 2001, Roanoke, VA.

Copyright 2001 ACM 1–58113–345–6/01/0006?\$.50.00.

These value—added services could take the form of cross—archive search engines, linking systems, and peer—review systems. OAI is becoming widely accepted and there are many archives currently or soon—to—be OAI compliant. *Arc* (<http://arc.cs.odu.edu>) is the first federated search service based on the OAI protocol, and its concept originates from the Universal Preprint Service (UPS) prototype [2]. We encountered a number of problems in developing *Arc*. Different archives have different format/naming conventions for specific metadata fields that necessitate data normalization. Arbitrary harvesting can over—load the data provider making it unusable for normal purposes. Initial harvesting when a data provider joins a service provider requires a different technical approach than periodical harvesting that keeps the data current.

2. Architecture

The *Arc* architecture is based on the Java servlets—based search service that was developed for the Joint Training, Analysis and Simulation Center (JTASC) [3]. This architecture is platform independent and it can work with any web server. Moreover, the changes required to work with different databases are minimal. Our current implementation supports two relational databases, one in the commercial domain (Oracle), and the other in public domain (MySQL). The architecture improves performance over the original UPS architecture by employing a three—level caching scheme [3]. Figure 1 outlines the major components; however, for brevity we discuss only the components relevant to this paper.

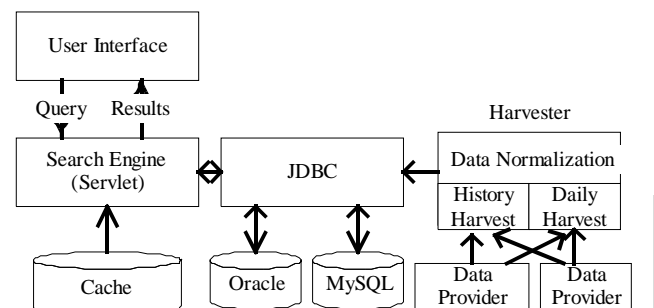


Figure 1. *Arc* Architecture

2.1 Harvester

Data providers are different in data volume, partition definition, service implementation quality and network connection quality. All these factors influence the harvesting procedure. Historical and newly—published data harvesting have different requirements. When a data provider joins a service provider for the first time, all past data (historical data) needs to be

harvested, followed by periodic harvesting to keep the data current. Historical data are high-volume and more stable, the harvesting process generally runs once, and a chunk-based harvest is preferred to reduce accessing time and data provider overhead. To harvest newly published data, data size is not the major problem but the scheduler must be able to harvest new data as soon as possible and guarantee completeness – even if data providers provide incomplete data for the current date. The OAI protocol provides flexibility in choosing the harvesting strategy; theoretically, one data provider can be harvested in one simple transaction, or one is harvested as many times as the number of records in its collection. But in reality only a subset of this range is possible; choosing an appropriate harvesting method has not yet been made into a formal process. We defined four harvesting types in *Arc*: (a) bulk-harvest of historical data (b) bulk-harvest of new data (c) one-by-one-harvest of historical data (d) one-by-one-harvest of new data. From our tests, these four strategies in combination can fulfill various requirements of a particular collection.

2.2 Database Schema

OAI uses Dublin Core (DC) as the default metadata set. All DC attributes are saved in the database as separate fields. The archive name and partition are also treated as separate fields in the database for supporting search and browse functionality. In order to improve system efficiency, most fields are indexed using full-text properties of the database, which makes high performance queries over large dataset possible. The search engine communicates with the database using JDBC and Connection Pool.

2.3 Search Interface specification

The search interface supports both simple and advanced search as well as result sorting by date stamp, relevance ranking and archive. Simple search allows users to search free text across archives. Advanced search allows user to search in specific metadata fields. Users can also search/browse specific archives and/or archive partitions in case they are familiar with specific data provider. Author, title, abstract search are based on user input, the input can use boolean operators (AND, OR, NOT). Archive, set, type, language and subject use controlled vocabularies and are accumulated from the participating archives' source data.

3. Results

We maintain two test sites, the public *Arc* cross archive service and another site for the OAI alpha test only. So far, in the public service, we have seven archives available for search (Table 1). In the OAI alpha test site, we have nine archives across different domains, including documents from Heinonline.org, Open Video Project, Language Resource Association, Library of Congress and other organizations. The alpha test data is not open to the public.

Table 1. Collections Harvested by *Arc* (by Feb 7, 2001)

Archive Name	URL	Records
arXiv.org e-Print Archive	arxiv.org	151650
Cognitive Science Preprints	cogprints.soton.ac.uk	999
NACA	naca.larc.nasa.gov	6352
Networked Digital Library of Theses and Dissertations	www.ndltd.org	2401
Web Characterization Repository	repository.cs.vt.edu	131
NCSTRL in Cornell	www.ncstrl.org	2080
NASA Langley Technical Report Server	techreports.larc.nasa.gov/ltrs	2323
Nine Other Collections For the OAI alpha test	arc.cs.odu.edu/help/archives.htm	9467

4. Lessons learned

Little is known about the long-term implications of a harvest-based digital library, but we have had the following initial experiences. The effort of maintaining a quality federation service is highly dependant on the quality of the data providers. Some are meticulous in maintaining exacting metadata records that need no corrective actions. Other data providers have problems maintaining even a minimum set of metadata and the records harvested are useless. We have not yet fully addressed the issue of metadata normalization. Some normalization was necessary to achieve a minimum presentation of query results. However, we did so on an ad hoc basis with no formal definition of the relationship mappings. A controlled vocabulary will be of great help for a cross archive search service to define such metadata fields as 'subject' or even 'organization'. XML syntax errors and character-encoding problems were surprisingly common and could invalidate entire large data sets. We also faced a trade-off in frequency of harvests: too many harvests could over burden both the service and data providers, and too few harvests allow the data in the service provider to potentially become stale.

5. REFERENCES

- [1] Van de Sompel, H. and Lagoze, C. The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine*, 6(2), February 2000. <http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html>
- [2] Van de Sompel, H., Krichel, T., Nelson, M. L., Hochstenbach, P., Lyapunov, V. M., Maly, K., Zubair, M., Kholief, M., Liu, X. and O'Connell, H. The UPS Prototype: An Experimental End-User Service across E-Print Archives, *D-Lib Magazine* 6(2), February 2000. <http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html>
- [3] Maly, K., Zubair, M., Anan, H., Tan, D. and Zhang, Y. Scalable Digital Libraries based on NCSTRL/Dienst. In *Proceedings of the 4th European Conference on Digital Libraries – ECDL 2000* (Lisbon, Portugal, September 2000), pp. 169–179.