2002

Federated Searching Interface Techniques for Heterogeneous OAI Repositories

Xiaoming Liu
Old Dominion University

Kurt Maly
Old Dominion University

Mohammad Zubair
Old Dominion University

Qiaoling Hong
Old Dominion University

Michael L. Nelson
Old Dominion University

See next page for additional authors

Repository Citation

Liu, Xiaoming; Maly, Kurt; Zubair, Mohammad; Hong, Qiaoling; Nelson, Michael L.; Knudson, Frances; and Holtkamp, Irma, "Federated Searching Interface Techniques for Heterogeneous OAI Repositories" (2002). *Computer Science Faculty Publications*. 35.
https://digitalcommons.odu.edu/computerscience_fac_pubs/35

Original Publication Citation

Liu, X., Maly, K., Zubair, M., Hong, Q., Nelson, M.L., Knudson, F., & Holtkamp, I. (2002). Federated searching interface techniques for heterogeneous OAI repositories. *Journal of Digital Information, 2*(4), 1-8.

Authors
Xiaoming Liu, Kurt Maly, Mohammad Zubair, Qiaoling Hong, Michael L. Nelson, Frances Knudson, and Irma Holtkamp

Journal of Digital Information, Vol 2, No 4 (2002)

Federated Searching Interface Techniques for Heterogeneous OAI Repositories

Xiaoming Liu, Kurt Maly, Mohammad Zubair, Qiaoling Hong, Michael L. Nelson*, Frances Knudson and Irma Holtkamp****
Old Dominion University, Norfolk, Virginia USA
Email: {liu_x, maly, zubair, hong_q}@cs.odu.edu
*NASA Langley Research Center, Hampton, Virginia USA
Email: m.l.nelson@larc.nasa.gov
**Los Alamos National Laboratory, Los Alamos, New Mexico USA
Email: {fknudson, isholtkamp}@lanl.gov

Abstract

Federating repositories by harvesting heterogeneous collections with varying degrees of metadata richness poses a number of challenging issues: (1) how to address the lack of uniform control for various metadata fields in terms of building a rich unified search interface, and (2) how easily new collections and freshly harvested data in existing repositories can be incorporated into the federation supporting a unified interface? This paper focuses on the approaches taken to address these issues in Arc, an Open Archives Initiative-compliant federated digital library. At present Arc contains over 1M metadata records from 75 data providers from various subject domains. Analysis of these heterogeneous collections indicates that controlled vocabularies and values are widely used in most repositories. Usage is extremely variable, however. In Arc we solve the problem by implementing an advanced searching interface that allows users to search and select in specific fields with data we construct from the harvested metadata, and also by an interactive search for the subject field. As the metadata records are incrementally harvested we address how to build these services over frequently-added new collections and harvested data. The initial result is promising, showing the benefits of immediate feedback to the user in enhancing the search experience as well as in increasing the precision of the user's search.

1 Introduction

The lack of interoperability is one of the significant problems facing digital libraries. One major objective of digital library interoperability is to provide a unified search interface. For the purpose of this paper, a unified search interface is defined as an interface that can seamlessly search across multiple repositories. Many repositories have significant investment in controlled metadata fields. This includes controlled vocabularies (thesauri, subject heading lists, etc.), controlled values (a type of encoded schema, usually a string formatted in accordance with a formal notation or parsing rules), and other locally controlled value or text. This paper discusses several metadata fields used in Dublin Core (DC) (Weibel *et al.* 1998) while emphasizing controlled vocabularies and values. Controlled metadata is crucial for effective search and retrieval of Internet resources (Desai 1997). French *et al.* (2001) point out that controlled metadata is of little use if it is not used effectively in query formulations. Our focus is how to build a rich, unified search interface that can exploit the controlled metadata across heterogeneous collections. The problem is solved by an advanced searching interface that allows users to search and select in specific fields with data we construct from the harvested metadata, and also by an interactive search for the subject field.

Independent of the metadata control issue, there are two ways to implement a coherent set of digital services across heterogeneous digital repositories: a distributed searching approach (Levy *et al.*1996; Gravano *et al.* 1997; Dushay *et al.* 1999) and a harvesting approach (Bowman *et al.* 1995; Lagoze and Van de Sompel 2001; Schwartz 1996; Koster). In the harvesting approach, the harvesting step is well understood by the Web crawler community and is easy to implement. The data are usually harvested on the service provider side, so we have the luxury of pre-building advanced services without relying on real-time interactive access to the remote archives. However, building a rich unified search interface over harvested metadata brings about new challenges. Many information-rich repositories have major investments in detailed metadata, which frequently includes some forms of controlled vocabularies and/or controlled values. To build better services, we need to understand how metadata control is used in these repositories, and determine if we can exploit them in a unified interface. Furthermore, we need to know how easily new collections and freshly harvested metadata can be built into the unified interface.

One straightforward approach is to build a keyword search similar to typical Web search engines. Web search engines represent a well-proven, successful technology based on harvesting and keyword searching. Keyword searching is a useful way to assume little about the semantics of a document, which works well for the heterogeneous, unstructured data sources that make up the Web, but when structured metadata is available, it fails to exploit the additional semantics.

Another approach to address the lack of a unified controlled metadata is to create a standard and map each repository's controlled metadata to the standard (Koch *et al.* 2001). For controlled vocabularies, this approach can be improved by a meta-thesaurus based solution like UMLS (Unified Medical Language System) (Lindberg *et al.* 1993), which "preserves the meanings, hierarchical connections, and other relationships between terms present in its source vocabularies, while adding certain basic information about each of its concepts and establishing new relationships between concepts and terms from different source vocabularies". Both approaches introduce significant human effort to maintain the relationships. Adding new collections to the federation leads to the complexity of updating relationships. Therefore neither is feasible in our scenario.

We describe how we solve these problems in Arc (Liu *et al.* 2001), a fully implemented system that provides uniform access to over 1M metadata records from 75 Open Archives Initiative (OAI)-compliant data providers (Lagoze and Van de Sompel 2001). OAI is a major effort to promote interoperability through the concept of metadata harvesting. The OAI framework supports data providers (repositories, archives) and service providers. Service providers develop value-added services based on the information collected from data providers. Data providers are simply collections of harvestable metadata that may or may not contain additional services and content. OAI is designed to make it as easy as possible for digital repositories to become data providers. As a consequence, service providers have to work much harder than data providers, and the quality of the service providers is directly proportional to the lengths they go to address issues that lie outside the scope of the OAI protocol. OAI is becoming widely accepted and there are many archives currently or soon-to-be OAI compliant.

Arc collects data from several different communities, ranging from museum to eprint collections. As participating archives add new records to their collections, their metadata records are harvested by Arc. Analysis of these heterogeneous collections indicates that metadata control is widely used in most repositories, especially in certain metadata fields. Usage is extremely variable, however. From our study it is clear that no single approach would allow effective use of the manifold metadata control we encountered. In Arc we solve the problem by implementing an advanced searching interface that allows users to search and select in specific fields with data we construct from the harvested metadata, and also by an interactive search for the subject field. As the metadata records are incrementally harvested we address how to build these services over frequently-added new collections and harvested data.

The rest of the paper is organized as follows. Section 2 analyzes the metadata variability in Arc collections. In Section 3, we discuss the new interactive search approach. Section 4 discusses related works, and Section 5 analyzes the initial experience and discusses future work.

2 Metadata Variability

A metadata field can be based on either controlled or free text. We consider three types of metadata control: controlled vocabularies, controlled values, and other locally defined metadata. Controlled vocabularies are typically used for subject access and can control synonyms, variant spellings, as well as providing broad term, narrow term and other subject relationships. Controlled vocabularies include thesauri and classification schema. Controlled values are usually a string formatted in accordance with a formal notation or parsing rules (e.g. "2000-01-01" as the standard expression of a date). These include values of a "fixed or set length" (e.g. "eng" or "en" as the standard expression of ISO 639-2 three-character vs ISO 639-1 two-character language code for "English"). Locally defined metadata may contain a mix of locally controlled text strings or values for a given metadata field/element used in records from a given repository and an "encoded schema" may be implied but is not clearly identified and available for public use.

OAI uses unqualified DC as the default metadata set to enable minimal interoperability. Although OAI supports other metadata formats, our discussions are based on DC because it is the common metadata set supported by all OAI compliant repositories. Over the past several years DC has developed as a *de facto* standard for simple cross-discipline metadata. It defines 15 metadata elements: creator, title, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights. Among the 15 DC fields, some, such as description, are obviously free-text based. The subject field tends to be based on controlled vocabularies, and other fields, such as type, date, format and language may be based on either controlled values or locally defined values. In contrast to qualified DC, unqualified DC does not include an encoding scheme to aid in the interpretation of an element, so there is no definite way to decide whether a metadata field is controlled without consulting the original data providers. However, by studying the harvested metadata, in most cases the difference between controlled and free text input is obvious so we consider the tendency shall be correct. We manually examined the subject, language, format, date and type fields for further study of metadata variability. The meaning of "subject", "date" and "language" are fairly straightforward; the definition of "type" is "The nature or genre of the content of the resource", and of "format" is "The physical or digital manifestation of the resource".

We harvested from 75 data providers registered at the OAI homepage and from other resources. To understand how metadata fields are used, consider Table 1, which has been constructed based on the collections in Arc.

Table 1. Sample data of subject, format, language and date fields in four archives (excerpt)

	CogPrints	MIT etheses	arXiv	NCSTRL
Subject	Complexity-Theory Computational-Linguistics Computational-Neuroscience	Computer-animation Computer-architecture Computer-composition	Computational Physics Computer Vision and Pattern recognition Computers and Society	N/A
Format	N/A	application/pdf image/gif	N/A	N/A
Language	N/A	N/A	FRENCH French German	English French German
Date	1950-01-01 1953-01-01 1954-01-01	1900-01-01 1903-01-01 1921-01-01	1991-01-15 1991-02-05 1991-02-25	1958-12-01 1959-03-01 1959-12-01
Type	Book Chapter Conference Paper Conference Poster	Thesis	e-print	Proceeding Proceedings

Table 1 contains an excerpt from our analysis of how metadata is used in four archives. The complete data are available at (http://arc.cs.odu.edu/stat). Table 2 lists the number of records harvested and the number of distinct subject, type, format and language fields used in each archive. Table 2 also lists whether consistent formatting is used in the date field. In columns 4-7, a 'zero' value means this metadata field is not used, either because the metadata is not available in the repository, or the repository simply ignores it because it is constant across all records (e.g. English language archives may simply leave the DC language field empty). In column 3, dealing with date field formatting, the value "Y" means consistent formatting is used; "N" means free input is used; "N/A" means this field is never used. In Table 2, if the number of records is significantly larger than the number of distinct values in one metadata field, it suggests that a controlled metadata is used. This is not always true, however, so we manually verified these results.

Table 2. Metadata variability in Arc (to April 3, 2002). Detailed information about each archive is available at http://arc.cs.odu.edu:8080/oai/info.jsp

Archive	Number of Records	Consistent Format in the "Date" Field	Number of Unique Values in "Subject" Field	Number of Unique Values in the "Type" Field	Number of Unique Values in the "Format" Field	Number of Unique Values in the "Language" Field
8657690236	798	Y	53	1	0	0
AIM25	3962	N	2424	1	1	0
anlc	5	Y	2	1	2	0
anu	114	Y	22	6	0	0
aps	422	N/A	5	0	180	0
arXiv	182996	Y	121	1	0	12
bmc	220	Y	0	13	0	1
caltechCSTR	504	Y	8	2	0	0
caltecheerl	140	N	1	1	0	0
caltechETD	30	Y	10	1	4	1
cav2001	111	Y	103	1	0	0
CBOLD	89	Y	136	1	20	3
CCSDthesis	99	Y	16	1	0	0
CDLCIAS	36	Y	9	4	0	0
CDLDERM	5	Y	3	2	0	0
cdlib1	224	Y	324	1	1	0
CDLTC	1	Y	1	1	0	0
CEIAT	19	Y	5	3	2	1
celebration	186	N	160	1	0	0
cimi	196920	N	41934	872	14088	7
cogprints	1396	Y	70	10	0	0
conoze	77	Y	0	1	0	1
CPS	346	Y	49	1	0	0
CSTC	64	Y	46	3	0	1
dfki	71	Y	0	57	0	9
dlpscoll	44749	N	7331	1	0	1
DUETT	223	Y	34	1	9	1
EKUTuebingen	485	N	3572	4	123	2
eldorado	312	Y	99	1	4	6
elra	184	Y	0	8	0	43
ENUMERATE	82	N	0	1	1	0
epubwu	11	Y	11	1	1	2
etdcat	62097	Y	75965	18	22710	46
ethnologue	486	Y	484	0	0	0
Formations	21	Y	16	2	0	0
GenericEPrints	1	Y	2	1	1	0
glasgow	15	Y	11	8	0	0
HKUTO	8365	Y	88	1	1	4
hsss	11	N	11	2	1	2
HUBerlin	118	N	7	1	1	2
in2p3	2580	Y	78	5	0	3
ioffe	337	Y	0	0	0	1
lacito	65	Y	0	4	0	10
lcoa1	97159	N	4344	143	0	29
LDC	212	Y	2	4	2	26
LSUETD	71	Y	30	1	4	1
LTRS	2629	N	70	117	0	0
mathpreprints	76	N/A	0	0	0	0
mit.etheses	6288	Y	1339	1	5	0
MONARCH	490	Y	471	11	0	6
NACA	7492	N	0	7483	0	0
NCSTRL	21213	N	0	60	0	7
ndltd	6	Y	4	2	0	0
Nottingham	41	Y	9	5	0	0
NSDL-DEV-CU	2559	N	1735	8	857	9
OpenVideo	1658	Y	389	1	3	2
ota	1245	Y	0	1	44	52
perseus	1394	N/A	0	1	0	0

physdoc	407	Y	397	1	8	13
rdn	387	N/A	674	0	0	24
RIACS	35	Y	5	1	0	0
sammelpunkt	109	Y	29	11	116	0
sceti	47	N	130	0	0	1
scout	50	N	175	0	0	0
SUUB	125	N	97	1	0	2
tkn	321	N	401	25	2	0
Tropicos	517400	Y	0	0	0	0
UBC	2	Y	3	3	1	2
UDLAthesis	95	Y	59	2	1	3
uiLib	29443	N	3353	5	688	4
UKETD	26	Y	9	1	1	0
UKOLN-ejournals	113	N/A	0	0	0	1
USF	28	Y	13	1	2	1
UUdiva	1536	Y	7387	1	2	6
VTETD	3138	Y	170	1	1	1
yea	86	Y	279	2	32	9

Figure 1 shows the percentage of archives that use controlled values in each metadata field. Figure 1 indicates that metadata control is widely used among archives, especially in the type, format, language and date fields. About half use controlled vocabularies in the subject field. In the archives labeled "without metadata in specific field", many use a constant value (e.g. "English" in the LANGUAGE field based on other factors such as an assumption that records in "English-based" repositories represent only publications in the English language.)

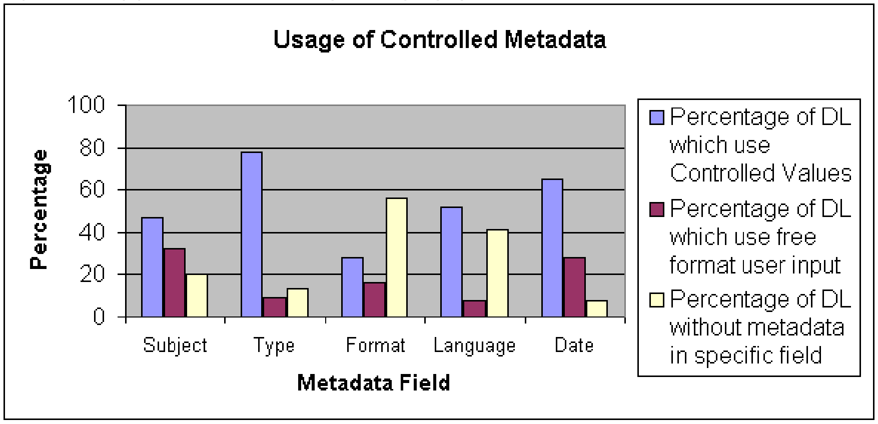


Figure 1. Controlled Metadata

In many circumstances, even if controlled metadata are used, each archive may employ its own semantics for these fields. Archives may have different semantics for the same field, and they frequently use different standards, such as subject classification methods. However, data providers have invested significant human and machine resources to use controlled metadata, and service providers should try to re-use these rich metadata.

One straightforward approach is to use a standard and map the controlled metadata from an individual archive to the standard. We could then reflect this in the search interface by showing the standard as a selectable option. This approach has three major limitations:

1. The standard may differ in terms of levels and semantics from that of an individual archive leading to low-precision searches.
2. Adding a new collection to the federation leads to complexity in updating mapping tables.
3. Significant manual effort may be required to define the standard and the mapping tables. Moreover, any new archive may differ significantly from the standard necessitating an update to the unified scheme.

From Table 1 we also observed that while the number of subject fields is large, the number of different language, type and format fields is limited in most archives. This leads directly to our design decision to create a browse interface for language, type and format fields. An interactive search interface is designed for subject search. Because most date fields follow strict controlled values, we implement the date field as a free input with restricted format.

3 Approaches

Our solution is based on the user-centric approach where users engage in a series of interactions with the federation service to communicate their queries. There are two phases of interactions. In the first stage, a user searches the controlled value, and at the second stage continues resource discovery based on the results from the first stage. We built browse capacity and interactive search interfaces based on the user-centric approach.

In the metadata harvesting approach, there are no pre-defined authority files available and the unified interface has to be built over harvested data that are added on a regular basis, so the search interface has to be adaptive to the frequently changing metadata. Figure 2 shows the components of the system and how it works. The harvester keeps harvesting data from sources.

Another process periodically collects key metadata fields from harvested metadata, builds an index and refreshes the search interface. Users interact with the search interface to identify their preferred controlled metadata and then execute a search.

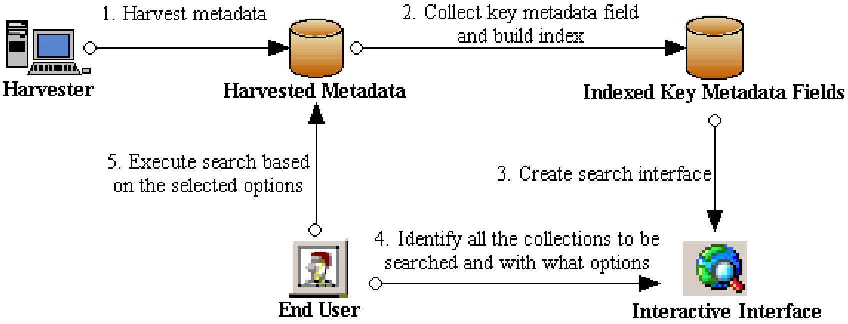

1. Harvest metadata
2. Collect key metadata field and build index
Harvester
Harvested Metadata
Indexed Key Metadata Fields
5. Execute search based on the selected options
3. Create search interface
4. Identify all the collections to be searched and with what options
End User
Interactive Interface

Figure 2. Building a Search Interface Based on Harvested Metadata

In the metadata harvesting approach, synchronization between data provider and service provider is very important. In our case, since the search interface is built over harvested data, it must be adaptive to the frequently changed data. We implemented an interface builder for this objective. The interface builder is responsible for creating a new interface when new archives and records are added. The interface builder is resource-expensive and cannot run just-in-time. Instead it periodically builds a cached interface on the server side, and the user will always see this cached version. This interface builder does not change the layout of the interface nor the type of fields the user can choose to interact with; rather it creates values the user can choose from when selecting a field for queries.

3.1 Keyword Search

Keyword search allows users to search all metadata fields across archives. It is implemented by accumulating and indexing all metadata fields together. Keyword search provides a simple and familiar way to conduct search across all archives, and the input can include Boolean operators (AND, OR, NOT). It is probably the only way to search across extremely variable sources without major work, but it cannot exploit the rich metadata set defined by source archives.

3.2. Advanced Search

Advanced search provides a way for a user interactively to pick up the controlled values defined by specific archives via the search interface. The searcher picks an interesting archive, then the system creates a series of selectable options for each metadata field, and the user selects the exact controlled value and executes the precise search.

In the implementation, the metadata fields are accumulated from the archives' source data. One background process periodically checks the harvested data and recreates the browse list. The advanced search capacity fits the user who is familiar with specific archives, but it does not scale well for a large number of archives because the browse list becomes too large to use.

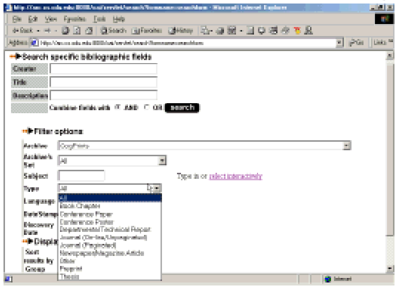


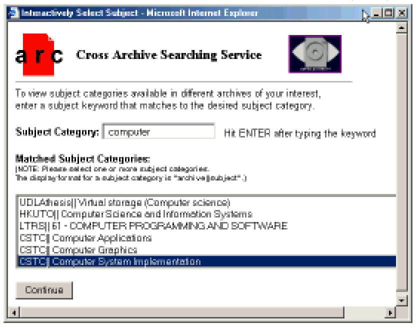
Figure 3. Advanced Search Interface (full size animated GIF image)

3.3 Interactive Approach

In this approach, the users provide some simple initial descriptions of their queries by means of a series of keywords. The system will then present the user with contextual metadata information from those archives that have relevant records. Users can then opt to add to the search query with richer metadata elements chosen from those presented by the system. The key to this approach lies in the interaction between the user and the system. This interaction provides increasing detail to the user by obtaining detailed values from the harvested metadata. Consider the example of the subject classification maintained by arXiv for physics and a subject classification maintained by the Human Development collection. Assume that the user types "accelerator", the system would find this term in the arXiv classification under high-energy-physics: proton accelerator, and in the Human development collection under: university education: science: accelerated learning. Users can then choose which more closely fits their view of the subject.

3.3.1 Prototype Implementation

Based on the user-centric approach, we have implemented an interactive interface for Arc to help users select the subject category. The interface is illustrated in Figure 4 (note that Figure 4 only shows the subject selection interface; the rest of the interface is similar to Figure 3). To view subject categories available in different archives, a user enters a subject keyword that closely matches the desired subject category. The input can include Boolean operators (AND, OR, NOT). Next, the user is shown matched subject categories from different archives. The user either selects one or more of the matched categories or further refines the matching list by typing more words in the field. This way a user is able to select the desired subject categories, which are then used to construct the search query for the Arc database.



Interactively Select Subject - Microsoft Internet Explorer

arc Cross Archive Searching Service

To view subject categories available in different archives of your interest,
enter a subject keyword that matches to the desired subject category.

Subject Category: computer Hit ENTER after typing the keyword

Matched Subject Categories:
(NOTE: Please select one or more subject categories.
The display format for a subject category is "archive||subject".)

UDLAthesis||Virtual storage (Computer science)
HKUTO|| Computer Science and Information Systems
LTRS|| 61 - COMPUTER PROGRAMMING AND SOFTWARE
CSTC|| Computer Applications
CSTC|| Computer Graphics
CSTC|| Computer System Implementation

Figure 4. Interactive Subject Selection Interface (full size animated GIF image)

To support the interactive subject selection interface, we created a subject table in the database. The subject table is constructed by extracting the subject field(s) from each archive. The table consists of two fields: archive and subject. Once the user enters a keyword and hits enter in the subject selection interface, the keyword is sent to a servlet at the back-end. The servlet then connects to the database using JDBC (Reese 2000) and searches the subject table for the keyword using an Oracle full-text search. The matched records are returned to the user and displayed in a multi-selection list.

The interactive subject selection interface improves the search precision by giving users the flexibility of selecting the archive and subject category of his choice.

3.3.2 Use-case Scenario

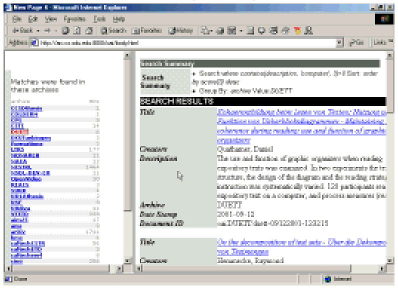
We show the effectiveness of our approach by considering a few use-case scenarios, which demonstrate that the interactive subject selection interface improves the search precision by giving the user the flexibility to select the archive and subject category of choice. In Table 3, once the user searches for the subject keyword "science", the interactive subject selection interface returns 613 matched subject categories in 39 archives whose subjects include the word "science". If the user refines the query to "computer science", the search interface matches 60 subject categories in 19 archives. Next the user selects the archives and subjects of choice.

Table 3. Number of Matched Archives and Subjects Using Interactive Search

Keyword typed by user	# of matched archives	# of matched subjects
science	39	613
computer science	19	60
computer science or computer engineering	20	68
computer network	4	19
physics	27	215
nuclear physics	6	32

3.4 Displaying the Search Result

Our experience proves that rich metadata sets not only provide a way to give users a powerful search interface, but also help users to review the search results. Users have the flexibility of sorting and grouping by rank, date stamp, subject or archive. In the result display page, the left frame shows all groups and hit numbers, and the right frame shows summary information about each document in the selected group (Figure 5). Users can also traverse different pages if multiple search pages exist. When users are interested in a document, they can view the detail page, and follow the link to the full text document that resides in the data provider's repository.



Matches were found in these archives

Search Summary

SEARCH RESULTS

Title
Creator
Description
Archive DUETT
Date Stamp 2001-08-12
Document ID

Title
Creator

Figure 5. Arc Search Result Page (full size animated GIF image)

4 Related Work

Although many projects (Levy *et al.* 1996; Gravano *et al.* 1997; Zubair *et al.* 2000; Maa *et al.* 1997; Mahoney and Giacomo 2001; Dushay *et al.* 1999) have tried to provide uniform access to heterogeneous collections, almost all such systems use a distributed searching approach. Their major task is how to select different sources, issue queries to heterogeneous sources, and merge the query results in limited time, usually one user session. The systems differ from the metadata harvesting approach used in Arc. The method introduced in Goldman and Widom (1998) is an interactive system for semi-structured data that helps the inexperienced user by focusing on a semi-structured graph-based database for Web data. Entry Vocabulary (Gey *et al.* 2001) is another technology that enhances searching by mapping from the user's ordinary language to the metadata of the digital sources. French *et al.* (2001) demonstrate a technique for mapping user queries into a controlled indexing vocabulary, with the potential to radically improve document retrieval performance. Both assume the existence of one unique classification scheme, which does not exist in our scenario. In an earlier study of Arc (Liu *et al.* 2001), the issue related to lack of controlled values was addressed in an *ad hoc* way. At that time the number of collections was rather limited, and we found the method presented did not scale well with the addition of OAI-compliant archives.

5 Conclusions and Future Work

We built the Arc search interface based on the above approach and the initial results are promising. Working with over 1M records in Arc, the advanced search interface html page is only around 74K. This approach also allows daily updating of this html page, which covers all vocabularies from 75 data providers. This interface can be accessed quickly with the speed of a conventional home Internet connection. For the interactive search, the user has the flexibility of continually refining queries so the system will scale to a larger number of data providers. After removing test queries from our own site, 8053 queries were conducted in five months: among them 6137, or 76%, were keyword searches; another 1916, or 24%, were advanced searches, indicating that users still prefer to use keyword search. In the NCSTRL (http://www.ncstrl.org) project, which is based on Arc, a usability evaluation (Shivakumar *et al*. 2002) indicated the interface is easy to understand and not difficult to use. The system functionality seems appropriate and the user interface is aesthetically pleasing. The study also addressed some potential usability problems that could aid future redesign and development. A focus group study at Los Alamos National Laboratory indicated the interactive interface holds promise. The benefits of immediate feedback to the user hold great promise in enhancing the search experience as well as increasing the precision of the user's search. Making this interface more intuitive will be part of our future study.

It is clear from our experiment that most archives tend to use controlled metadata, but the metadata are extremely variable from archive to archive. We have implemented two user-centric search interfaces, advanced searching and interactive searching, providing a unified search interface across heterogeneous collections and exploiting the rich controlled metadata.

Our approach still makes individual archives visible to the searcher, but from our study it is becoming evident that one unified interface, which exploits the rich source metadata and is transparent to participating archives, is feasible. Controlled values are widely used in many archives and for fields such as type and language we could map the data to a standard without significant manual effort, with the help of approximate word matching and other algorithms (French *et al.* 1997). In the future we shall compare our current solution with regard to the subject field, which is important in cross-archive searching, to one where we map to a unified schema by automatic categorization algorithms. We shall also study the possibility of improving the interactive search by using reverse-engineered text categorization (Gey *et al.* 2001) that is used to supply mappings from an ordinary language vocabulary to a specialist vocabulary.

References

Bowman, C. M., Danzig, P. B., Hardy, D. R., Manber U. and Schwartz, M. F.

(1995) "The Harvest Information Discovery and Access System". *Computer Networks and ISDN Systems*, 28 , 119-125

Desai, B.C. (1997) "Supporting discovery in virtual libraries". *Journal of the American Society for Information Science*, 48, 190-204

Dushay, N., French, J. C. and Lagoze, C. (1999) "A Characterization Study of NCSTRL Distributed Searching". *Cornell University Computer Science Technical Report*, TR99-1725 http://historical.ncstrl.org/tr/ps/cornellcs/TR99-1725.ps

French, J. C., Powell, A. L., Gey, F. and Perelman, N. (2001) "Exploiting A Controlled Vocabulary to Improve Collection Selection and Retrieval Effectiveness". In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, Atlanta, Georgia, USA, pp.199-206

French, J. C., Powell, A. L., Schulman, E. and Pfaltz, J. L. (1997) "Automating the construction of authority files in digital libraries: a case study". In *Proceedings of Research and Advanced Technology for Digital Libraries, first European conference*, Pisa, Italy, pp. 55-71

Gey, F. C., Buckland, M., Chen, A. and Larson, R. (2001) "Entry Vocabulary -- a Technology to Enhance Digital Search". In *Proceedings of the First International Conference on Human Language Technology*, San Diego, USA

Goldman, R. and Widom, J. (1998) "Interactive query and search in semistructured databases". In *WebDB'98, Proceedings of the International Workshop on the Web and Databases*, pp. 42-48

Gravano, L., Chang,K., Garcia-Molina, H., Lagoze,C. and Paepcke, A. (1997) "STARTS: Stanford proposal for internet meta-searching". In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 207-218

Koch, T., Neuroth, H. and Day, M. (2001) "Renardus: Cross-browsing European subject gateways via a common classification system (DDC)". *IFLA satellite meeting: Subject Retrieval in a Networked Environment*, OCLC, Dublin, OH, USA http://www.lub.lu.se/~traugott/drafts/preifla-final.html

Koster, M. "The Web Robots Page" http://www.robotstxt.org/wc/robots.html

Lagoze, C. and Van de Sompel, H. (2001) "The Open Archives Initiative: Building a low-barrier interoperability framework". In *Proceedings of the First ACM/IEEE Joint Conference on Digital Libraries*, Roanoke, VA, pp. 54-62 http://www.cs.cornell.edu/lagoze/papers/oai-final.pdf

Lee-Smeltzer, K. H. (2000) "Finding the needle: controlled vocabularies, resource discovery, and Dublin Core". *Library Collections, Acquisitions, & Technical Services*, 24, 205-215

Levy, A. Y., Rajaraman, A. and Ordille,J. (1996) "Querying heterogeneous information sources using source descriptions". In *Proceedings of the 22nd International Conference on Very Large Data Bases*, pp. 251-262

Lindberg, D., Humphreys, B. and McCray, A. (1993) "Unified medical language systems". *Methods of Information in Medicine*, 32(4), 281-291

Liu, X., Maly, K., Zubair, M. and Nelson, M. L. (2001) "Arc - An OAI Service Provider for Digital Library Federation". *D-Lib Magazine*, 7(4) http://www.dlib.org/dlib/april01/liu/04liu.html

Maa, M.-H., Esler, S. L., and Nelson, M. L. (1997) "Lyceum: A Multi-Protocol Digital Library Gateway". *NASA* TM-112871 http://techreports.larc.nasa.gov/ltrs/PDF/1997/tm/NASA-97-tm112871.pdf

Mahoney, D. and Giacomo, M. D. (2001) "Flashpoint @ LANL.gov: A simple smart search interface". *Issues in Science and Technology Librarianship*, Summer, 2001 http://www.istl.org/istl/01-summer/article2.html

Reese, G. (2000) *Database Programming with JDBC and Java* (Sebastopol, CA: O'Reilly & Associates)

Shivakumar, P. (2002) "Sample NCSTRL Usability Evaluation Report". *Virginia Tech Computer Science Technical Report*, TR02-08

Schwartz, M. (1996) "Report of W3C Distributed Indexing and Searching Workshop" http://web3.w3.org/Search/9605-Indexing-Workshop

Weibel, S., Kunze, J., Lagoze, C. and Wolfe, M. (1998) "Dublin Core metadata for resource discovery". Internet RFC-2413 http://www.ietf.org/rfc/rfc2413.txt

Zubair, M., Maly, K., Ameerally, I. and Nelson, M. L. (2000) "Dynamic Construction of Federated Digital Libraries". In *Proceedings of WWW9 Conference*, Amsterdam, The Netherlands http://www9.org/final-posters/poster17.html