National Conference on Scientific Achievements of SC & ST Scientists & Technologists
14–16 April 2009, National Aerospace Laboratories, Bangalore-17

A Comparative Study of the Search and Retrieval Features of OAI Harvesting Services

V. Indrani[1] and K. Thulasi[2]

[1]Information Centre for Aerospace Science and Technology, National Aerospace Laboratories,
Bangalore-560017, INDIA
E-mail: indrani@css.nal.res.in

[2]National Centre for Science Information, Indian Institute of Science, Bangalore-560012, INDIA

ABSTRACT

Several OAI service providers are coming up providing cross-search services by harvesting metadata from OAI compliant repositories. OAI facilitates quick discovery of content and free exchange of information among repositories through Service Providers. In order to achieve interoperability in their operations, Service Providers need to incorporate a generalized set of search and browse features in their search interface. Few parameters are drawn to compare the search and retrieval features of Service Providers and arrived at a useful checklist for Service Providers to achieve homogeneity and standardization while designing their search interface.

Keywords: *Service Providers Features, Comparative Study of OAI harvesting Services, Metadata Fields in OAI Harvesting.*

1. INTRODUCTION

Objective of this article is to compare the features supported by the service providers for searching browsing and presentation of results. Open Archive Initiative (OAI) is a tool. This is all about moving metadata around and main to focus on interoperability. OAI has a protocol to harvest metadata from other archives. OAI divides the world into two participants one is metadata providers called as data providers and another one has harvester called as service providers. Data providers refer to entities that possess data and metadata and are willing to share metadata with others via well-defined OAI protocols. Service Providers are entities that harvest metadata from Data providers in order to provide higher-level service to users. Search and Retrieval features are used by the Archives (Data providers and Service providers) either to retrieve or to expose their metadata. This article compares the search and retrieval features of the OAI service providers.

2. AIM AND SCOPE

OAI stresses on interoperability in technology and its operations as well and also it is equally important for Service Providers to incorporate a generalized set of search and browse features in their search interface, to make it more interoperable among service providers, repositories and users. List of criteria is identified and classified under broad headings a. Purpose and Scope b. Software c. Volume and Growth d. Usage e. Metadata f. Search and Browse g. Display Options and h. Additional Services. Based on the list of criteria, a comparative study of the Search/Browse Interface of a few important Service Providers was carried out, in relation to searching, browsing and presentation of results. Few service providers like ARC, OAIster, SAIL, Archon, Metalis and cite base among other registered service providers that exist today are compared for the study. Arc, OAIster and SAIL are leading service providers harvesting metadata from several archives covering major subject disciplines. Cite base include citation information besides standard metadata elements to provide citation search services. ARCHON and Metalis are two subject specific Service Providers in Physics and Library science respectively.

Table 1: Showing Comparison of the Features of Different Service Providers Based on the Criteria

Search and Retrieval Features	*Arc*	*OAIster*	*SAIL*	*Cite base*	*Archon*	*Metalis*
1. Purpose						
1. 1.1 Cross Archive Search	yes	yes	yes	yes	yes	yes
1. 1.2 Citation based service	No			yes	yes	
2. Scope						
2. 2.1 Discipline based harvesting	No				yes	yes
2.2 Multiple discipline	yes	yes	yes	yes		
2.3 Resource type harvesting						
2.3.1 Technical Reports						
2.3.2 Patents						
2.3.3 Thesis						
2.3.4 Others						
2.3.5 All types	yes	yes	yes	yes	yes	yes
3. Software						
3.1. Own Software	yes		yes	yes	yes	yes
3.1.1 Available as open source	yes		yes	yes	yes	yes
3.2 Commercial Software		yes				
3.3 Database						
3.3.1 MySQL	yes	yes	yes	yes	yes	yes
3.3.2 Oracle	yes					
3.3.3 Others						
3.4 Platform						
3.4.1 Linux Operating System	yes	yes	yes	yes	yes	yes
3.4.2 Java	yes	yes	yes	yes	yes	yes
3.4.3 Perl	yes	yes		yes		yes
3.4.4 Others						
4. Volume and Growth						
4.1 Frequency of Harvesting						
4.1.1 Weekly		yes	yes	yes		yes
4.1.2 Bi weekly	yes				yes	yes
4.1.3 Others						
4.2 Records	71,56,192	55,32,970	6,42,530	2,00,000	3,81,270	
4.3 Archives Harvested	180	495	107	3	5	9
5. Service Usage Statistics						
5.1 No. of Searches, Records			yes			
5.2 Most accessed archives			yes			
5.3 Most accessed clients			yes			
6. Metadata						
6.1 Unqualified Dublin Core	yes	yes	yes	yes	yes	yes
6.2 Qualified Dublin Core						
6.3 Any Other metadata						
7. Search and Browse						
7.1 Simple Search	yes	yes	yes		yes	yes
7.2 Advanced Search	yes		yes	yes	yes	yes
7.2.1 Field based						
7.2.1.1 Author/Title	yes	yes	yes	yes	yes	yes
7.2.1.2 Title	yes	yes		yes	yes	yes
7.2.1.3 Abstract	yes	yes	yes	yes	yes	yes
7.2.1.4 Subject		yes				
7.2.1.5 Archive						
7.2.1.6 Date						
7.2.1.6.1 Deposit date/Date stamp						
7.2.1.6.2Discovery dt.						
7.2.2 Phrase searching	yes	yes			yes	yes
7.2.3 Boolean	yes	yes	yes	yes	yes	yes
7.2.4 Equation search					yes	
7.2.5 Process Result Set	yes	yes			yes	
7.2.6 Duplicate detection						

(Contd…)

(Table 1 contd…)

Search and Retrieval Features	*Arc*	*OAIster*	*SAIL*	*Cite base*	*Archon*	*Metalis*
7.3 Filter option						
7.3.1 Archive						
7.3.1.1 All archives	yes		yes		yes	yes
7.3.1.2 Archive name	yes		yes		yes	yes
7.3.2 Subject	yes		yes		yes	yes
7.3.3 Resource type	yes	yes			yes*	yes
7.3.4 Date Stamp	yes		yes		yes	yes
7.3.5 Discovery Date	yes		yes		yes	yes
7.4 Browse						
7.4.1 Archive	yes	yes			yes	
7.4.2 Title						
7.4.3 Author			yes			
7.4.4 Any other/Deposit date			yes			
7.5 Citation Search						
7.5.1 Citation Author				yes		
7.5.2 Paper				yes	yes	
7.5.3 Year				yes		
7.6 Search History						
7.6.1 Saved Searches			yes			
7.7 Annotations					yes	
8. Display Option						
8.1 Sorting:						
8.1.1 Title	yes	yes			yes	yes
8.1.2 Author		yes				yes
8.1.3 Date stamp		yes		yes		
8.1.4 Discovery date	yes	yes		yes	yes	yes
8.1.5 Archives	yes				yes	
8.1.6 Subject	yes				yes	
8.1.7 Relevance Ranking:		yes		yes		
8.1.7.1 Hit frequency		yes				
8.1.7.2 Weight Hit frequency		yes				
8.1.7.3 Citation, Hits, Score				yes		
8.2 Display Results						
8.2.1 Archives	yes	yes	yes		yes	
8.2.2 Summary	yes		yes	yes	yes	yes
8.2.2.1 Title	yes		yes	yes	yes	yes
8.2.3 Detail	yes	yes	yes	yes	yes	yes
8.2.3.1 Author	yes	yes	yes	yes	yes	yes
8.2.3.2 Title	yes	yes	yes	yes	yes	yes
8.2.3.3 Contributor		yes				
8.2.3.4 Year (Discovery)	yes	yes	yes	yes	yes	yes
8.2.3.5 Publisher	yes	yes	yes	yes		yes
8.2.3.6 Resource type	yes	yes	yes	yes		yes
8.2.3.7 Resource Format	yes	yes	yes			
8.2.3.8 Language		yes	yes			yes
8.2.3.9 Abstract	yes	yes	yes	yes	yes	yes
8.2.3.10 Subject	yes	yes	yes	yes	yes	yes
8.2.3.11 URL	yes	yes	yes	yes	yes	yes
8.2.3.12 Note		yes				
8.2.3.13 Record ID			yes	yes	yes	
8.2.3.14 Citation info				yes	yes	
8.2.3.15 Similar Authors (clickable)				yes	yes	
8.2.3.16 Similar Subjects (clickable)	yes				yes	
8.2.3.17 Institution		yes	yes			
9. Additional services						
9.1 Alerting services			yes			
9.2 Act as data provider	yes	yes				
9.2.1 Base URL	yes	yes				

The basic idea is to bring out a list of parameters for comparison of search and browse features with suggestion for Service Providers to include the same in order to:

(a) Achieve homogeneity and standardization while designing their search interface
(b) Help users to search and identity resources efficiently and effectively while searching from different Service Providers search interface.

Also a checklist that would be useful for Service Providers, while designing their search/browse interface, and would also facilitate quick access and efficient retrieval of records. This could be useful to current and prospective service providers in improving or designing their search interface who plan to set up new OAI-based Service Provider.

3. ANALYSIS OF SERVICE PROVIDERS BASED ON THE ABOVE COMPARISON OF CRITERIA

Presently, Archon and Cite base offer citation search services among others. Arc and OAIster act as both Data Providers and Service providers. These Service Providers harvest all resource types like journals, technical reports, and conference proceedings and do not concentrate on any specific ones. While OAIster and Cite base have single search interface, the rest support both simple and advance search interface. Cite base include citation searching and Archon Equation searching. Archive names in the dropdown menu should be arranged strictly in alphabetical order. Just as corresponding Archive Set values get displayed with particular Archive, so also corresponding subject should be displayed instead of subjects included in all Archives, thus reducing search time. Cite base includes query-processing time with response being pretty fast. None of them use Proximity operators like WITH, NEAR which increases precision in searching. It is useful to include Browsing by broad topical categories or Subjects, Resource Types besides Institutions or Deposit date or Author. Search within selected browse categories will be useful as provided by SAIL and Metalis. Currently there is a limit in the number of records for Grouping or Sorting. Cite base and Metalis have no grouping of archives for displaying records. Only SAIL supports Saving Search history for setting up Alerts, Saving records using standard bibliographic tools, viewing latest updates of records harvested and offers detailed Usage Statistics. Archon's Annotations field is unique enabling users to make some notes on the respective record. Arc, Archon, Cite base and OAIster support relevance ranking of results. Archon's Linking is extensive compared to others. It includes, author with links to his other articles, Show Equations(all equations from the result set is shown), Similar subject (all other articles in the current result set with same subject), Citation Links showing list of citing references as well as cited references. Arc, Archon, Metalis and SAIL provide extended services through OpenURL field, by providing links to other services and metadata formats. None of these Service Providers are able to detect Duplicate records while harvesting from various Data Providers.

There is no uniformity in rendering values for metadata elements by Archives. For example, Arc assigns URL instead of Institution name for metadata Source DL unlike others who assign the repository name without link. OAIster renders values for Resource type and resource format interchangeably. Some archives have names like Yea, tkn, pkp, that can be expanded to be more meaningful and explicit. Thus values for Metadata Subjects, Set, Resource type, Resource formats and Deposit date, Discovery date, Date, Harvest date, Date stamp, Accession Date among Archives have not been normalized correctly by Service Providers.[5]

4. CHECKLIST FOR SERVICE PROVIDERS

Based on the analysis of search and browse interface of these Service Providers, The following checklist that may be considered by other service providers, while designing their search/browse interface. Navigation Links—Navigation in the search/browse interface can include links to Home page, Simple search, Advanced Search, Browse, Alerting services, Usage Info, Help/FAQ (Query examples), Latest Updates (Weekly, Monthly, 3 months updates), Related links to other service providers, Administration (to include registration for Login/user id and others), Additional information like OAI related institutions, Reference articles on OAI, Trouble shooting tips, Contact and Copyright information Browse Interface—Browse features can include browse by archive, institution, deposit date, author, subjects or broad topical categories, resource type, equations/formulae, latest updates (weekly, monthly).

Simple Search Interface—Searching on Author, Title and Abstract/Description Advanced Search Interface—Searchable Fields: Archive, Title, Abstract, Author (permuted names), Subject, Resource type, Date stamp, Discovery date, Archive set, Institution hosting archives; Besides keywords, should also support Equations/Formulae based searching; Search within multiple archives to be allowed; Search based on broad standard subjects/Topical categories; Combining search to Title and Abstract fields in order to retrieve only those records with abstracts; Use of Boolean operators AND, OR NOT within a field as well as across the fields; Besides author or creator, contributor and others can also be included based on the type of resource; Lateral searching of records from the search result; Case-sensitivity/Capitalization ignored, Word variations supported, punctuations to be ignored, parenthesis for grouping words; Natural language searching as in Google can be considered; Filtering/Limiting Fields: Filtering option: limiting to language, resource type.

Result set processing options—Ability to refine the search made or build the searches, inclusion of 'Search summary box'. Sorting Fields—Sorting of records by Archive, Discovery year, Subject, proximity, institution frequency, Title, Author, Date, Relevance ranking. Hit frequency or Weighted hit frequency; Default sort order can be title; No limit for sorting. Display/Saving records—Customizing Display of no. of results per page; Select/Mark/Unmark the records for display or for saving/export to some bibliographic management tool; Highlighting of search words in results; Title and KWIC among other display formats; Make HTML embedded in search results records viewable and linkable; Ability to save records during a session, download and email them; Ability to view all records without restrictions.

OpenURL and Z39.50 compliancy for use with other federated search engines.

Usage Statistics—Include list of most accessed archives, most important clients, no. of simple, advanced searches done, browse pages accessed annually.

Duplicate records detection—Implement automatic checking of duplicate records by Service Providers while harvesting metadata records.

Standardization of Archive names—Archive name followed by Institution hosting the same as well as broad subject category will make it more explicit and meaningful.

Alerting services—Alerting registered users with latest records based on saved search query; List of latest institutions/archives harvested monthly/fortnightly.

Cross-archive citation search service—Include Linking of references for each article.

Help—Context-specific help with Query Examples will be more useful/Detailed FAQ/Trouble Shooting Tips etc.

5. CONCLUSIONS

Based on the study, observation is made that the search interface of OAI Service Providers has few features as compared to extensive search features incorporated in bibliographic databases. This may be because the resources in the archives are freely accessible unlike licensed bibliographic databases. Users always tend to do quick and general searches rather than do a perfect search. The more specific the search features adopted by each Service provider, the more difficult it becomes for users, to understand and perform searches. Since they provide access to collection in the archives that are decentralized as well as each archive following their own rules in rendering information related to various metadata fields, users face difficulty in performing efficient search and retrieval from individual Service Providers. Standardization in rendering information for all metadata elements is also very essential. The archives included by individual Service Providers can be mutually exclusive. Eg General/ Comprehensive (OAIster or Arc), Subject wise (Metalis), Resourcetypewise (NCSTRL), Countrywise Service Providers etc. This will reduce unnecessary proliferation of Service Providers as well as prevent different Service providers wasting their resources in harvesting the same records from same set of archives. The archives also need to submit only to one specific Service Provider based on the nature of their resources, instead of registering with multiple Service Providers as is the case now. This will also help users enormously by saving their search time. Since Service Providers facilitate one point access to highly valuable information residing in various archives harvested by them, the search/browse interface should be as simple and at the same time include all the necessary search and retrieval features, so users can carry out their searches efficiently and effortlessly.

REFERENCES

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

[2] Hitchcock, S. The Open Citation Project Final (Year 3) *Report to JISC*. (2002). http://opcit.eprints.org/finalreport/final-report111.pdf

[3] Liu, X., Maly, K., Zubair, M.: Federated Searching Interface Techniques for Heterogeneous OAI Repositories. *Journal of Digital Information. 2(*2002). http://jodi.tamu.edu/Articles/v02/i04/Liu/

[4] Arc http://arc.cs.odu.edu

[5] METALIS. http://metalis.cilea.it/index.html

[6] OAIster. http://oaister.umdl.umich.edu/o/oaister

[7] SAIL-eprints. http://eprints.bo.cnr.it

[8] Indrani, V, Rajasekhar T B, Thulasi K, Filbert Minj *Comparison of Search and Retrieval features of OAI Harvesting Services*. Indian Institute of Science, Bangalore (January 2005).