

OAI Service Providers

Gerry McKiernan

The Open Archives Metadata Harvesting Protocol . . . is simply an interface that a networked server . . . can employ to make metadata describing objects housed at . . . [a] server available to external applications that wish to collect this metadata. (Lynch, 2002)

The mission of the Open Archives Initiative (OAI) (www.openarchives.org) is to develop and promote “interoperability standards that aim to facilitate efficient dissemination of content.” It is based on original efforts that sought to enhance access to e-print repositories such as arXiv.org “as a means of increasing the availability of scholarly communication.” The primary focus of the OAI has been the development of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) ([OAI/openarchivesprotocol.html](http://www.openarchives.org/protocol.html)).

The OAI-PMH is based on a simple and powerful model “whereby repositories (data providers) make metadata . . . available via a well-defined protocol. The exposure of the metadata allows other organizations (service providers) to harvest it and then aggregate it, post-process it, and refine it with the goal of developing services that add value. . . . Some examples of these external services are cross-repository searching, current-awareness, and reference linking” (Lagoze and Van de Sompel, 2003: 119). In this *Sci-5* column, we profile five

Gerry McKiernan, AB, MS, is Associate Professor, Science and Technology Librarian, and Bibliographer, Iowa State University Library, Ames, IA (E-mail: gerrymck@iastate.edu).

operational or experimental OAI service providers that offer access and value-added features and functionalities to major or emerging science and technology Open Archives data providers.

CITE: Clifford A. Lynch, "Metadata Harvesting and the Open Archives Initiative," *ARL Bimonthly Report* 217 (August 2001): 1-9. Also available at: <<http://www.arl.org/newsltr/217/mhp.html>> (30 August 2003).

Carl Lagoze and Herbert Van de Sompel, "The Making of the Open Archives Initiative Protocol for Metadata Harvesting," *Library Hi Tech* 21 no. 2 (2003): 118-128.

Also available at <<http://www.cs.cornell.edu/lagoze/papers/The%20Making%20of%20the%20Open%20Archives%20Initiative.pdf>> (31 August 2003).

* * *

WHAT? *Arc: A Cross Archive Search Service*

WHERE? <http://arc.cs.odu.edu/>

WHEN? October 2000

WHY? *Arc* is an experimental research service that serves as a platform for demonstrating the scalability of the OAI-PMH and as a vehicle for providing access to OAI-compliant repositories through a unified search interface.

HOW? *Arc* is the oldest federated search service based on the OAI-PMH. As of September 1, 2003, *Arc* contained nearly 6,475,000 records harvested from more than 160 repositories. Among its more common or uncommon repositories are:

- FOREX: Forschungs- und Expertendatenbank (www.forex.uni-bremen.de)
- Mathematics Preprint Server (www.mathpreprints.com)
- Informedia Digital Library (Carnegie Mellon University) (www.infsearch.cs.cmu.edu)
- National Science Digital Library (nsdl.org)
- SciELO: The Scientific Electronic Library Online (www.scielo.org)

Arc offers two major search options: 'Simple Search' and 'Advanced Search.' From the 'Simple Search' interface users can group results by 'Archive' (default), 'Discovery Year,' or 'Subject.' In addition, results can also be sorted by 'relevance ranking' (default) or 'discovery date.'

After execution, search results are displayed from the first repository that includes matching records, in a detailed format, and in relevancy order (default). The detailed format includes title, author(s), subject, discovery date, and OAI identifier. A hotlink to similar records in the *Arc* collection ('Similar

Subject') is also included for selected entries. A hyperlinked list of all repositories that contain matching records, including the first, is displayed in alphabetical order (default) to the left of the initial search results. Adjacent to each repository name are the number of matching records for a repository. The results from any of these other repositories can be subsequently displayed by clicking the repository entry of interest.

The full record for any entry can be accessed by clicking the hyperlinked title. The full record can include the name of the source repository, item identifier, author name(s), abstract, subject(s) (if available), item type, format, source, language, rights statement, OAI information, OAI identifier, and 'date stamp.' The source repository and identifier are externally linked, with the latter providing access to a record for the item within the repository. If available, the user can link to the full text of an associated record from within the repository record.

Arc also provides an advanced search option that offers field searching (i.e., 'Author,' 'Title,' and 'Abstract'). Terms can be searched in combination ('All the specified terms')(default) or individually ('Any of the specified terms') by accepting or selecting an associated radio button. Users can also use explicit Boolean operators (AND, OR) with fields and search by phrase by enclosing terms within double quotes. Results can be limited by selecting from several 'filter options' notably 'Archive,' 'Archive Set,' 'Subject,' 'Date Stamp,' or 'Discovery Date.' As in the 'Simple Search' option, users can group results by 'archive' (default), 'discovery year,' or 'subject,' and sort by 'relevance ranking' (default) or 'discovery date.' The contents of each of the component repositories of the *Arc* collection can also be individually browsed (arc.cs.odu.edu:8080/oai/results.jsp).

The *Arc* harvester and search engine software is available free-of-charge from SourceForge (sourceforge.net/projects/oaiarc/) and is released under the NCSA Open Source License.

WHO? The Digital Library Research Group, Old Dominion University (Norfolk, Virginia) (dlib.cs.odu.edu).

CITE: Xiaoming Liu, Kurt Maly, Mohammad Zubair, and Michael L. Nelson, "Arc—An OAI Service Provider for Digital Library Federation," *D-Lib Magazine* 7 no. 4 (April 2001). Available at <<http://www.dlib.org/dlib/april01/liu/04liu.html>> (10 September 2003).

* * *

WHAT? *Citebase*

WHERE? <http://citebase.eprints.org/>

WHEN? May 2001

WHY? *Citebase* “allows researchers to search across free, full-text research literature eprint archives, with results ranked according to many criteria (e.g., by citation impact), and then to navigate that literature using citation links and analysis.”

HOW? The repositories harvested for *Citebase* are:

- arXiv.org (UK mirror only)
- BioMed Central (biomedcentral.com)
- Cogprints (cogprints.ecs.soton.ac.uk)

As of September 1, 2003, *Citebase* included 266,500 source items and contained nearly 6.7 million references, with more than 1.4 million linked to their respective full text. More than 90% of the source items in *Citebase* (245,000) are derived from the arXiv.org UK mirror.

Citebase offers three search types: ‘Metadata’ ‘Citation,’ and ‘OAI Identifier.’ In a ‘Metadata’ search, the user can search by:

- ‘Author(s)’;
- ‘Title/Abstract Keyword(s)’;
- ‘Publication title’; and/or
- ‘Creation Date.’

More than one author can be searched by separating the names with a semi-colon (;) (e.g., ‘Witten E ; Nathan Seiberg’), with or without spacing between the names and the semi-colon. For a title/abstract search, users can use standard Boolean operators (e.g., AND, OR, NOT). The title of the source publication (for example, journal) can be searched using the publication title as cited by the author, which in most cases will be the standardized abbreviated publication title (e.g., ‘Phys.Rev.B.’). Any or all author, title/abstract keyword, or publication title searches can be limited to the year of ‘creation,’ or a range of years.

The results from a ‘Metadata’ search can be displayed in ‘descending’ (default) or ‘ascending’ order and ranked by one of several criteria:

- ‘Citations (Paper)’ (default);
- ‘Citations (Author)’;
- ‘Date (Creation)’;
- ‘Date (Update)’; or
- ‘Hits (Author).’

For each entry in the default display, the author(s), title, ‘creation’ date, identifier, abstract, and ‘comment’ text are provided. A number appropriate to the selected ranking option is also included and is located to the left of an author name.

A full record is provided for all entries regardless of ranking format and includes a variety of bibliographic data and novel features and functionalities (see Tables 1 and 2).

The graph of a document’s citation/hit history includes a table that provides the ‘citations identified,’ the total number of Web hits, and the mean number of hits for an author in the *Citebase* database for identified citations and Web hits. References are listed as in original source documents (e.g., numbered, alphabetical, or alphanumeric). For those references with associated full-text, a hotlinked term (‘journal’ or ‘eprint’) is noted to the left of a cited reference. Selecting of ‘eprint’ citations will retrieve the full *Citebase* record for the item with its associated bibliographic data and various citation-related features and functionalities (see Tables 1 and 2), while selection of ‘journal’ citations will retrieve a brief record that typically includes basic bibliographic data and a link to the full text of the cited document.

Users can also search *Citebase* using a known ‘Identifier’ (e.g., ‘oai:arXiv.org:hep-th/9304011’). As with results from a ‘Metadata’ search, results in an ‘Identi-

TABLE 1. Bibliographic Data Provided in a Full *Citebase* Record

- full document title;
- author names;
- full abstract
- ‘Comment’;
- a link to a full-text PDF version of the document;
- a hotlinked identifier and harvest date;
- a ‘creation date’; and
- document type.

TABLE 2. Value-Added Features and Functionalities Provided from Within a Full *Citebase* Record

- a graph of the document’s ‘citation/hit history’;
- the document’s cited references (‘This Article’s Reference List’);
- brief record data for the ‘Top 5 Articles Citing this Article;’
- a listing of ‘All Articles Citing this Article’ in brief record format, with entries listed in descending order by number of times the citing paper itself has been cited (‘Citations (Paper)’);
- a listing of the ‘Top 5 Articles Co-cited with this Article’; and
- a listing of ‘All Articles Citing this Article.’

fier' search can be displayed in descending or ascending order and ranked by one of several criteria, including one unique to this search option, 'co-citedness,' the co-occurrence of two (or more) different references in the same document. In an 'Identifier' search, there are three search and display options: a user can request that the standard full *Citebase* record be displayed for a document by executing an 'Abstract' search, request a listing of documents in brief format with which a identified document has been 'Co-cited With,' or request a brief format display of those *Citebase* documents that have cited it ('Cited By'). A 'Citation' search option is also available in *Citebase* and is intended to permit users to retrieve the full text of a document using a standard citation.

Citebase is an experimental demonstration service and users are cautioned not to use it for academic evaluation purposes. Coverage and functionalities in *Citebase* are limited to "citing and cited papers that their authors have already archived in the source eprint archives," to "cited papers that can currently be successfully linked," and "for arXiv, for now, on the usage/hit data for its UK-site only." Currently, Web log usage data ('hits') date from August 1999 to the present.

WHO? *Citebase* was developed by Tim Brody (tdb01r@ecs.soton.ac.uk) as part of the Open Citation Project (opcit.eprints.org), a project of the Intelligence, Agents and Multimedia Group, Department of Electronics & Computer Science, at the University of Southampton, UK.

CITE: Steve Hitchcock, Donna Bergmark, Tim Brody, Christopher Gutteridge, Les Carr, Wendy Hall, Carl Lagoze, and Stevan Harnad, "Open Citation Linking: The Way Forward." *D-Lib Magazine* 8 no. 10 (October 2002). Available at: <<http://www.dlib.org/dlib/october02/hitchcock/10hitchcock.html>> (6 September 2003).

* * *

WHAT? *my.OAI*

WHERE? <http://www.myoai.com>

WHEN? March 2002

WHY? *my.OAI* is "a full-featured search interface to a selected list of metadata databases from the Open Archives Initiative."

HOW? As of September 1, 2003, *my.OAI* contained nearly 6.6 million records harvested from fifteen repositories. Among its more common or uncommon repositories are:

- CERN Document Server (cds.cern.ch)
- Citebase (citebase.eprints.org)
- Historical NCSTRL Collection (historical.ncstrl.org)
- Open Video Project (www.open-video.org)
- Project Euclid (projecteuclid.org).

my.OAI provides two different, yet similar search forms. The default (expanded) form includes one additional search field option not provided in a contracted version, as well as three limit options. Using either version, the user can:

- search in any field;
- search for words in specified field;
- combine search terms from fields in either an AND or OR Boolean combination, or search for an exact phrase;
- select one or more (or all repositories) to search;
- use a pull-down menu, specify the number of records to be displayed per page; and/or
- sort documents.

The user can display search results with records in a brief record format ('View search results') (default), view the number of records ('hits') retrieved for each selected repository ('View meta search results'), or view the results grouped by repository ('View search results grouped by database').

In the default display format, results are listed in a brief HTML record format in descending order by relevancy. Above the listing is a restatement of the executed search strategy and a concatenated listing of selected repositories and a link to an alternative XML display option for the search results. A brief record within *my.OAI* generally includes the following fields:

- a relative relevance rating number;
- author name(s), 'publication' year, and title;
- description (abstract) (if available) (default);
- format display options;
- 'External Link' (link to an original source record, if available);
- date;
- repository name; and
- 'Create Annotation' option.

The author, publication year and title are collectively hotlinked to a full record for the item, which typically provides the title of the item, the name of the 'creator' (author(s)), subject terms or phase (if available), 'description' (abstract), 'publication' date, type of document, the Uniform Resource Locator (URL) for the source item, and 'Links.'

In addition to this conventional display option, the pull-down menu offers the user several advanced features and functionalities that include:

- viewing the documents with similar documents ('View selected documents with similar documents');
- viewing the documents with recommended documents ('View selected documents with recommended documents'); and
- viewing the documents in their 'native' format (e.g., text or XML format).

my.OAI can be accessed either as a guest or as a registered user. A registered user can utilize a number of special features unavailable to guest users, most notably the ability to set preferences, save searches or records for later use, e-mail selected records, or append comments (annotations) to documents. Registered users can also activate a *my.OAI* e-mail alerting service to receive updates that match a saved search strategy. In addition, registered users can save and store search results in a personal folder on the *my.OAI* server.

my.OAI offers one of the widest arrays of search options and operators of any non-commercial or commercial interface (www.myoai.com/search/html/HelpSearch.html).

WHO? François Schiettecatte, Principal, FS Consulting, Inc., Salem, Massachusetts.

* * *

WHAT? *Open Archives Initiative Information in Engineering, Computer Science, and Physics* (Grainger Engineering Library at the University of Illinois at Urbana-Champaign)

WHERE? <http://g118.grainger.uiuc.edu/engroai/>

WHEN? September 2002

WHY? *Open Archives Initiative Information in Engineering, Computer Science, and Physics* provides access to major local, national, and international OAI-compliant repositories in computer science, engineering, physics, and related subject areas.

HOW? As of September 1, 2003, the *Open Archives Initiative Information in Engineering, Computer Science, and Physics* service contained more than 484,000 records harvested from 13 repositories. Among its more common or uncommon repositories are:

- arXiv.org
- Institute of Physics journals in physics and related disciplines (www.iop.org/EJ/)
- California Institute of Technology (Caltech) Electronic Theses and Dissertations (etd.caltech.edu)
- University of Illinois at Urbana-Champaign Engineering Documents Center Collection (shiva.grainger.uiuc.edu/engdoc/)
- Wolfram Research Functions (functions.wolfram.com)

A user can search the by 'Author,' 'Title/Subject/Abstract' (default), 'Report Number/Journal Source,' 'Title,' 'Subject,' 'Abstract,' 'Publisher,' 'Date,' 'Language,' or 'Any Field' by selecting the field name from a pull-down menu found adjacent to a primary search box and entering appropriate terms. Terms can be combined with other terms by entering these into a secondary search box. A tertiary search box is also available. The user can combine the terms (or phases) from two (or three) search boxes choosing from one of three Boolean operations: 'also must contain' (AND), 'or may contain' (OR), or 'but not contain (AND NOT) from available pull-down menus.

A search can be limited to the entire *Open Archives Initiative Information in Engineering, Computer Science, and Physics* collection ('All Collections' (default)), or to a specific repository by selecting the one of interest from pull-down. Search results can be sorted by relevance ('Relevance'), by collection ('Collection'), or not at all ('None' (default)). Upon the execution of a search, the results are displayed as a numbered set of brief records. Each brief record typically contains the collection name ('Collection'), 'Identifier,' 'Title,' and 'Creator.'

In addition, at the base of each record there is a link that allows that user to 'View Complete Metadata Record' or to 'Add to a Book Bag.' At least one identifier is hotlinked and provides access to a brief source repository record. From within the source repository the user can then select from available full text access options (e.g., arXiv.org: "Full-text: PostScript, PDF, or Other formats." The 'Add to Book Bag' function allows the user to collect records (in brief format) in a separate collection that can be subsequently printed or saved (currently in XML format). From the service search page, users can access their respective 'Book Bag' collection as well their session 'Search History,' which allows the user to re-execute ('Redo') or modify (and subsequently re-execute) ('Modify Search') any previous search query.

The full record ('View Complete Metadata Record') for an item can include 'Identifier,' 'Datestamp,' 'SetSpec,' 'Title,' 'Creator,' 'Subject' (if available), 'Description' (Abstract), 'Date' and document 'Type.'

WHO? William H. Mischo (w-mischo@uiuc.edu), Head, Grainger Engineering Library Information Center, and Director, Beckman Institute Library, University of Illinois at Urbana-Champaign (gateway.library.uiuc.edu/granger/).

* * *

WHAT? *SAIL-eprints* (Search, Alert, Impact and Link)

WHERE? <http://eprints.bo.cnr.it/>

WHEN? April 2003

WHY? *SAIL-eprints* (*Search, Alert, Impact and Link*) is “an electronic open access service provider for finding scientific or technical documents, published or unpublished, in Chemistry, Physics, Engineering, Materials Sciences, Nanotechnologies, Microelectronics, Computer Sciences, Astronomy, Astrophysics, Earth Sciences, Meteorology, Oceanography, . . . [Agriculture], and related . . . [subjects].”

HOW? *SAIL-eprints* “has been designed primarily to collect information on scientific documents (metadata) authored by [Consiglio Nazionale delle Ricerche [Italian National Research Council] (CNR), Italy] . . . researchers and deposited as preprints or postprints in CNR institutional open access archives.” In addition, metadata from other data providers that cover identical or related scientific fields is also gathered. As of September 1, 2003, *SAIL-eprints* contained nearly 319,000 records harvested from 28 distinct repositories (eprints.bo.cnr.it/cgi-bin/info.pl).

Among the more common or uncommon repositories are:

- arXiv.org
- BioMed Central (www.biomedcentral.com)
- DSpace™ at MIT (libraries.mit.edu/dspace-mit/)
- Indian Institute of Science (eprints.iisc.ernet.in)
- Organic Eprints (orgprints.org).

SAIL-eprints offers two search interfaces: a ‘Simple Search’ and an ‘Advanced Search.’ The *SAIL-eprints* ‘Simple Search’ allows users to search only the title and/or abstract metadata gathered from all harvested sites. By default, brief records are displayed for a maximum of 20 matching items per page, although the user can increase the number to 100 items in groups of twenty records. After execution, a listing of repositories that contain matching results is

presented on the left side of the page. To display the results from a given repository, the user clicks on the repository name (e.g., 'arXiv'), which lists retrieved results in a brief record format. Among other data and information, this format provides the title, subject(s) (if available in the original), author(s), deposit date, and an excerpt of the context of the search term(s).

From within the original, brief record listing, the user can display the detailed record for an item by clicking the associated 'Show Detail' hotlink, the last field in the brief record format. The detailed record includes the name of the repository (e.g., 'arXiv'), a 'Record ID' (e.g., 'oai:arXiv.org:hep-lat/9909157'), harvest and deposit dates, an identifier (e.g., 'http://arXiv.org/abs/hep-lat/9908013 ; Nucl. Phys. B575 (2000) 255-266'), title, author name(s), subject and abstract (if available in the original), item publication date, item type (e.g., 'text') and a note on 'Extended Services' (e.g., 'No Resolver'). The last function is an experimental OpenURL service that currently provides links to the Google search engine and the CNR OPAC for select records. If available, the full text of the associated item can be retrieved by clicking the hotlinked identifier from within the brief record and then the linked repository record document options.

In an 'Advanced Search,' the user can search a specific repository (e.g., 'arXiv.org'), a repository categorized within a specific discipline (e.g., 'Aeronautics,' 'Chemistry,' 'Mathematics'), or the repository of a particular institution or organization (e.g., 'CALTECH (California Institute of Technology)'). More than one repository can be searched within the repository option by pointing, holding, and clicking individual entries. All repositories, or disciplines, and/or institutions, can also be selected by clicking the 'Select ALL/None' hotlink located below each respective listing. In addition to selecting individual entries within each group, the user is required to accept the default (for the repository grouping) or click the radio button located above the associated listing of either the 'Discipline' or 'Institutions' grouping.

After choosing one or more individual repositories, a discipline category, or institutional repository, the user proceeds to the second step in the advanced search option—entry of the names of specific authors, and/or title and/or abstract keywords in one of three fields with associated pull-down menus for these options (i.e., 'author,' or 'title/abstract'). Field terms may be combined in either an OR (default) or AND Boolean statement by accepting or selecting the associated radio button beneath the search options page. As with the 'Simple Search' option, brief records are displayed to a maximum of 20 matching items per page as a default, although as noted, the number can be increased.

Users can also browse the collection by author surname or deposit date ('Browse Indexes By *Author *Deposit Date') or by 'Latest Updates.' Access statistics are also available, providing data and bar charts on usage, and aggregated

search and browsing activity (http://eprints.bo.cnr.it/cgi-bin/show_stat.pl). In addition, *SAIL-eprints* offers an e-mail alerting service to registered users.

WHO? Silvana Mangiaracina (mangiaracina@area.bo.cnr.it), librarian of the Biblioteca di Area Della Ricerca Di Bologna, Consiglio Nazionale delle Ricerche, Italy (<http://biblio.bo.cnr.it/>).

* * *

NOMINATIONS

Members of the science and technology community are invited to nominate quality science and technology Web sites and resources for potential review in *Sci-5*. Of greatest interest are sites with uncommon but useful content, and those with innovative features and functionalities. Nominations should be sent to Gerry McKiernan (gerrymck@iastate.edu).