

## Gerry McKiernan

**Open archives initiative service providers. Part I: science and technology**

The Open Archives Metadata Harvesting Protocol [is] an important new infrastructure component for supporting distributed networked information services. [It offers] a mechanism that enables data providers to expose their metadata ... and [provides] a fascinating array of new services and system architectures for a diverse set of communities (Lynch, 2001, p. 1).

**Open archives initiative protocol for metadata harvesting (OAI-PMH)**

As stated in its mission statement, the purpose of the Open Archives Initiative (OAI) ([www.openarchives.org](http://www.openarchives.org)) is to develop and promote “interoperability standards that aim to facilitate the efficient dissemination of content.” At its root was a vision that sought to “stimulate the growth of open e-print repositories” as an alternative method for the “rapid dissemination of research results.” At the core of the OAI is the Open Archives Protocol for Metadata Harvesting (OAI-PMH) ([www.openarchives.org/OAI/openarchivesprotocol.html](http://www.openarchives.org/OAI/openarchivesprotocol.html)), a protocol that permits service providers to harvest, aggregate, post-process, and refine metadata harvested from local repositories, and enables them to develop value-added services such as cross-repository searching, current-awareness, and reference linking (Lagoze and Van de Sompel, 2003, p. 119).

Since the launch of the arXiv e-print server more than a decade ago, several dozen data providers (repositories) have been established ([www.openarchives.org/Register/BrowseSites.pl](http://www.openarchives.org/Register/BrowseSites.pl)). More recently, a number of service providers ([www.openarchives.org/service/listproviders.html](http://www.openarchives.org/service/listproviders.html)) have emerged that provide not only access, but also enhanced features and functionalities to

a wide variety of select science and technology repositories[1].

**Arc: a cross archive search service**

Arc: a cross archive search service ([arc.cs.odu.edu](http://arc.cs.odu.edu)) 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 (Liu *et al.*, 2001). Formally announced in October 2000, Arc is the oldest “federated search service based on the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). It includes a harvester that harvests OAI-PMH 1.x and 2.0 compliant repositories, a search engine together with a simple and advanced search interface, and an OAI-PMH layer over harvested metadata.”

As of September 1, 2003, ARC contained nearly 6,475,000 records harvested from more than 160 repositories that include:

- [arXiv.org](http://arXiv.org).
- ECS EPrints Service (Electronics and Computer Science Department, University of Southampton, UK) ([eprints.ecs.soton.ac.uk](http://eprints.ecs.soton.ac.uk)).
- FOREX: Forschungs- und Experimentdatenbank ([www.forex.uni-bremen.de](http://www.forex.uni-bremen.de)).
- Informedia Digital Library (Carnegie Mellon University) ([www.infsearch.cs.cmu.edu](http://www.infsearch.cs.cmu.edu)).
- Internet Scout Project OAI Repository ([scout.cs.wisc.edu](http://scout.cs.wisc.edu)).
- Mathematics Preprint Server ([www.mathpreprints.com](http://www.mathpreprints.com)).
- National Science Digital Library ([nsdl.org](http://nsdl.org)).
- Project Euclid (Cornell University) ([projecteuclid.org](http://projecteuclid.org)).
- SciELO: The Scientific Electronic Library Online ([www.scielo.org](http://www.scielo.org)).
- Virginia Tech Imagebase ([imagebase.lib.vt.edu](http://imagebase.lib.vt.edu)).

Arc offers two major search options: “Simple Search” and “Advanced Search.” In a “Simple Search” the user can “search all bibliographic fields” that include not only the author, title, and abstract fields, but additional fields such as archive name, discovery date, language, and subject type, if available ([arc.cs.odu.edu/service\\_help.html](http://arc.cs.odu.edu/service_help.html)).

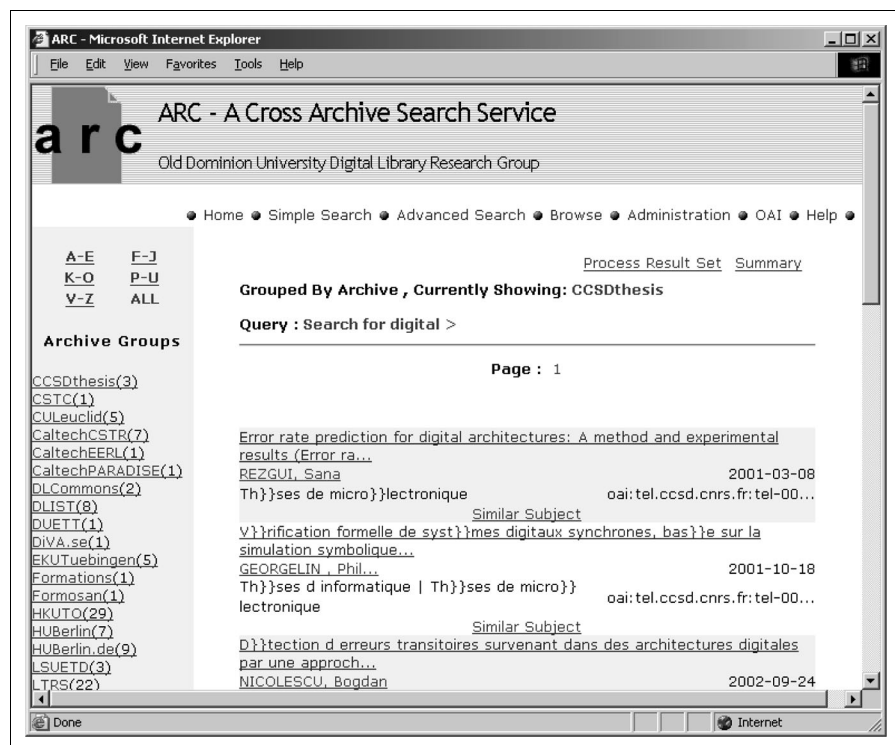
Within Arc, a search is performed such that it will match any and all records that include terms that begin with the respective letters (e.g. the letters “comp” will match “computer” and “computation” as well as “comprehensive.” A phrase search can be performed by enclosing the terms in double quotes.

From the “Simple Search” interface users can group results by “archive” (default), “discovery year”, or “subject”, by accepting or selecting either of these options from a pull-down menu. In addition, results can also be sorted by “relevance ranking” (default) or “discovery date” by accepting or choosing either from a different pull-menu.

After execution, search results are displayed in a detailed format from the first repository that includes matching records in relevancy order (default). The detailed format provides basic bibliographic data for the item, notably title, author(s), subject, discovery date, and OAI identifier (see Figure 1). A hotlink to similar records in the Arc collection (“Similar Subject”) is also included for select 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 enclosed in parentheses. The results from any of these other repositories can be subsequently displayed by clicking the repository entry of interest. The full name of most repositories can be

**Figure 1**

A screen print from Arc "Simple Search" showing select relevant records in a detailed format from the first matching repository ("CCSDthesis"). A partial listing of other repositories that contain relevant records is located along the left border



displayed by rolling-over an individual repository entry. The total number of matching records in all relevant Arc repositories ("Total Size") is displayed at the end of the alphabetical listing.

The user can display only the title of entries by selecting the "Summary" option found near the upper right-hand corner of the results display, or refine a search by selecting the "Process Results Set" found to the left of the "Summary" hotlink. The "Process Results Set" option allows the user to "re-organize the result set by grouping or sorting", "refine the result set by Author, Subject, Title or Archive", or "refine the result set by Discovery Date . . . ."

The full record for any entry can be accessed by clicking the hyperlinked title, which is subsequently displayed in a pop-up window. The full record includes the name of the source repository, item identifier, author name(s), abstract, subject(s) (if available) item type (e.g. "text"), format (e.g. "paper") source, language, rights statement, OAI information, OAI

identifier, and "date stamp." The source repository and identifier are usually externally linked, with the latter typically providing access to a record for the item within the repository. If available, the user can link to the full text for an associated record from within the repository record.

An author name is internally linked and provides access to an appropriate segment of an author index from which the user can review, select, and search for other items authored by an individual in the Arc collection. The full record display also includes a "Service" field that provides a link to an annotation function intended to permit users to append comments to a record, and as a "link to other services and metadata format" ("DP9" icon). This latter function links to the DP9 service (see below), which provides a brief Dublin Core record as well as links to document metadata in alternative formats. For select records, a link to document references is also offered in the "Service" field.

Arc also offers users an "Advanced Search" option that permits specific field searching (i.e. "Author", "Title", and "Abstract"). Terms can be searched in combination ("All the specified terms") (AND) (default) or individually ("Any of the specified terms") (OR) by accepting or selecting the associated radio button for the option. Users can also use explicit Boolean operators (AND, OR) within 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". A specific archive can be selected from the alphabetical listing of archives displayable in a pull-down menu; at this time, only one archive (or the entire collection) can be selected at a time. A range of dates from either the "Date Stamp" or "Discovery Date" options can be selected by using a horizontal slide function. Users can group results by "Archive" (default), "Date," or "Subject," and sort by "Date", "Subject" or "Title." Within the Arc 'Advanced Search' users can also search within search results by selecting the 'Search Last Results' option. 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).

Arc was developed by the Old Dominion University Digital Library Research Group (dlib.cs.odu.edu), which in addition to the development of Arc is engaged in building and demonstrating novel digital library services. Among these are DP9 (an open source gateway service that allows general search engines such as Google to index OAI-compliant archives (128.82.7/dp9)), 99Kepler ("a self-contained, self-installing software system that functions as an Open Archives Initiative data provider") (kepler.cs.odu.edu:8080/kepler/index.html), Archon (an OAI-compliant federated digital library with an emphasis on physics for the National Science, Mathematics, Engineering, and Technology Education Digital Library (archon.cs.odu.edu)), and an OAI-based framework for the Networked Computer Science Technical Reference Library (NCSTR) (www.ncstrl.org).

The Arc harvester and search engine software is available free-of-charge from SourceForge ([sourceforge.net/projects/oaiarc/](http://sourceforge.net/projects/oaiarc/)) and is released under the NCSA Open Source License. Arc is based on Java Servlet technology and requires JDK1.4, Tomcat 4.0x, and a RDBMS server (tested with Oracle and MySQL).

## Citebase

Citebase ([citebase.eprints.org](http://citebase.eprints.org)) “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.” The data providers harvested for Citebase are:

- arXiv.org (UK mirror only).
- BioMed Central ([biomedcentral.com](http://biomedcentral.com)).
- Cogprints ([cogprints.ecs.soton.ac.uk](http://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. Of its source items, more than 90 percent (245,000) were derived from the arXiv.org UK mirror. The current version of Citebase was established in May 2001, preceded by experimental trials (Brody, 2002; Hitchcock *et al.*, 2002).

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
- “Creation Date”.

Author names can be specified by using the following notation: family\_name, given\_name, or given\_name family\_name. More than one author can be searched by separating their 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). “Searches can also be specified in a similar manner to most WWW search engines, with the exception of phrase

searches ...” While select plural suffixes (i.e. “s” or “es”) and the “ing” suffix are automatically removed for root searching, word stemming per se is not currently supported. 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 form of the 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); This number is defined as the total number of citations identified by Citebase to a document.
- “Citations (Author)”;
- This number is defined as the total number of citations identified by Citebase to Citebase indexed authored documents divided by the number of documents that an individual has authored and which are indexed in Citebase.
- “Date (Creation)”;
- The date given may be that provided by the author, or may be the date when a record for the document was added to a repository.
- “Date (Update)”;
- or The most recent date when a change was made to the record for the document, not necessarily the document itself.
- “Hits (Author)”
- This number is defined as the mean number of author hits for a named author. The number is calculated by dividing the total number of ‘hits’ to documents authored by an individual by the total number of documents authored by that same individual.

The retrieved results are displayed in an HTML format as a default. Alternatively, users can display the results in either the XML, “Refer” or a “BibTeX” format by selecting either of these options from above the search results listing.

For each entry in the default display, the author(s), title, creation, 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 the author name. For example, in the case of the “Citations (Paper)” ranking option, the number of citations to a listed document is displayed; in the case of “Hits (Paper)” option, the total number of hits for the document is provided. For each entry in the default HTML format listing, a link to a document’s “Abstract/Citations” and PDF version (if available) is also provided (see Figure 2).

A full record is provided for all entries regardless of ranking format and provides a variety of bibliographic data:

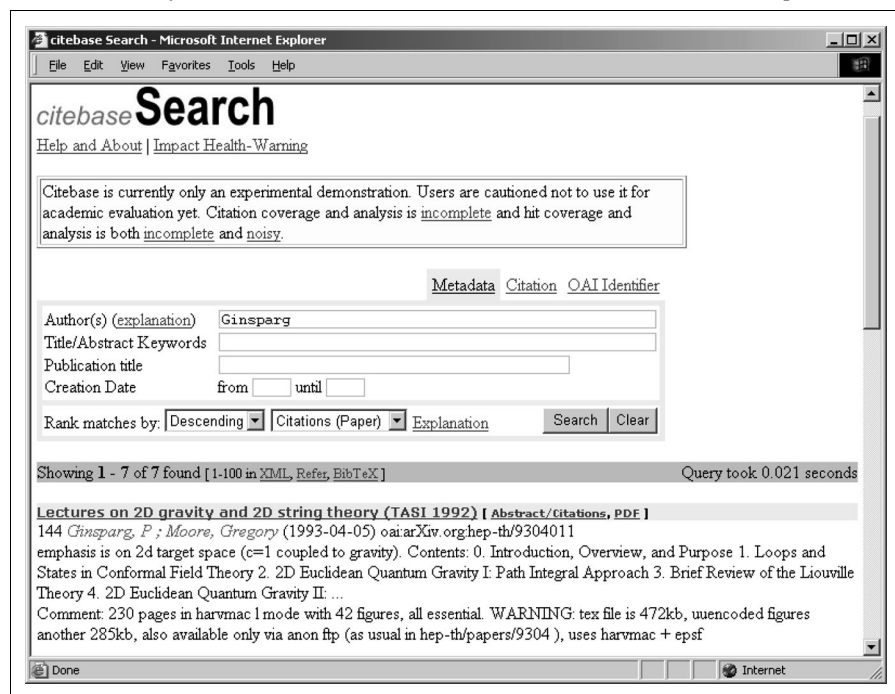
- full document title (e.g. “Lectures on 2D gravity and 2D string theory (TASI 1992)”);
- author names (e.g. “Ginsparg, P.; Moore, Gregory”);
- full abstract;
- “Comment” (“230 pages in harvmac 1 mode with 42 figures, all essential. WARNING: tex file is 472kb, uuencoded figures another 285kb, also available only via anon ftp (as usual in hep-th/papers/9304), uses harvmac + epsf”);
- a link to a full-text PDF version of the document (indicated by a green icon labeled “PDF”);
- a hotlinked identifier and harvest date (e.g. “oai:arXiv.org:hep-th/9304011 (2003-05-28)”);
- a “creation date” (“1993-04-05”); and
- document type (e.g. “text”).

and novel features and functionalities:

- a graph of the a 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 Cited with this Article”; and
- a listing of “All Articles Citing this Article”.

**Figure 2**

Search results from a "Metadata" author search in Citebase with sample record



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 the original source document (e.g. numbered, alphabetical, or alphanumerical). For references with associated full-text, a hotlinked term ("journal" or "eprint") is noted to the left of a citation. 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 lists above), 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 in one or more formats.

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 "Identifier" 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: 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 (e.g. Turing, A.M. (1950) "Computing Machinery and Intelligence", *Mind*, 49:433-460).

Citebase is an experimental demonstration service and users are cautioned not to use it for academic evaluation purposes. Its coverage and functionalities 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. "Citebase offers both a human user interface ... and an Open Archives (OAI)-based machine interface for further harvesting by other OAI services" (e.g. my.OAI (see below).

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, that was originally funded by the Joint NSF - JISC International Digital Libraries Research Programme (www.dli2.nsf.gov/internationalprojects/intlprojects.html).

## my.OAI

my.OAI (www.myoai.com) is "a full-featured search interface to a select list of metadata databases from the Open Archives Initiative" formally launched in March 2002. As of September 1, 2003, my.OAI contained nearly 6.6 million records harvested from 15 repositories that include:

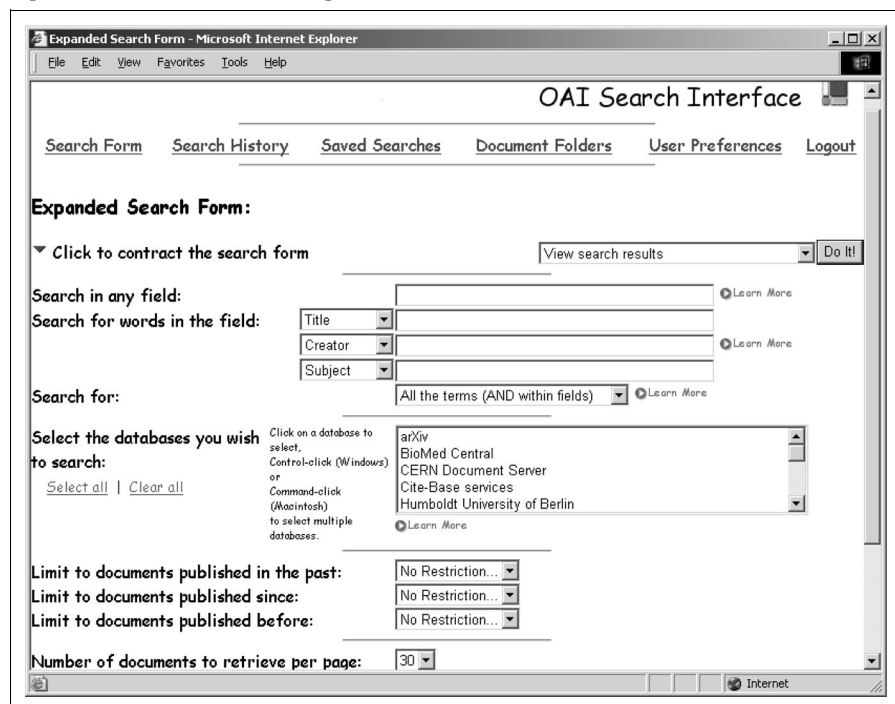
- arXiv.org.
- CERN Document Server (cds.cern.ch).
- Citebase (citebase.eprints.org).
- Digital Library of MIT Theses (theses.mit.edu).
- Historical NCSTRL Collection (historical.ncstrl.org).
- Humboldt University of Berlin Document Server (dochost.rz.hu-berlin.de).
- Langley Technical Report Server (techreports.larc.nasa.gov/ltrs/).
- National Advisory Committee for Aeronautics server (naca.larc.nasa.gov).
- 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 (see Figure 3). Using either version, the user can:

- search in any field;
- search for words in specified field ("title", (default) "creator" (author), "subject", "description" (abstract), "contributor" (author), "publisher" (associated organization), or "identifier");
- combine search terms from fields in either an AND or OR Boolean

**Figure 3**

The my.OAI “Expanded Search Form” showing search boxes, a partial listing of repositories, and date limit options



combination, or search for an exact phrase;

- select one or more (or all) repositories) to search;
- using a pull-down menu, specify the number of records to be displayed per page (“5”, “10” (default), “15”, “20”, “25”, or “30”); or
- sort documents (“Relevance” (default), “Date – Newest first, Date – Oldest first”).

Within the expanded search form the user can limit search results by the following three variants of date:

- “Limit to documents published in the past” (“Week”, “Month”, “3 Months”, “6 Months”, “9 Months”, or “Year”);
- “Limit to documents published since” a selected year (1990-2003);
- “Limit to documents published before” a selected year (1990-2003).

After entering search terms (or phrases), the user can select the format in which the results are displayed by choosing from options made available in a pull-down menu located above the

search term entry fields. The user may 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”). From within the “View meta search results” display, users can then select the specific repositories from which search results are displayed.

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, 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 typically includes the following fields:

- a relative relevance rating number;
- author name(s), “publication” year, and title;
- description (abstract) (if available) (default);
- format display options (“HTML” or “XML Text”) and their associated file size;

- “External Link” (link to an original source record, if available);
- date;
- repository name (in parentheses); and
- “Create Annotation” option.

The last option – displayed only to a user who has previously established a personal my.OAI account – provides a link to a separate window that permits the user to append comments to a selected record. Access to the annotation can be limited by the user for his or her personal use (“Just Me”) or made publicly available (“Everyone”). A link to the record annotation (“View Annotations”) is provided when the record is next retrieved.

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 (month, day, and year), type of document (e.g. “text”), the Uniform Resource Locator (URL) for the source item, and “Links”. The latter option enables the user to directly search the Web by the title of the document (“Search Google with the title”) or to perform a search using the abstract text (“Search Google with the description”).

From within either the merged or grouped search results, the user can select individual records by checking a box located to the left of each brief record, or mark all results using a “Select All” feature. After choosing one or more records, the user may display the selected record(s) in the full record format by accepting and executing the display option found as the default in a pull-down menu.

In addition to this conventional display option, a 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”); or

- viewing the documents in their “native” format (e.g. text or XML format).

In viewing “selected documents with similar documents,” records that are determined to be “similar” to the selected record(s) are displayed below each of these originally selected records respectively. In viewing “selected documents with recommended documents”, the other potential useful documents identified by the “prior search and retrieval patterns of other users” are displayed in a similar manner.

my.OAI can be accessed either as a guest or as a registered user. A registered user can utilize a number of special features that are 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 personalized folder on the my.OAI server. Users can access their respective folders, “Search History”, “Saved Searches”, “User Preferences” form, or a new “Search Form”, from the top (or bottom) of nearly all my.OAI pages.

Users may personalize a variety of my.OAI features and functionalities from a “User Preferences” form ([www.myoai.com/search/Search.cgi/UserPreferences](http://www.myoai.com/search/Search.cgi/UserPreferences)), including:

- (1) “User Information”;
  - “User login”
  - “User name”
  - “User e-mail address”
- (2) “Search Preferences”;
  - “Search form default”
  - “Search operator default”
  - “Number of documents to retrieve per page”
  - “Search sort order”
  - “Maximum number of entries to save in search history”
- (3) “Database Selection Defaults”;
- (4) “Data Summary Preferences”;
  - “Document summary type”

- “Document summary length in words”
- (5) “Document Retrieval Preferences”; and
    - “Number of similar documents retrieved”
  - (6) “Saved Search Defaults”;
    - “Activate saved search by default”
    - “Save search frequency”
    - “Saved search delivery format”
    - “Saved search delivery method”.

my.OAI offers one of the widest arrays of search options of any non-commercial or commercial interface ([www.myoai.com/search/html/HelpSearch.html](http://www.myoai.com/search/html/HelpSearch.html)). Among its conventional search features are:

- key word;
- standard Boolean operators (e.g. AND, OR, NOT; nested statements);
- proximity (e.g. NEAR/WITHIN);
- phrase searching (enclosed in quotes);
- wildcard characters (e.g. ‘\*’ (truncation), ‘?’ (single character replacement), ‘@’ (single letter replacement), ‘#’ (single digit replacement));
- automatic stemming.

Among its more sophisticated and novel search functions are:

- “Soundex”; This function allows the user to search for documents containing terms with the same soundex key as the one for the term specified. Soundex functionality based on “an algorithm for encoding a word so that similar sounding words encode the same” (wombat.doc.[ic.ac.uk/foldoc/foldoc.cgi?soundex](http://ic.ac.uk/foldoc/foldoc.cgi?soundex)).
- “Metaphone”; This function allows the user to search for documents containing terms with the same metaphone key as the one for the term specified. Metaphone functionality is based on “an algorithm for encoding a word so that similar sounding words encode the same.” It is similar to soundex in purpose, but is more accurate (wombat.doc.

[ic.ac.uk/foldoc/foldoc.cgi?metaphone](http://ic.ac.uk/foldoc/foldoc.cgi?metaphone)).

- “Phonix”; This function allows the user to search for documents containing terms with the same phonix key as the one for the term specified. Phonix is similar to soundex, but is based on a more complex algorithm ([www.jonathan.net/DOCS/phonix&soundex.doc](http://www.jonathan.net/DOCS/phonix&soundex.doc)).
- “Typo”; and The typo function allows the user to search for documents containing terms that contain simple typographical errors such as missing or juxtaposed letters.
- “Thesaurus” The thesaurus function allows the user to expand the search using the thesaurus optionally defined in the my.OAI search engine.

my.OAI also offers several “modifier” functions that allow modification of Boolean, or sorting default values, or the default display of search results. In early September 2003, Rich Site Summary (RSS), a “lightweight XML format designed for sharing headlines and other Web content” was implemented on my.OAI.

Case sensitivity in my.OAI is atypical: “search terms entered in mixed case will be searched for as such; for example a search for ‘Animal’ will return documents containing the term ‘Animal’ [initial capital letter] but not documents containing the term ‘animal’ [lower case capital letter] whereas a search for ‘animal’ [lower case initial letter] will return all documents containing that term ...” regardless of case. In addition, author names are normalized using a simple scheme where the author initials are appended after the author name, separated by an underscore, (e.g. the author name “William H Brown”, would be searched as “Brown\_WH”).

my.OAI is made available by FS Consulting, Inc., Salem, Massachusetts, François Schiettecatte, Principal ([webmaster@myoai.com](mailto:webmaster@myoai.com)). It is “built using the MPS Information Server, MPS Information Server Perl Interface and the MPS Information Server Search Interface developed by FS Consulting, Inc.” ([www.fsconsult.com](http://www.fsconsult.com)). A Perl-based OAI Harvester used to harvest all the data searchable in my.OAI is available free of charge ([www.myoai.com/downloads/](http://www.myoai.com/downloads/)).

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

Established in September 2002, the Open Archives Initiative Information in Engineering, Computer Science, and Physics service ([g118.grainger.uiuc.edu/engroai/](http://g118.grainger.uiuc.edu/engroai/)) provides access to major local, national, and international OAI-compliant repositories in computer science, engineering, physics, and related disciplines. 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 that include:

- arXiv.org.
- Institute of Physics journals in physics and related disciplines ([www.iop.org/EJ/](http://www.iop.org/EJ/))
- California Institute of Technology (Caltech) Electronic Theses and Dissertations ([etd.caltech.edu](http://etd.caltech.edu))
- Caltech Earthquake Engineering Research Laboratory Technical Reports ([caltecheerl.library.caltech.edu](http://caltecheerl.library.caltech.edu))
- Digital Library of MIT Theses ([theses.mit.edu](http://theses.mit.edu))
- Langley Technical Report Server ([techreports.larc.nasa.gov/ltrs/](http://techreports.larc.nasa.gov/ltrs/))
- National Advisory Committee for Aeronautics Server ([naca.larc.nasa.gov](http://naca.larc.nasa.gov))
- University of Illinois at Urbana-Champaign Engineering Documents Center Collection ([shiva-grainger.uiuc.edu/engdoc/](http://shiva-grainger.uiuc.edu/engdoc/))
- Wolfram Research Functions ([functions.wolfram.com](http://functions.wolfram.com))
- Wolfram Research MathWorld ([mathworld.wolfram.com](http://mathworld.wolfram.com))

A user can search the service 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 others by entering these secondary terms into a secondary search box and selecting an appropriate field from an identical

pull-down menu. A tertiary search box is also available. The user can combine the terms (or phrases) using one of three Boolean operations: "also must contain" (AND), "or may contain" (OR), or "but not contain" (AND NOT).

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 and can be sorted by relevance ("Relevance"), by collection ("Collection"), or not at all ("None" (default)).

Upon the execution of a search, results are displayed as a numbered set of brief records. Each brief record typically contains:

- Collection name ("Collection") (e.g. "ArXiv: E-Print Archive in Physics and Related Disciplines");
- "Identifier" (e.g. "<http://arXiv.org/abs/cond-mat/0211397>");
- "Title" (e.g. "Patterned nanostructure in AgCo/Pt/MgO(001) thin film"); and
- "Creator" (e.g. "Liu, Zhi-Rong").

There can be more than one creator as well as identifier within a record. At least one identifier is hotlinked and provides access to a brief record within the source repository. 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.")

In addition to basic bibliographic data, links are available that allow a user to "View Complete Metadata Record" or to "Add to a Book Bag" are located at the base of each record. 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) (see Figure 4). From the service search page, the user can access his or her respective "Book Bag" collection as well their session "Search History", which allows the user to re-execute ("Redo") or modify ("Modify Search") of any previous search strategy.

The full record ("View Complete Metadata Record") for an item contains:

- "Identifier" (e.g. "oai:arXiv.org:cond-mat/0210325");

- "Datestamp" (e.g. "2003-02-5");
- "SetSpec" (e.g. "physics:cond-mat");
- "Title";
- "Creator";
- "Subject" (if available) (e.g. "Materials Science");
- "Description" (Abstract);
- "Date" (e.g. "2002-11-19");
- "Type" (e.g. "text"); and
- Secondary "Identifier" (e.g. "<http://arxiv.org/abs/cond-mat/0211397>").

A full record can include not only multiple creators and identifiers, but multiple descriptions as well. Search terms in the full record, as well as the brief record, are bolded to assist in scanning their context.

The total number of records harvested from each constituent repository for the Open Archives Initiative Information in Engineering, Computer Science, and Physics service are available ("Check the Status of Latest OAI Harvest") ([g118.grainger.uiuc.edu/engroai/LastHarvest.asp](http://g118.grainger.uiuc.edu/engroai/LastHarvest.asp)). There are plans to create a secondary provider site to allow other service providers to harvest from the service and to incorporate metadata and link to other relevant repositories.

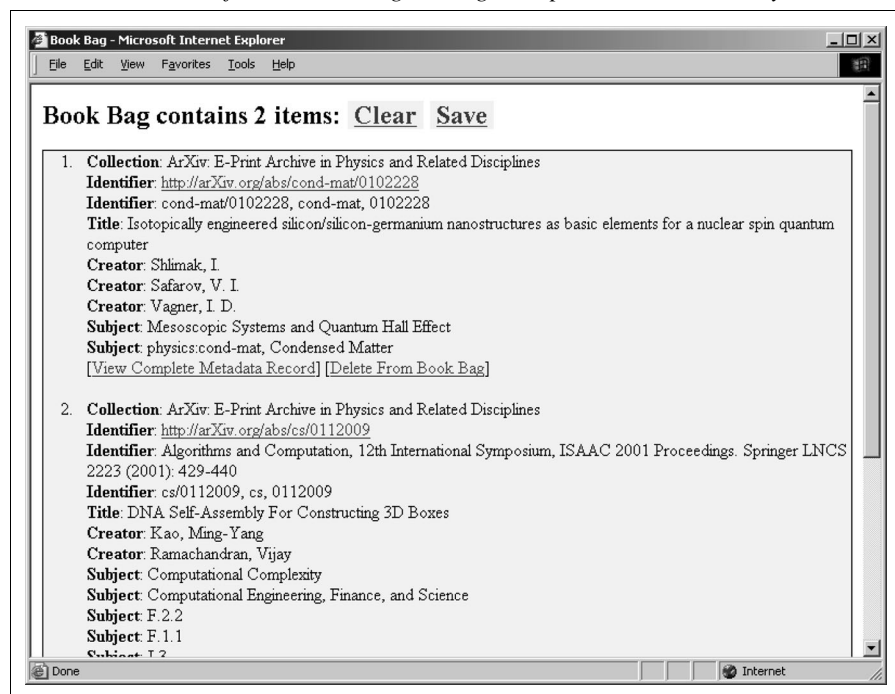
## SAIL-eprints (Search, Alert, Impact and Link)

Formally established in April 2003, SAIL-eprints ([eprints.bo.cnr.it](http://eprints.bo.cnr.it)) 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]." 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

**Figure 4**

Records for two sample documents placed in a “Book Bag” collection of the Open Archives Initiative Information in Engineering, Computer Science and Physics service



records harvested from 28 distinct data providers (eprints.bo.cnr.it/cgi-bin/info.pl) that include:

- arXiv.org.
- BioMed Central (www.biomedcentral.com).
- CNR ISOF (Istituto per la Sintesi Organica e la Fotoreattività) Eprints Server (isof-eprints.bo.cnr.it)
- Digital Archives and Library (Virginia Polytechnic Institute and State University) (scholar.lib.vt.edu)
- DSpace<sup>TM</sup> at MIT (libraries.mit.edu/dspace-mit/)
- Indian Institute of Science (eprint-s.iisc.ernet.in)
- Multidisciplinary Theses Server (Centre pour Communication Scientifique directe) (tel.ccsd.cnrs.fr)
- Organic Eprints (orprints.org)
- PhysNet Document Server (www.physnet.net)
- University of Melbourne ePrints Repository (UMER) (eprints.unimelb.edu.au)

Owners of OAI-compliant repositories who wish to be harvested by SAIL-eprints are provided with a form (eprints.bo.cnr.it/cgi-bin/info.pl) by which they can register their sites.

SAIL-eprints offers two user-friendly 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 20 records (“40”, “60”, “80”, “100”) by selecting the desired number from a pull-down menu. After execution, a listing of repositories that contain matching results is presented on the left side of the screen. To display the results from a given repository, the user clicks on the repository name (e.g. ‘arXiv’), which then subsequently 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 in which the search term(s) occur.

From within the brief record display listing, the user can mark select records or all records on a page. Marked results can be “exported” in an EndNote or HTML format; the user can also export all results without marking individual entries or pages (“Export ALL in ...”).

When exported in the HTML format, the selected results will be concatenated. Records in the HTML format include only basic data and information about each item, (title, subject (when available), author(s), and abstract (when available)). Each brief record, however, also includes an embedded hotlinked to a detailed record (“Show Details”) above the title of each entry in this format.

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)”) (see Figure 5). More than one repository may 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” grouping or “Institutions” grouping instead.

After choosing one or more individual repositories, a discipline category, or institutional repository, the user proceeds to the second step in the advanced search process – entry of the name of a specific author, and/or title and/or abstract keywords in one of three fields with associated pull-down menus for these field 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 for a maximum of 20 matching items per page as a default, although, as noted the user can increase the number. In either a “Simple Search” or an “Advanced Search,” the user can use special characters (“\*” or “%”) alone or in combination to search for word variants (e.g. hous\* matches with “house”, “houses”; m%n matches with “man”; “men”; lin%e\* matches with “linked”, “linger”).

**Figure 5**

In an "Advanced Search" in SAIL-eprints, the user can search a specific repository (left frame), repositories categorized within a specific discipline (middle frame), or the repositories associated with a particular institution or organization (right frame)



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 (eprints.bo.cnr.it/cgi-bin/show\_stat.pl). SAIL-eprints also offers an e-mail alerting service to registered users.

#### ACKNOWLEDGEMENT

The author is most grateful to the following individuals for granting permission to use screen images from their respective projects: Figure 1 – Michael L. Nelson, Old Dominion University; Figure 2 – Tim Brody,

University of Southampton, UK; Figure 3 – François Schiettecatte, FS Consulting, Inc.; Figure 4 – William H. Mischo, University of Illinois at Urbana-Champaign; Figure 5 – Silvana Mangiaracina, Biblioteca di Area Della di Bologna, Consiglio Nazionale delle Ricerche, Italy.

#### NOTE

- 1 This is the first of three eProfiles on select Open Archives Initiative service providers. The second in the series will focus on those in the social sciences and humanities, while the third will review those that provide access to a broad array of subjects and disciplines.

#### REFERENCES

Brody, T. (2002), "Citebase search: citation-impact ranking search service over e-print archives", PowerPoint presentation given at Academic Libraries of Open and Continuous Access, the 11th Pan-Hellenic Conference of Academic Libraries, Library of the Technological Educational Institute, Larissa, Greece, November 7, available at: [www.ecs.soton.ac.uk/~tdb01r/presentations/Greece%20-%202002-11-07.ppt](http://www.ecs.soton.ac.uk/~tdb01r/presentations/Greece%20-%202002-11-07.ppt) (accessed 13 September 2003).

Hitchcock, S., Bergmark, D., Brody, T., Gutteridge, C., Carr, L., Hall, W., Lagoze, C. and Harnad, S. (2002), "Open citation linking: the way forward", *D-Lib Magazine*, Vol. 8 No. 10, October, available at: [www.dlib.org/dlib/october02/hitchcock/10hitchcock.html](http://www.dlib.org/dlib/october02/hitchcock/10hitchcock.html) (accessed 13 September 2003).

Lagoze, C. and Van de Sompel, H. (2003), "The making of the Open Archives Initiative protocol for metadata harvesting", *Library Hi Tech*, Vol. 21 No. 2, pp. 118-28, available at: [www.cs.cornell.edu/lagoze/papers/The%20Making%20of%20the%20Open%20Archives%20Initiative.pdf](http://www.cs.cornell.edu/lagoze/papers/The%20Making%20of%20the%20Open%20Archives%20Initiative.pdf) (accessed 13 September).

Liu, X., Maly, K., Zubair, M. and Nelson, M.L. (2001), "Arc – an OAI service provider for Digital Library Federation", *D-Lib Magazine*, Vol. 7 No. 4, April, available at: [www.dlib.org/dlib/april01/liu/04liu.html](http://www.dlib.org/dlib/april01/liu/04liu.html) (accessed 14 September 2003).

Lynch, C.A. (2001), "Metadata harvesting and the Open Archives Initiative", *ARL Bimonthly Report*, No. 217, August, pp. 1-9, available at: [www.arl.org/newsltr/217/mhp.html](http://www.arl.org/newsltr/217/mhp.html) (accessed 13 September 2003).

**Gerry McKiernan** ([gerrymck@iastate.edu](mailto:gerrymck@iastate.edu)) is a Science and Technology Librarian and Bibliographer, Iowa State University Library, Ames, Iowa, USA.