

Open archives initiative service providers. Part III: general

This is the third and last in a series that describes Open Archives Initiative service providers. It profiles not only services that offer access to a variety of resources, but also recently initiated projects. The first in the series reviewed select services that offer access to a range of science and technology repositories, while the second focused on those that provide access to social science and humanities OAI-compliant resources.

CILEA Open Archives Platform

Established in 1974, the Consorzio Interuniversitario Lombardo per la Elaborazione Automatica (CILEA) is a consortium of universities in the Lombardy region of north-central Italy based in Segrate, an eastern suburb of Milan. It provides information and communication technology services on behalf of its members, other universities, public organizations, and businesses, as well as “professional advice for both the planning and the dissemination of advanced technologies in the fields of high performance computing, networking services and informatics.” Supercomputing, networking, library automation and digital libraries, database development, and bioinformatics are among the major activities of CILEA (www.cilea.it/redazione/struttura/cilea.pdf). In May

2003, the consortium launched an 18-month initiative to promote the establishment of institutional archives, and the creation of the CILEA Open Archives Platform (www.cilea.it/servizi/g/aepic/OA/doc/OA_platform_project_plan_eng_1_3.pdf), a project that aims to develop an Italian national platform providing a central access point to research papers collected by Italian Open Archive repositories”. In order to establish a “critical mass” of quality data from Open Archives”, CILEA will be involved in the “design, implementation, and running of independent Data Providers, whether institutional, disciplinary, or individual . . .” (Mornati, 2003a, b).

The “. . . national platform will also provide advanced data search and retrieval tools, data aggregation, time stamping, protection against plagiarism, and long-term digital preservation. In addition, the availability of usage statistics and citation linking will provide measurement tools for research impact, that are useful . . . [for] research management, for assessment and funding purposes . . .” The platform architecture will offer a common search and retrieval interface, disciplinary access by subject, and complementary services to different and independent repositories. The platform architecture is structured into two main objects: a collection of service providers and a portal. To sum up, a cluster of eprints archives, called data providers, which contain documents and associated metadata deposited by users at each institution, will provide the raw data, and a cluster of additional services, called service providers, will add value to these data, and expose them to users through a portal, that will allow results customisation.”

The metadata and full text (where available) of research papers will be harvested from select OAI-compliant archives, most notably those with coverage of the Italian research

literature, and subsequently enhanced by external Web services. Among the notable processing activities, features, and functionalities are:

- categorization of metadata by broad subject classification based on the Italian official academic research areas classification scheme (Elenco dei settori scientifico-disciplinari (www.murist.it/atti/2000/alladm001004_01.htm);
- central cacheing and indexing functionalities;
- citations parsing in the document text and extraction of machine-readable OpenURLs (www.niso.org/committees/committee_ax.html);
- development of cross-walks for data conversion, enabling the import and export to/from databases adopting different metadata standards; and
- full-text search capabilities.

Portal

In addition to support of institutional repositories and the creation of a data service provider, project personnel expect to create a “portal” that will “provide a single point of access to the centralised data, and to resources relevant to open archive issues, together with a range of additional services, such as a user profile management system, tailored e-mail alerting services, aggregated newsfeed . . .” among other services. The portal will also host a list of Italian open archives initiatives, a test-bed of tools for developers, a virtual reference desk for participants, and an online forum for discussion of national policy issues relating to open archives.

Susanna Mornati (mornati@cilea.it) serves as the Project Leader of the CILEA Open Archives Platform. The platform architecture for the CILEA Open Archives Platform is based on the ePrints UK model (see below) and suggestions made by Antonella de Robbio, Librarian, Biblioteca del

The author is most grateful to the following individuals for granting permission to use screen images from their respective projects: Figures 1 and 2: Marieke Guy, UKOLN, University of Bath, UK; Figure 3: Hussein Suleman, University of Cape Town; Figure 4: Kat Hagedorn, OAIster Project, University of Michigan; and Figure 5: John Willinsky, Public Knowledge Project, University of British Columbia, Canada.

Seminario Matematico, Università degli Studi di Padova, Italy (De Robbio, 2003). The project emerged from a feasibility study on scientific electronic publishing released in August 2002 (Comba, 2002) and a conference on scientific communication organized by CILEA and the Università degli Studi di Milano held in May 2003 (*Comunicazione Scientifica ed Editoria Elettronica*, 2003).

ePrints UK

Launched in July 2002, ePrints UK (www.rdn.ac.uk/projects/eprints-uk/) is a two-year project that seeks to “develop a national service provider repository of e-print records . . . derived by harvesting metadata from institutional and subject-based e-prints archives using the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). The project also aims to provide access to these institutional assets through the eight Resource Discovery Network (RDN) faculty level hubs . . . and the Education Portal. . .” (Martin, 2003) (see Figure 1). Funded by the Joint Information Systems Committee (JISC) (www.jisc.ac.uk) under its “Focus on Access to Institutional Resources (FAIR) Programme,” the broader goal of ePrints UK is to create “a series of national, discipline-focused services through which the higher and further

education community can access the collective output of e-print papers . . . particularly those provided by UK universities and colleges” (www.rdn.ac.uk/projects/eprints-uk/).

ePrints UK is a collaborative project of UKOLN (www.ukoln.ac.uk), the Resource Discovery Network (RDN) (www.rdn.ac.uk) based at King’s College London, the OCLC Office of Research (www.oclc.org/research/projects/mswitch/epuk.htm) and the University of Southampton, OpCit Project (<http://opcit.eprints.org>). The RDN is “a collaboration of over 70 educational and research organisations, including the Natural History Museum, London, UK, and the British Library. In contrast with search engines, the RDN gathers resources . . . [that] are carefully selected, indexed and described by specialists . . . [from its] partner institutions” (www.rdn.ac.uk/about/). “RDN is a co-operative network consisting of a central organisation, the Resource Discovery Network Centre (RDNC) and a number of independent service providers called ‘hubs.’”

There are currently eight hubs (www.rdn.ac.uk/about/#HUBS) covering a range of subjects and disciplines:

- ALTIS (hospitality, leisure, sport and tourism) (www.altis.ac.uk).
- Artifact (arts and creative industries) (www.artifact.ac.uk).

- BIOME (health and life sciences) (<http://biome.ac.uk>).
- EEVL (engineering, mathematics and computing) (www.eevl.ac.uk).
- GEsorce (geography and environment) (www.gesorce.ac.uk/home.html).
- Humbul (humanities) (www.humbul.ac.uk).
- PSIGate (physical sciences) (www.psigate.ac.uk).
- SOSIG (social sciences, business and law) (www.sosig.ac.uk).

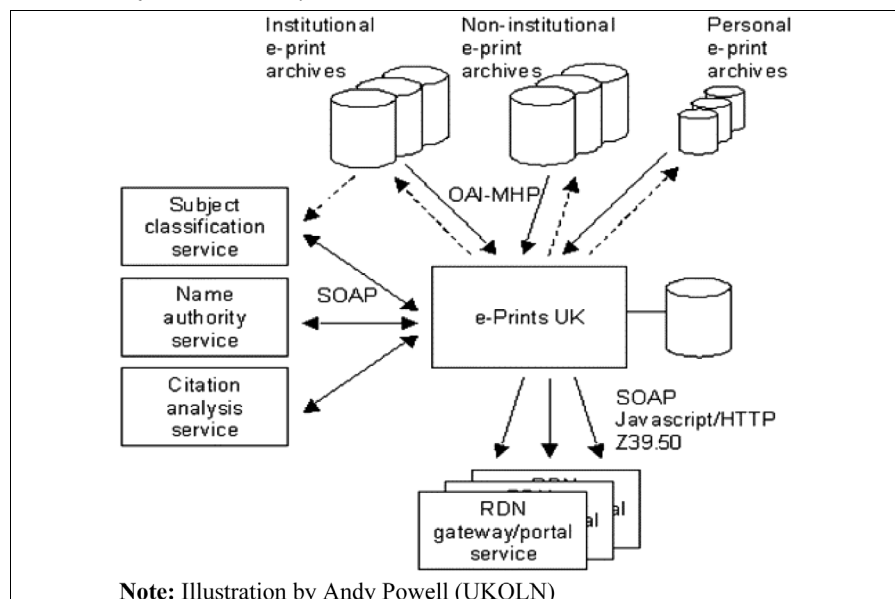
All hubs are collectively accessible from the RDN homepage (www.rdn.ac.uk).

The ePrints UK project deliverables (www.rdn.ac.uk/projects/eprints-uk/docs/proposal/) include:

- *WP1 – Central Database of UK e-Print Archive Metadata Records.* This work package will develop, install and configure the OAI harvesting software that will gather metadata records from e-print archives into a central database to hold the metadata records.
- *WP2 – Interfaces to Central Metadata Database.* This work package will develop a number of interfaces to the central database developed during WP1. These will include a central Web interface, a Simple Object Access Protocol (SOAP) interface and a Z39.50 interface and the integration of the SOAP interfaces with eight existing RDN hub Web sites.
- *WP3 – Name Authority and Subject Classification Web Services.* This work package will develop the name authority and subject classification Web services and integrate them with the database developed as part of WP1. This work will involve the design of appropriate SOAP interfaces to the Web services and associated software development. This work will be undertaken by UKOLN, in close collaboration with technical staff at the OCLC Office of Research.

The enhanced metadata records returned by the subject classification Web service will include Dewey classmarks and possibly other subject data as well. In combination with other metadata, these classmarks will be used to

Figure 1
Schematic of ePrints UK system architecture



partition the e-Prints UK database into subject-focused sub-sets, for presentation out through the RDN hub interfaces.

- *WP4 – Citation Analysis Web Service.* This work package will develop the citation analysis Web service. This work will be based on existing software developed by the Open Citation project at the University of Southampton.
- *WP5 – Supporting Studies.* This work package will carry out a number of supporting studies throughout the project (notably reports on impact assessment, collection development, business and intellectual property, and research assessment).
- *WP6 – Evaluation.* This work package will evaluate e-Prints UK in various ways. The project will undertake an evaluation of the subject classification Web service; taking a statistical survey approach and analysing the Dewey classmarks assigned to 400 metadata records.
- *WP7 – Publicity, Promotion and Events.* This work package will deliver ... [several] regional workshops that will promote the establishment of e-print archives by UK universities and colleges and provide a forum for the dissemination and discussion about best practice.

It is important to note that not only will metadata from target repositories be harvested in the ePrints UK project, but also the full text of associated documents, when available (see Figure 2).

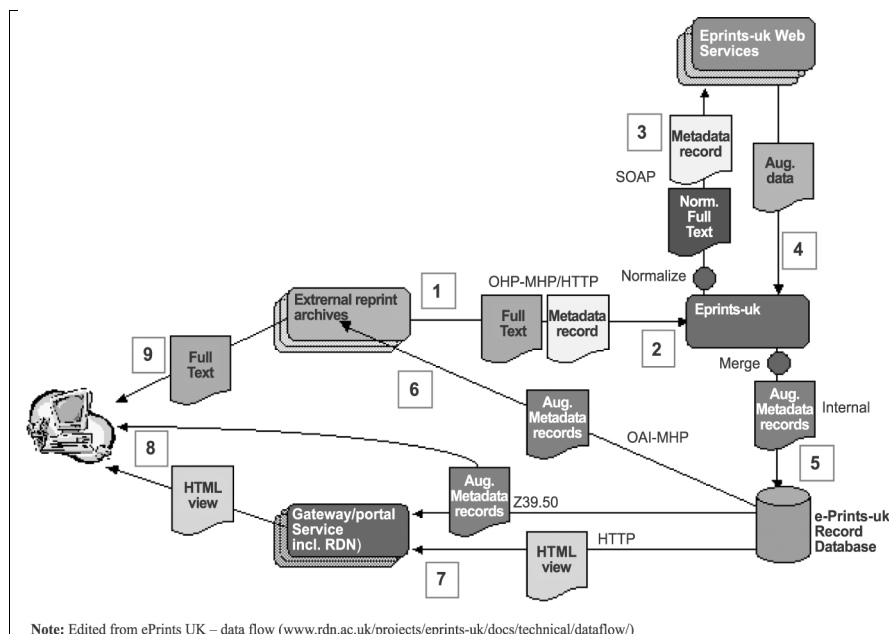
The success of the ePrints UK project is significantly dependent on the widespread availability of significant and suitable content. Currently there are a number of UK institutional repositories (www.rdn.ac.uk/projects/eprints-uk/repositories/) (see Table I), but few subject-based e-print repositories located in the UK (see Table II) (Day, 2003b).

As of May 2003, nearly 2,500 records had been harvested from select UK repositories for the ePrints UK collection (Day, 2003a).

ePrints UK has been implemented using the Arc software (oiarc.sourceforge.net) developed by Old Dominion University Digital Library

Figure 2

Schematic of ePrints UK project planned data flow with annotations describing associated activities



Group (dlib.cs.odu.edu), while indexing is performed using the Cheshire II system software (cheshire.lib.berkeley.edu) (Cliff, 2003). Marieke Guy (m.guy@ukoln.ac.uk) serves as the ePrints UK Project Manager, while Michael Day (m.day@ukoln.ac.uk) serves as the Research Officer (Metadata), and Pete Cliff (p.d.cliff@ukoln.ac.uk) is the Systems Developer.

NDLTD Union Catalog (Electronic Thesis/Dissertation OAI Union Catalog based at OCLC)

“The Networked Digital Library of Theses and Dissertations (NDLTD) (www.ndltd.org) is a loose federation of member institutions and organizations that publish *Electronic Theses and Dissertations (ETDs)*. Since its 1996 inception, NDLTD has grown to include over 130 individual member institutions and many consortia from countries around the world ...” [The] NDLTD ... supports and encourages the production and archiving of *ETDs*. While many current NDLTD member institutions and consortia have individual collections accessible online, there has until recently been no single mechanism to aggregate all *ETDs* to provide NDLTD-wide services ... With the emergence of the Open Archives Initiative (OAI) that has

changed” (Suleman and Fox, 2003, pp. 219-20).

“While not a primary goal, it has long been a desire of NDLTD to unite the various member sites into a single collection to support the researcher or student seeking *ETDs*. ... [T]o support multiple service providers ... [a] union archive was created with the purpose of collecting metadata from remote sites and republishing it as a single collection.” Among the advantages of a central merged collection of metadata are the development of multiple services at a central site without the need to harvest metadata multiple times, and the development of experimental services that “do not negatively impact the production servers of individual members by repeatedly requesting transfers of the same metadata.” “The union archive is designed to function as both a provider of services (harvester) and a provider of data. It harvests metadata from the remote sites, stores them in an internal database and then republishes these metadata through its own OAI data provider” (Suleman and Fox, 2003, p. 221).

NDLTD Union Catalog based at OCLC (<http://rocky.dlib.vt.edu/~etdunion/cgi-bin/OCLCUnion/UI/index.pl>) is “purely an experimental service” that is “built by harvesting metadata from

Table I*A listing of institutional repositories in the UK (as of September 2003)*

Armagh Observatory Preprints/Reprints Series 2003 http://star.arm.ac.uk/preprints/	28
University of Bath: Mathematics Group: Preprints for the Mathematics Group www.maths.bath.ac.uk/MATHEMATICS/preprints.html	186
University of Bath: eprints@bath http://eprints.bath.ac.uk/	6
Bristol Centre for Applied Non-linear Mathematics www.enm.bris.ac.uk/anm/publications.html	134
University of Cambridge: Isaac Newton Institute for Mathematical Sciences Preprints series www.newton.cam.ac.uk/preprints.html	41
University of Cambridge: Statistical Laboratory: MCMC Preprints www.statslab.cam.ac.uk/~mcmc/	469
University of Cardiff: School of Mathematics: Hoyle-Wickramasinghe Preprint Series on the Internet www.cf.ac.uk/math/wickramasinghe/contents.html	11
University of Durham: Geometry and Arithmetic Preprints http://fourier.dur.ac.uk:8000/pure/preprint.html#viewnote	82
University of Edinburgh: Theoretical and Applied Linguistics http://archive.ling.ed.ac.uk/	90
University of Glasgow: ePrints @ Glasgow http://eprints.lib.gla.ac.uk/	9
University of Glasgow: ERPAePRINTS http://daedalus.lib.gla.ac.uk/	36
Lancaster University: Department of Mathematics and Statistics Spatial and Computational Statistics Network: Network preprints www.maths.lancs.ac.uk/dept/stats/essn/preprints.html	81
University of Leicester: White Dwarf Group Preprint Server www.star.le.ac.uk/wd/preprint.html	16
University of Leicester: X-ray Astronomy Group http://ledas-www.star.le.ac.uk/Preprint/	172
Loughborough University: Department of Mathematical Sciences: Preprint Archive www.lboro.ac.uk/departments/ma/preprints/	194
University of Manchester Institute of Science and Technology (UMIST): Department of Physics www.umist.ac.uk/departments/physics/research/preprint.htm	114
University of Nottingham: Nottingham ePrints http://eprints.nottingham.ac.uk/	46
Open University ePrints http://libeprints.open.ac.uk/	3
University of Oxford: Mathematical Institute RAND-APX Thematic Network: Preprint series www.maths.ox.ac.uk/rand-apx/preprint.html	15
University of Southampton: Department of Electronics and Computer Science http://eprints.ecs.soton.ac.uk/	7,158
University of St Andrews: Astronomy Group Preprint Server http://star-www.st-and.ac.uk/astronomy/preprints.html	57
University of Wales, Bangor: School of Informatics: Maths Preprints www.informatics.bangor.ac.uk/public/mathematics/research/preprints/preprint.html	66

open archives of electronic theses and dissertations” [from the NDLTD]. “The underlying technology is based on layered open archives with data being harvested from source archives and then stored in a Union Catalog. This Union Catalog is then front-ended with a search engine for demonstration purposes . . .” (<http://rocky.dlib.vt.edu/~etdunion/about.html>).

As of mid-December 2003, the following harvested sites had more than 500 records in the Union Catalog:

- California Institute of Technology (Caltech) (772).
- CCDS theses-EN-ligne (Centre pour la Communication Scientifique Directe) (715).
- Hong Kong University Theses (9,970).
- Humboldt-Universität zu Berlin (1,234).
- Louisiana State University (646).
- Digitale Hochschulschriften der LMU (Ludwig-Maximilians-Universität München) (940).
- Massachusetts Institute of Technology (MIT) (8,671).
- National Sun Yet-Sen University (Taiwan) (4,105).
- North Carolina State University (1,468).
- Universitat Autònoma de Barcelona (596).
- University of Tennessee Library (3,751).
- Uppsala Universitet (970).
- Virginia Polytechnic Institute and State University (Virginia Tech) (4,675).

Other sites with sizeable numbers of records in the NDLTD Union Catalog include East Tennessee State University (332), Georgia Institute of Technology (Georgia Tech) (468), PhysNet (319), Universitat Politècnica de Catalunya (226), and the Wirtschaftsuniversität Wien (351). As of mid-December 2003, more than 50,200 records were harvested from these and other sites for the NDLTD Union Catalog (<http://alcm.oclc.org/ndltd/servlet/OAIHandler?verb=ListSets>).

Search and browse

The NDLTD Union Catalog can be searched or browsed from a single interface. Users can perform a “Quick

Table II*A list of subject repositories based in the UK in order by the number of records*

Subject-based repository	Number of records (March 2003)	Web address
CogPrints: (Cognitive Science Eprint Archive)	1,709	cogprints.ecs.soton.ac.uk
Psychology [e-journal]	720	psycprints.ecs.soton.ac.uk
Formations Media Studies Archive	21	formations2.ulst.ac.uk

search” free text search or a “Quick browse.” In the quick search mode, all record fields are searched; alternatively, users can search a specific field by using a correct Dublin Core field name and colon before the search term or name (e.g. “contributor:fox”) (<http://oai.dlib.vt.edu/~etdunion/oclchtml/search.html>). In the browse mode, users can browse the catalog by institutional source and sort the results by institutional name (“university”), “year”, “title”, “author”, or accept the “default”, a display of all records in no defined or discernible order. In addition, the catalog can be browsed by year. The NDLTD Union Catalog not only includes records and access to the full text of dissertations not only from the twentieth and twenty-first centuries, but from the nineteenth and eighteenth centuries as well.

Results from a search (or browse) are displayed in a numbered list. Each record includes the title of the dissertation/thesis, author name, abstract, date, host institution, and hotlinks to “more info”, “go to document”, and “find similar documents” (see Figure 3). “More info” provides a link to a “full metadata record” that provides comprehensive data and information about the work, including such Dublin Core fields as “title”, “creator”, “subject”, “description”, “publisher” (home institution), “contributor”, “date”, publication “type” (e.g. “thesis”), “format”, “identifier”, “source”, “language”, “rights”, and “relation”, if appropriate. “Go to document” provides access to the record for the item from the institutional source server from which the user can retrieve an item’s full text, if available, while the “find similar documents” retrieves theses/dissertations similar in some manner to the source item. The go to document and find similar document features are also available at the bottom of a full metadata record. Within a browse

results display, users can sort results by the general sort options (i.e. “university”), “year”, “title”, “author”, or accept the “default” option (see Figure 3).

The NDLTD Union Catalog based at OCLC is “powered by OCLC’s OAI Cat Repository Framework”, a Java servlet Web application open source project that provides an OAI-PMH v2.0 repository framework (<http://alcme.oclc.org/oaicat/index.html>). In addition to the NDLTD OCLC-based Union Catalog, an identical, but less comprehensive version is also available (<http://oai.dlib.vt.edu/~etdunion/cgi-bin/index.pl>). An experimental advanced search interface is also being developed by the OCLC office of research that allows users to search the NDLTD Union Catalog by several key fields (i.e. “title word”, “creator”,

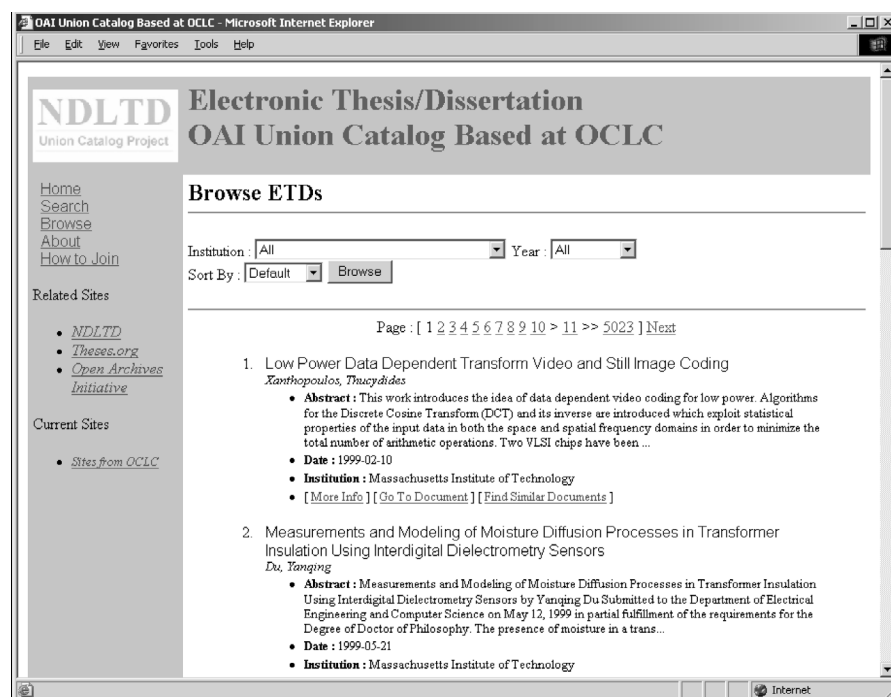
“contributor”, “abstract”) singularly, or in combination (<http://alcme.oclc.org/ndltd/SearchbySru.html>).

OAister

OAister (www.oaister.org) is a project of the University of Michigan Digital Library Production Services that seeks “... to create a collection of freely available, difficult-to-access, academically-oriented digital resources ... that are easily searchable by anyone” (see also Hagedorn (2003)). OAister is global in scope, providing cross-repository searching of metadata describing publicly-available digital objects, such as electronic books, online journals, audio files (e.g. wav, mp3), images (e.g. tiff, gif), video (e.g. mpeg, QuickTime), and reference texts (e.g. dictionaries, directories), with a focus on providing ready access to digital resources that reside in the “Hidden Web”. Formally launched in June 2002 with a seed collection of 275,000 records harvested from more than 50 institutional sites (McKiernan, 2003, p. 64), OAister contained more than 2.2 million records harvested from more than 240 sources in early December 2003 (www.oaister.org).

Figure 3

Screen print showing the brief record format in NDLTD Union Catalog based at OCLC



Users can search all record fields concurrently, or limit a search to keyword, title, author, and/or subject. While queries may be restricted to resource type (i.e. text, image, audio, video), other search operations are few: two (or more) terms are searched as an implied Boolean “AND” and users can truncate a term using an asterisk (*) to search on term variations.

Results can be pre-sorted by “title” (default), “author/creator”, “date ascending” or “date descending”. In addition, users have the options of sorting results by “hit frequency” or “weighted hit frequency”. The former “counts the number of instances of the words and phrases ... entered and orders them from highest count to lowest count”, while the latter operates in a similar manner as “hit frequency” but “gives more weight to instances of words and phrases in certain fields ... and will display a score ... ” in the research results. Sorting is not currently available for sets with more than 1,000 results (www.oaister.org/oaister/help.html).

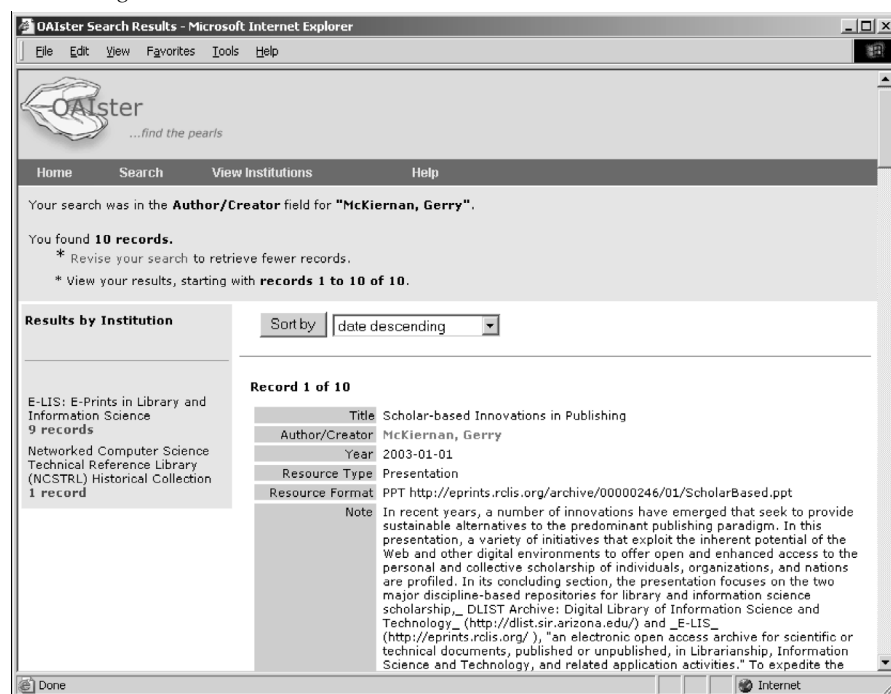
Records retrieved from a search in OAIster can include some or all of the following fields:

- Title.
- Author/creator.
- Contributor.
- Publisher.
- Year.
- Resource type.
- Resource format.
- Language.
- Note.
- Subject.
- URL.
- Rights.
- Institution.

Displayed to the left of the record(s) of initial search results (“results by institution”) are the name(s) of the source repositories containing relevant records (e.g. “E-LIS: Prints in Library and Information Science”; “Networked Computer Science Technical Reference Library (NCSTRL) Historical Collection”) and the number of records from each (e.g. “9”; “1”) (see Figure 4). Clicking the hotlinked number will retrieve and display only those records from the associated

Figure 4

Screen print of a partial record in OAIster. Results have been sorted in “date descending” order



repository. Within search results, records can be re-sorted (“sort by”) by selecting from the identical options offered in the search interface. To enable users to readily scan search results, query terms and phrases are highlighted in a bold red-maroon color.

Collections

An alphabetical browsable, annotated list of harvested collections that provides the formal name of the collection (e.g. “behavioral and brain sciences (BBS) EPrints online archive”), the number of records harvested from the source (e.g. “151”), a hotlink to the collection or the homepage of the originating organization, and brief description of the nature and scope of the select collection, is also available. Common and uncommon collections within OAIster with more than 1,000 harvested records include:

- American Numismatic Society (ANS) (www.annumsoc.org) (3,127).
- Bibliothekservice-Zentrum Baden-Württemberg, Germany, Virtueller Medienserver (www.bsz-bw.de/di-glib/medserv/) (24,987).

- conoZe: Intelligere ut Credas, Credere ut Intelligas (www.conoze.com) (1,771).
- Digital Library for Earth Systems Education (DLESE) (www.dlese.org) (5,100).
- Indiana University Digital Library Program (<http://dlib.indiana.edu>) (2,729).
- Internet Archive (www.archive.org) (25,764).
- Project Euclid (<http://projecteuclid.org>) (6,155).
- PubMed Central (PMC) (www.pubmedcentral.gov) (126,301).
- RePEc (search papers in economics) (www.repec.org) (53,003).
- SciELO (Scientific Electronic Library Online) (www.scielo.br) (24,523).

The OAIster project makes use of the Open Archives Initiative Metadata Harvesting Protocol (OAI-MHP) (www.openarchives.org/OAI/openarchivesprotocol.html) as the framework for retrieving digital resources. The University of Illinois at Urbana-Champaign (UIUC) was contracted to develop and provide the “harvester” mechanism that collects, aggregates, and updates the metadata from cooperating compliant repositories, while the

University of Michigan constructed the indexing and presentation tools for organizing the harvested data, and developed and provided an associated search engine service (<http://oai.granger.uiuc.edu/michigan.htm>) (see also Hagedorn (2003)).

Planned system enhancements for OAIster include Boolean searching and use of proximity operators; downloading and e-mailing of records; sorting by proximity and institution frequency; browsing by broad topical categories; and removal of duplicate records from preliminary search results.

Kat Hagedorn (khage@umich.edu) is currently the OAIster Librarian and served as the OAIster project manager and librarian during the grant period, while Michael Burek has served as the OAIster programmer. Project personnel are willing to assist sites in converting their collection metadata to enable harvest and incorporation within OAIster (www.oaister.org/o/oaister/dataproviders.html). The OAIster project is one of seven OAI Metadata Harvesting Protocol projects funded by the Andrew W. Mellon Foundation in summer 2001. The Research Libraries Group (RLG), Emory University, the Woodrow Wilson International Center for Scholar, in Washington, DC, the University of Virginia, and the Southeastern Library Network, Inc. (SOLINET) were the other institutions that received support (Waters, 2001).

Public knowledge project Open Archives Harvester

“The Public Knowledge Project (PKP) is a federally-funded research initiative located at the University of British Columbia in Vancouver, Canada, ‘dedicated to exploring whether and how new technologies can be used to improve the professional and public value of scholarly research’” (www.pkp.ubc.ca). As part of its efforts, the PKP “is currently developing and testing a number of online research management systems to improve the scholarly and public quality of academic research. These systems are designed not to only assist in the management and publishing of scholarly work, but also to improve the indexing of research in online environments and create a richer context of connections for any given

study, connections both within the scholarly literature and to the larger world of related online information” (www.pkp.ubc.ca/about/what.html).

Currently the PKP online systems include:

- Education Research Index.
- electronic theses and dissertations.
- open journal systems.
- Open Archives Harvester.
- open conference systems; and
- research support tool.

“The PKP Open Archives Harvester is a free metadata indexing system developed by the PKP through its federally funded efforts to expand and improve access to research. The PKP OAI Harvester allows ... [one] to create a searchable index of the metadata from Open Archives Initiative-compliant archives ... The PKP OAI Harvester is currently compatible with versions 1.1 and 2.0 of the OAI Harvesting Protocol.” Users are able not only to download the harvester software (www.pkp.ubc.ca/pkp-harvester/download.html) but also to access an operating service provider – the Open Archives Harvester. As of mid-December 2003, the PKP Open Archives Harvester (www.pkp.ubc.ca/harvester/) contained nearly 6,600 records from 39 archives.

The common and uncommon harvested sites within the PKP Open Archives Harvester include:

- Centre for the Study of Historical Consciousness (www.cshc.ubc.ca).
- DRH 2003 (Digital Resources in the Humanities) (www.hcu.ox.ac.uk/ocs/).
- Digitale Hochschulschriften der Ludwig-Maximilians Universität München (<http://edoc.ub.uni-muenchen.de>).
- DLIST (Digital Library of Information Science and Technology) (<http://dlist.sir.arizona.edu>).
- DSpace@Erasmus (<http://ep.eur.nl>).
- *Electronic Journal of Probability* (www.math.washington.edu/~ej-pecp/).
- Organic Eprints (<http://orgprints.org>).
- Portraits of Literacy Conference (www.pkp.ubc.ca/literacyconference/)

- PsyDok: Volltextserver der Virtuellen Fachbibliothek Psychologie (<http://psydok.sulb.uni-saarland.de>.)
- Sammelpunkt (<http://sammelpunkt.philo.at:8080/>).

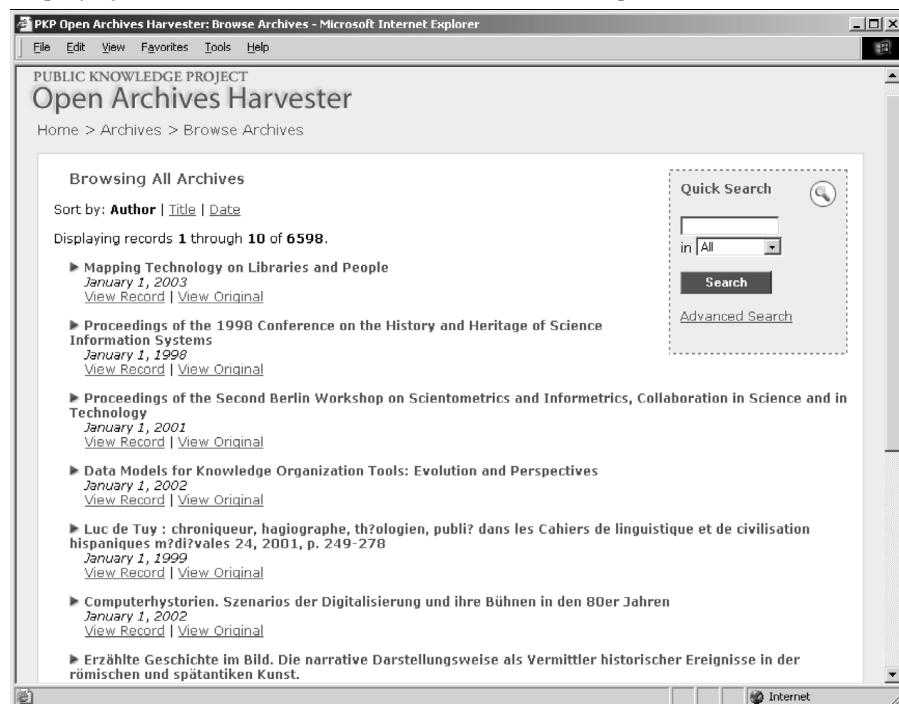
The Open Archives Harvester provides three general search or browse options: a basic search, “advanced search”, and a “browse archives”. In the basic search, the user can perform a free text search in any field in a record (“all”) or restrict the query to one of four fields: “author(s)”, “title”, “abstract”, “index terms” by selecting from a pull-down menu. Using the “advanced search” search, users can search all databases concurrently (“search databases”), perform a keyword search in a selected collection or all collections (“search all categories for”), or search in a variety of record categories or fields (e.g. “author(s)”, “title”, “abstract”, “date”, “index terms”, “language”, “sponsor(s)”, “type”). The Open Archives Harvester offers a number of uncommon, yet potentially useful, search field options, notably the ability to search by author affiliation, by “sponsor(s)” and “type” of document (i.e. “refereed articles”, “reviewed papers”, “dissertations”, “forums”, “research materials”, and “creative work”). Users can search or limit a query by “discipline(s)”, “subject(s)”, “approach/method”, and “coverage” (www.pkp.ubc.ca/harvester/about.php) from the “advanced search” page. To enhance the selection of appropriate search terms and phrases a link to the Library of Congress Classification Outline (<http://lcweb.loc.gov/catdir/cpsolcco/lcco.html>) is provided above this set of search options. Index terms assigned to individual documents can be viewed from within the complete record format and also serve as candidate search terms.

Search and browse results

After executing a basic or advanced search option, results are displayed in no discernible order as brief records that include the title of the work, the author, date, and links that allows users to “view record” or “view original” (see Figure 5). In the browse display, records are also displayed in a similar brief format, but exclude the author(s)

Figure 5

Display of results in an “all archives” browse in the Open Archive Harvester



name(s). Although author names are not provided, by default records are in order by “author”. The results can re-sort the results by title or date by selecting the option of interest.

The full record for an item includes all available data and information about the item from the source, notably title, source archive name, item Uniform Resource Locator (URL), author(s) name(s), date, abstract, index terms, publisher, contributors, source, language, relation, type, format, and copyright information, when provided in the original document; the archive name and item URL are hotlinked to their sources. From with the results, users can perform a “quick search” or link to the “advanced search” option.

Research support tool

Located above each full record display is a “research tool kit” rectangular menu appropriate to the item. The research support tool (RST) has been designed to enable users to readily “link to related research, Web sites, and databases.” This rich context of connections is intended to provide a stronger basis for interpreting studies and pursuing a greater understanding of the field, whatever the ... [user’s]

background or interests (www.pkp.ubc.ca/demos/rsttour/index.html). For example, the RST offers users such features, added-value functionalities, or access options as:

- (1) For this ... [item]:
 - view metadata;
 - capture cite;
 - printer-friendly (display record in its source format).
- (2) Context:
 - author bio;
 - define terms;
 - e-journals;
 - related theory;
 - related studies;
 - pay-per-view;
 - online forums;
 - instructional;
 - gov policies; and
 - media reports.
- (3) Action:
 - e-mail author;
 - e-mail others; and
 - add to portfolio.

(See also www.pkp.ubc.ca/demos/rsttour/index.html).

A Web-based form (www.pkp.ubc.ca/harvester/add.php) that enables institutions to nominate their repository to be harvested by the PKP Open Archives Harvester.

Powerful mechanism

As observed by Lynch (2001, p. 9), “the Open Archives Metadata Harvesting Protocol opens many new possibilities which are yet to be explored.”

As he further notes: for data-intensive scholarly communities in which data are widely distributed rather than centralized into a few key community databases, this interface may offer a new way to translate rather abstract investments in metadata standardization into tangible opportunities to contribute to operational systems for locating information resources ...

Researchers who want to explore new ways of organizing, presenting, or using these large data resources will now have a standardized way of extracting content without much disruption or cost to existing operational systems. This may be a powerful mechanism for enabling the development of new applications and services that have never before been possible (Lynch, 2001, p. 9).

The variety of OAI service providers profiled in this series clearly demonstrates that the Open Archives Metadata Harvesting Protocol has not only opened “many new possibilities” for access to institutional and organizational digital resources, it has also enabled the development of a broad range of “new applications and services” that provide a multiplicity of value-added features and functionalities that significantly enhance their use.

REFERENCES

- Cliff, P. (2003), *ePrints UK – Architecture, 1.032002*. Based on Project Proposal, available at: www.rdn.ac.uk/projects/eprints-uk/docs/technical/architecturev1.032003/ (accessed 16 December).
- Comba, V. (2002), *AEPIC Academic E-Publishing Infrastructures – CILEA: Progetto di Editoria Elettronica per la Ricerca e la Didattica*, available at: http://eprints.rclis.org/archive/00000066/01/

AEPIC-CO511.pdf (accessed 10 December 2003).

Comunicazione Scientifica ed Editoria Elettronica: La Parola agli Autori: L'Utente-Autore nel Circuito della Comunicazione Scientifica: Editoria Elettronica e Valutazione della Ricerca (2003), Milan, May 20, available at: www.cilea.it/convegni/convegnoeditoria/presentazione.html (accessed 10 December).

Day, M. (2003a), *ePrints UK Biannual Progress Report*, available at: www.rdn.ac.uk/projects/eprints-uk/docs/biannual/eprintsuk-biannual-report.doc (accessed 28 November).

Day, M. (2003b), *Prospects for Institutional E-Print Repositories in the United Kingdom*, available at: www.rdn.ac.uk/projects/eprints-uk/docs/studies/impact/ (accessed 29 November).

De Robbio, A. (2003), "Auto-archiviazione per la ricerca: problemi aperti e sviluppi futuri", paper presented at *Comunicazione Scientifica ed Editoria Elettronica: La Parola agli Autori: L'Utente-Autore nel Circuito della Comunicazione Scientifica: Editoria Elettronica e Valutazione della Ricerca*, Milan, May 20, available at: <http://eprints.rclis.org/archive/00000180/03/OAI-20maggio2003.pdf> (accessed 10 December).

Hagedorn, K. (2003), "OAIster: a 'no dead ends' OAI service provider", *Library Hi Tech*, Vol. 21 No. 2, pp. 170-81.

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 (accessed 13 December).

McKiernan, G. (2003), "News from the field", *Journal of Internet Cataloging*, Vol. 6 No. 1, pp. 55-71.

Martin, R. (2003), "ePrints UK: developing a national eprints archive", *Ariadne*, No. 35, April, available at: www.ariadne.ac.uk/issue35/martin/ (accessed 29 November).

Mornati, S. (2003a), *CILEA Open Archives Platform - Project Plan*, available at: www.cilea.it/servizi/g/aepic/OA/doc/OA_platform_project_plan_eng_1_3.pdf (accessed 10 December).

Mornati, S. (2003b), "Open archives in Italia: una piattaforma nazionale", paper presented at *Biblioteche Digitali per la Ricerca e la Didattica: Esperienze e Prospettive*, Università di Parma, Parma, November 22, available at: http://eprints.rclis.org/archive/00000519/02/parma_mornati.pdf (accessed 13 December).

Pinfield, S. (2003), "Open archives and UK institutions: an overview", *D-Lib Magazine*, Vol. 9 No. 3, March, available at: www.dlib.org/dlib/march03/pinfield/03pinfield.html (accessed 30 November).

Suleman, H. and Fox, E.A. (2003), "Leveraging OAI harvesting to disseminate

theses", *Library Hi Tech*, Vol. 21 No. 2, pp. 219-27, also available at: <http://pubs.cs.uct.ac.za/archive/00000018/01/lht2002ra.pdf> (accessed 29 November).

Waters, D.J. (2001), "The metadata harvesting initiative of the Mellon Foundation", *ARL Bimonthly Report*, No. 217, August, pp. 10-11, also available at: www.arl.org/newsltr/217/waters.html (accessed 30 November).

FURTHER READING

McMillan, G. (2002), *ETDs and Libraries*, White Paper commissioned by the OCLC Office of Research, available at: www.oclc.org/research/projects/etd/mcmillan_etds_and_libraries.pdf (accessed 30 November).

Wilkin, J., Hagedorn, K. and Burek, M. (2003), *Creating an Academic Hotbot: Final Report of the University of Michigan OAI Harvesting Project*, available at: www.kathagedorn.com/mellon-harvesting-final.doc (accessed 1 December).

Gerry McKiernan (gerrymck@iastate.edu) is a Science and Technology Librarian and Bibliographer, Iowa State University Library, Ames, IA, USA.