



b/ITe

VOLUME 22, ISSUE 2 SUPPLEMENT

JULY/AUGUST 2005

ISSN 1541-7980

Kristin Yiotis is the winner of the 2005 Information Technology Division Student Paper Award. She is a student in the Library and Information Science Program at San Jose State University. Kristin was presented her award at the Business Meeting of the IT Division at the Annual SLA Conference in Toronto, CA on June 1, 2005. After finishing her degree, Kristin plans to continue with the open archives movement and its effect on scholar publication.

The Open Archives Initiative and Eprints Repositories

Kristin Yiotis

San Jose State University School of Library and Information Science

Abstract

The Open Access Initiative has brought about a revolution in the way scholarly articles are published. This article discusses Harnad's subversive proposal for electronic publishing that instigated the Open Archives and Open Access Initiatives. The researcher discusses self-archiving in local institutional repositories, interoperability and OAI compliance, protocols for harvesting metadata, OAI data providers and services providers in connection with the refereed literature-liberation movement. The aim of this movement is to shift the control of scholarly communications away from commercial publishers by reclaiming control of ownership and copyright of the scholars' own work.

The Open Archives Initiative and Eprints Repositories

The Open Access Initiative, the initiative to publish scholarly communications on the Internet freely available to all users, has gained momentum in the past few months. Recently, the United States Congress approved the National Institute of Health (NIH) proposal for making research articles based on NIH funding available to the public free of charge via PubMed Central, NIH's digital archive, within six months after publication in a peer-reviewed journal (Suber, 2004 November). On November 28, 2004, NBC Nightly News featured Julia Blixrud, Assistant Executive Director, ARL, in a story on the NIH Enhanced Public Access proposal (MSNBC News, 2004).

The early history of the movement is recorded at an ARL Web site that includes the original 1994 article by Stevan Harnad. Harnad posted "A Subversive Proposal" to an Internet discussion list based at Virginia Polytechnic Institute. Harnad, Professor of Cognitive Science at Princeton, NJ and University of Southampton, UK, was for many years a researcher and editor of *Behavioral and Brain Sciences*, a journal published by Cambridge University Press. In 1990, he introduced *Psychology* (Psychoquy 2004), the first peer-reviewed scientific journal on the Internet, and in 1997, the Cognitive Sciences Eprints Archive (CogPrints, 2004). In 1998, he started the American Scientist Open Access Forum, a high volume discussion list concerned with Open Access and Open Archives.

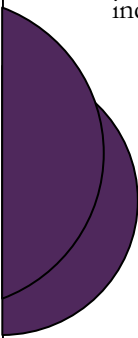
Harnad's 1994 proposal was the inspiration for the Open Archives Initiative. He proposed that all scholars publish their preprints of unpublished, unrefereed, original work on a globally accessible, local FTP archive, freely available to all other scholars. Once a work gets published everyone will substitute the published work for the preprint, his point being that scholars need not withdraw preprints from public viewing after refereed versions are accepted for paper publication. Once this becomes common, journal publishers will then be forced to restructure their costs for the electronic only versions to be truer to actual costs, which he estimated to be 25% less than the paper page costs. He also suggested that publishing costs be built in to the cost of research and be paid upfront in advanced rather than by the end user. If the current publishers don't restructure, then a new generation of electronic-only publishers would take over the market. The subversive part is that the originating scientist publishes, or self-archives, the preprint at his or her university or institution's repository, thereby making it freely available instantly to anyone with network access anywhere in the world. If the preprint isn't accepted for publication in a print journal, the author doesn't have to withdraw it, but if it does get published, everyone will use the published version. Crudely put, the publisher now seems superfluous. In Harnad's "Post Gutenberg Galaxy," money doesn't change hands and permission isn't a barrier (1994).

By 1994, the scientific community had already been using electronic files for archiving scientific literature. The first centralized archive, begun in 1991, was arXiv.org, a physics archives out of Los Alamos, New Mexico, now owned and operated by Cornell University (Gustafson, 2004). Self-archiving involves depositing a digital document at a publicly accessible, institutional Web site. Until standards emerged that made archives interoperable, institutional repositories were largely unsearchable; hence, self-archiving did not guarantee research impact, the sole reason scholars publish their findings. Interoperability guarantees that any user anywhere in the world can search archives in repositories also located anywhere (Harnad, 2001, May). The technical breakthrough that makes interoperability possible is XML.

Interoperability involves a single Web interface where the depositor enters tags for the metadata, date, author-name, title, journal-name, and then attaches the full-text document (Eprints.org, n.d., FAQ, section). Full text documents can be in different formats and locations, but the same metadata tags make them interoperable. The interoperable interface was developed by an international organization called The Open Archives Initiative (OAI, 2001). Eprints.org, out of Southampton University, UK, created self-archiving software that is OAI-compliant. Eprints software enables the whole open archives system to work technologically. The Open Archives Initiative established a registry for OAI-compliant archives that use Eprints software ePrints.org (Eprints.org, n.d., Institutional archives registry section). Any individual or institution running a UNIX operating system can download Eprints software for free, set up a self-archiving repository, and register with OAI (ePrints.org, 2004, Self-archiving and open archives section).

In 1999, the Open Archives Initiative convened in Santa Fe, New Mexico, to work out "a technical and organizational framework to support basic interoperability among e-print archives" (Open Archives Initiative, 2001). The framework instituted OAI compliance, enabling interoperability among Eprints archives, such that all archives can be harvested, integrated, navigated, and searched seamlessly, as if they were all in one global archive. The user could search, via a cross-archive search engine, a virtual archive and retrieve documents from university or institutional archives anywhere in the world, whether the archive were centrally located or distributed.

Today the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (2004) provides "an application-



independent interoperability framework based on *metadata harvesting* (Introduction section). The OAI-PMH framework allows for two types of participation: At the input end Open Archive Data Providers use OAI-PMH as a means of compiling metadata in a repository. At the output end Open Archive Service Providers use OAI-PMH to harvest metadata in response to a search (2004, Introduction section). Harvesting refers to the process of collecting metadata from repositories. A repository is a network accessible server running Eprints software that can process OAI-PMH requests. Interoperability requires that repositories use Dublin Core metadata. Metadata is assigned a unique identifier that identifies an item within a repository. OAI-PMH requests use the unique identifier to extract metadata from the item (2004, Definitions and concepts section).

One OAI Service Providers is ARC - A Cross Archive Search Service maintained by Old Dominion University Digital Library Research Group (2004). ARC is used to harvest OAI compliant repositories by making them accessible through a unified search interface. It is not a production service but searches other repositories. The total size of the all archive groups is 7,156,195 items (ARC, n.d., Archive groups section). The largest repository is OCLC's Experimental Thesis catalog, xtcat.oclc.org, 4,373,074 items (ARC, n.d., Archive groups section). You can search for a document in ARC by seeking keyword matches on all bibliographic fields or on specific bibliographic fields: author, title, type, language, archive, subject, accession date, discovery date, or abstract fields (ARC, n.d., Help section).

ARC maintains DP9- An OAI Gateway Service for Web Crawlers, from which users can access worldwide OAI compliant repositories (ARC, n.d., DP9 section). One such repository, number 301, is the Haverford College Senior Thesis Archive (ARC, n.d., Repository name: Haverford College section). If you click on the hyperlink for document oai:HaverfordCollegeThesis.OAI2:44, you will find the Dublin Core Metadata and abstract for Diana Postemsky's BA thesis, *Through the Looking-Glass: Reading and Reflecting from Wide Sargasso Sea to Jane Eyre*. These metadata tags: title, creator, subject, description, contributor, publisher, date, type, format, identifier, source, language, rights are completed by the depositor, as explained in the Help page at Haverford College Senior Thesis Archive (n.d., User documentation, Depositing records section). The thesis is available in full text as a PDF document at the Haverford College Library Senior Thesis Archive (Postemsky, 2003).

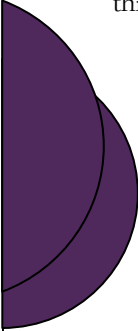
The goal of OAIster, another open archive service provider from the University of Michigan Digital Library Production Service (2004), is to create a collection of freely available, previously difficult to access, academically oriented, digital resources that are easily searchable by anyone. OAIster holds 740,751 records from 390 institutions' open archive repositories (University of Michigan, 2004). Perseus Digital Library Lookup Tool, out of Tufts University (2004), is another cross archive search service.

The software that supports cross archive interoperability is GNU Eprints developed at Massachusetts Institute of Technology (MIT) and the University of Southampton. GNU (pronounced "guh-noo"), developed by the GNU Project at MIT in 1985, is a free operating system that is upwardly compatible with UNIX (GNU Project, 2004). Free software means free to run, to study, to distribute, and to improve the program (GNU Project, 2004, Free software definition section). In 2001, the GNU Project developed the GNU Hurd as the GNU operating system kernel. Before that it was running on the Linux kernel. The kernel interfaces between the hardware and the programs that are run on it, including the operating system; the entire system is referred to as GNU/Hurd or GNU/Linux (Stallman, 2004).

Eprints archive-creating software was designed at the School of Electronics and Computer Science at the University of Southampton, United Kingdom, so that universities and research institutions worldwide could create their own Open Archives (Eprints.org, 2004). The software creates Eprints Archives that are interoperable and OAI-PMH compliant. Eprints software is downloadable from the eprints.org Web site. The software is free, uses only free software, and can be installed and maintained easily. It is modular, and written to be easily upgraded with each upgrade of the Open Archives protocol (Harnad, 2000).

Open Archive data providers administer repositories, like the Haverford College Senior Thesis Archive, that use Eprints software as a means of collecting metadata about the content in their repository. Open Archives service providers, like ARC, OAIster, and Perseus Lookup Tool, search the metadata of the data providers. Service providers are able to harvest, index, and search all the data providers providing a value-added services (OAI, n.d., Community section). Participants in the OAI community share information by implementing the OAI protocol and register at the ARC Community Web site as Data and Service Providers (OAI, n.d., Community section).

All Eprints archives can be citation-interlinked so that the research literature can be navigated by citation. CiteBase Search (n.d.) was developed by the OpCit Project (2004) as a search and citation analysis tool with data from August 1999 to the present. It interlinks all Open Access archives by enabling searches of the current article's reference list or all articles cited by



this articles, all articles citing this article or articles that have referred to the current article in their reference list, all articles co-cited with this article or articles that have been referred in the same citing article as the current article. This is useful because co-cited articles are likely to cover the same topic or argument. Searches can choose a metadata tag, a citation, or an OAI identifier search window (CiteBase Search, n.d.).

The Open Citation (OpCit) Project (2004) is a three-year collaboration between Southampton University, Cornell University and arXiv.org that ended in 2002. OpCit developed products and services that support OAI. One of its accomplishments was to explore the relationship between usage and impact for free online papers. The issue of impact is a serious one that has been studied extensively. Impact is a major reason why scholars publish research findings. Scholars fear that self-archiving and online publishing will limit the impact of their research. OpCit developed CiteBase as a tool to measure the usage and impact of OAI published papers. OpCit (2004, October) reported on studies tracking the effect of open access. Results showed that open access increases impact. Attelman (2004) reported findings that “across all four disciplines...philosophy, political science, electrical and electronic engineering and mathematics...freely available articles do have a greater research impact.”

OAI advocates calculate that the research impact of Eprints repositories will increase as researchers in all disciplines self-archive their research, both pre-refereeing preprints and refereed postprints. The Core Metalist of Open Archives Eprint Archives gives the present scale of open access archives and of author self-archiving (OpCit Project, 2003). To free the world's refereed research literature from its current access-barriers and impact-barriers was the initial idea of Harnad's Subversive Proposal (1994).

Harnad's call went out to researchers, universities, and libraries that researchers self archive present, future and past papers, that universities mount Eprints Archives, mandate them, and help in author start-up, and that libraries administer Eprints archives and help in author start up. The benefit for researchers is to increase impact; the benefit to the scholarly community is to enable free, unrestricted access to communications; the benefit to libraries is to redirect the 10-30% savings on the serials subscriptions budget to go for quality control and certification such as peer review and editing services and for consortial support for Eprints Archives (Harnad, n.d., Resolving the anomaly).

Harnad envisioned a new paradigm in the world of refereed scholarly publishing in which publishers will downsize to providing quality control and certification (QC/C) (n.d., Resolving the anomaly). Separating access from the quality control ensures that publishers have no jurisdiction over who gets to see what. The Subversive Proposal (1994) is designed to get around restrictive copyright legally. Publishing unrefereed preprints by self-archiving before submitting the paper to a journal enables the author to negotiate to hold, rather than transfer, copyright. If the author holds copyright, the author could self-archive the refereed postprint. If the author loses copyright, the author would self-archive the “corrigenda,” the differences between the preprint and the postprint (Harnad, n.d., Resolving the anomaly). Either way the entire article would be freely available or the author's research impact continue unfettered.

References

American-Scientist-Open-Access-Forum. (2004). *By thread*. Retrieved December 5, 2004, from <http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/index.html>.

Antelman, K. (2004, September). Do open-access articles have greater research impact? *College and Research Libraries* 65(5) 372-82. Retrieved December 5, 2004, from http://www.lib.ncsu.edu/staff/kantelman/do_open_access_CRL.pdf.

ARC-A cross archives search service. (2004). *Homepage*. Retrieved January 11, 2005, from <http://arc.cs.odu.edu/>.

ARC. (2004). *Community*. Retrieved December 5, 2004, from <http://www.openarchives.org/community/index.html>.

ARC. (2004). *DP9- An OAI gateway service for web crawlers*. Retrieved December 4, 2004, from <http://arc.cs.odu.edu:8080/dp9/index.jsp>

ARC. (2004). *Lithuanian electronic thesis and dissertation archive*. Retrieved December 4, 2004, from <http://arc.cs.odu.edu:8080/dp9/listidentifiers/etd.library.lt/response2.html>.

ARC. (n.d.). *Archive Groups*. Retrieved January 11, 2005, from <http://128.82.7.99:8080/oai/results.jsp>

ARC. (n.d.). *Help*. Retrieved December 4, 2004, from http://128.82.7.99:8080/oai/service_help.html#specific.

CogPrints: Cognitive Sciences Eprints Archive. (2004). *Welcome to Cogprints*. Retrieved December 5, 2004, from <http://cogprints.org/>.

Eprints.org. (2004). *Self-archiving and Open Access (OA) Eprint archives*. Retrieved December 5, 2004, from <http://www.eprints.org/>.

Eprints.org. (n.d.). *Citebase search*. Retrieved December 5, 2004, from <http://citebase.eprints.org/cgi-bin/search?type=metadata>.

Eprints.org. (n.d.) *Eprints.org FAQ*. Retrieved April 24, 2004, from <http://www.eprints.org/self-faq/#serch-archiving>.

Eprints.org. (n.d.). *Institutional archives registry homepage*. Retrieved January 11, 2005, from <http://archives.eprints.org/>.

GNU Project. (2004). *GNU operating system free software foundation*. Retrieved January 11, 2005, from <http://www.gnu.org/>.

GNU Project. (2004). *Free software definition*. Retrieved December 4, 2004, from <http://www.gnu.org/philosophy/free-sw.html>.

Gustafson, E. (2004). *IMS journals on arXiv*. Retrieved November 25, 2004, from http://stat-www.berkeley.edu/users/pitman/arxiv_article.html.

Hanard, S. (2001, May). The self-archiving initiative. *Nature Web Debates*. Retrieved December 4, 2004, from <http://www.nature.com/nature/debates/e-access/Articles/harnad.html>.

Harnad, S. (2001). The (refereed) literature-liberation movement. *The New Scientist*. Retrieved April 24, 2004 from <http://www.ecs.soton.ac.uk/~harnad/Temp/newscientist.htm>

Harnad, S. (2000). *Re: Eprints Open Archive software*. Retrieved December 4, 2004, from <http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/1057.html>.

Harnad, S. (1994). Scholarly journals at the crossroads: A subversive proposal for electronic publishing. *ARL Issues in Scholarly Communication*. Retrieved April 30, 2004, from <http://www.arl.org/scomm/subversive/sub01.html>.

Harnad, S. (n.d.). *Resolving the anomaly*. Retrieved December 5, 2004, from <http://www.ecs.soton.ac.uk/~harnad/Tp/2-Resolving-the-Anomaly/sld001.htm>.

Harnad, S. (n.d.). *How to get around restrictive copyright legally*. Retrieved December 5, 2004, from <http://www.ecs.soton.ac.uk/~harnad/Tp/2-Resolving-the-Anomaly/sld007.htm>.

Haverford College Senior Thesis Archive. (2004). Retrieved December 4, 2004, from <http://thesis.haverford.edu/view/departement/English.html>.

Haverford College Senior Thesis Archive. (n.d.). *User archive: depositing records*. Retrieved January 11, 2005, from <http://thesis.haverford.edu/help/#Depositing>.

Hitchcock, S. (2005, January). The effect of open access and downloads ('hits') on citation impact: a bibliography of studies. *OpCit: The open citation project*. Retrieved December 5, 2004, from <http://opcit.eprints.org/oacitation-biblio.html>.

Hitchcock, S. (2003). (2003). Core metalist of Open Access Eprint archives. *OpCit: The open citation project*. Retrieved January 11, 2005, from <http://opcit.eprints.org/explorearchives.shtml>.

Microsoft News (MSN). (2004). *MSNBC News*. Retrieved November 29, 2004 from <http://www.msnbc.msn.com/>.

Old Dominion University Digital Library Research Group. (2004). *ARC-A cross archives search service*. Retrieved January 11, 2005, from <http://arc.cs.odu.edu/>.

OpCit: The Open Citation Project. (2004). *Homepage*. Retrieved December 5, 2004, from <http://opcit.eprints.org/>.

Open Archives Initiative. (2004). *OAI Homepage*. Retrieved April 28, 2004, from <http://www.openarchives.org/index.html>.

Open Archives Initiative. (2004). *Protocol for metadata harvesting [OAI-PMH]*. Retrieved December 4, 2004, from <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

Open Archives Initiative. (2004). *OAI-PMH: Introduction*. Retrieved December 4, 2004, from <http://www.openarchives.org/OAI/openarchivesprotocol.html#Introduction>.

Open Archives Initiative. (2004). *OAI-PMH: Definitions and concepts*. Retrieved December 4, 2004, from <http://www.openarchives.org/OAI/openarchivesprotocol.html#DefinitionsConcepts>.

Open Archives Initiative. (2001). *Santa Fe Convention for the Open Archives Initiative (OAI)*. Retrieved December 4, 2004, from http://www.openarchives.org/meetings/SantaFe1999/sfc_entry.htm

Open Archives Initiative. (n.d.). *Community*. Retrieved January 11, 2005, from <http://www.openarchives.org/community/index.html>.

Postemsky, D. (2003). Through the looking-glass: Reading and reflecting from wide Sargasso Sea to Jane Eyre. *Haverford College Senior Thesis Archive*. Retrieved December 4, 2004, from <http://thesis.haverford.edu/archive/00000044/01/2003PotemskyD.doc.pdf>.

Psychology. (2004). *Homepage*. Retrieved Dec. 5, 2004, from <http://psycprints.ecs.soton.ac.uk/>.

Stallman, R. (2004). Linux and the GNU project. *GNU Project*. Retrieved December 4, 2004, from <http://www.gnu.org/gnu/linux-and-gnu.html>.

Suber, P. (2004, November). FY05 Omnibus Appropriations Conference Report, NIH, Office of Director. Mailing List SPARC-OAForum@arl.org Message #1317. Retrieved December 4, 2004, from <https://mx2.arl.org/Lists/SPARC-OAForum/Message/1317.html>.

Tufts University Perseus Digital Library (2004). *Perseus lookup tool*. Retrieved December 5, 2004, from <http://www.perseus.tufts.edu/cgi-bin/vor>.

University of Michigan Digital Library Production Service. (2004). *OAIster homepage*. Retrieved December 4, 2004, from <http://oaister.umdl.umich.edu/o/oaister/>.

ISSN 1541-7980

b/ITe (The Bulletin of the Information Technology Division. Electronic) is published six (6) times a years by the Information Technology Division of Special Libraries Association

Publisher: Holly Chong-Williams

Holly.chong-williams@thomson.com

Editor Shawn Livingston

sdlivi00@email.uky.edu

<http://www.sla.org/division/dite/bite>

