



Title	The Pacific Rim Library: A Surprising Pearl
Author(s)	Palmer, DT
Citation	Serials Review, 2009, v. 35 n. 3
Issue Date	2009
URL	http://hdl.handle.net/10722/54697
Rights	Creative Commons: Attribution 3.0 Hong Kong License

The Pacific Rim Library: A Surprising Pearl

The Pacific Rim Library (PRL) is an initiative of the Pacific Rim Digital Library Association (PRDLA). The project began in 2006 using the OAI-PMH paradigm and now holds over 300,000 records harvested from OAI data provider libraries around the Pacific. PRL's goal is to enable the sharing of digital collections amongst PRDLA members and the world, but greater unexpected benefits have been discovered. Through mirroring their metadata, PRL increases the chance that their data will be discovered in Google and other general search engines. With its many disparate collections, PRL is not a repository for traditional information discovery and retrieval. Initially users will bounce from a Google hit, to the PRL metadata record in Hong Kong, and then begin an intensive search on the original site which hosts the full digital object, in Vancouver, Honolulu, Wuhan, Singapore, or other PRDLA member location. *Serials Review* 2009; xx:xxx-xxx.

© 2009 Published by Elsevier Inc.

Keywords: OAI; Pacific Rim Library; Pacific Rim Digital Library Association; PRDLA; PRL; Information discovery and retrieval; Indexing; Metadata; Deep Web

Introduction

The Pacific Rim Digital Library Association (PRDLA) was formed in 1997 by fourteen libraries in the Pacific region. Now thirty-one libraries strong, the goal of this group is to "improve access to scholarly research materials through cooperative ventures."^{1,2} Over the years the association has sponsored many such endeavors. A task force was established in 2004 to study the best means for sharing PRDLA digital collections amongst themselves and the world. In its 2005 report, the task force recommended creating a Web-based Open Archives Initiative (OAI)³ and determined that establishing a Service Provider would be the quickest, easiest, and most effective way of accomplishing this goal.

Open Archives Initiative Model

In the OAI paradigm, there can be many geographically or categorically disparate Data Providers hosting their own local repositories, with digital objects of full-text items or other files. Each Data Provider will expose the metadata on these items; a Service Provider can then harvest, importing the metadata (and not the digital object) into a new repository. The Service Provider will apply a search engine on this new repository. Users can then do one search across all metadata of the several Data Providers. Upon choosing one record and clicking, the user leaves the Service Provider's repository and arrives on a page in one of the local Data Provider's repositories.

Developments and Refinements

In 2006, a new task force was funded by PRDLA to create an OAI pilot repository at The University of Hong Kong (HKU). The

working title of this repository was the "PRDLA Archive," which recently became "Pacific Rim Library" or "PRL" for short. (Hereafter, "PRL" will be used, although the name change has still not been enacted.) PRDLA decided that PRL would harvest OAI metadata from at least one locally created digital collection at each of the PRDLA member libraries. HKU became the OAI Service Provider, harvesting data from the many OAI Data Providers. Several PRDLA libraries then implemented the OAI protocol for metadata harvesting (OAI-PMH) for the first time. The PRL OAI data providers created XML compatible, UTF-8 compliant metadata using the simplest Dublin Core (DC) schema: "oai_dc" consisting of fifteen unqualified data elements. At that time, HKU had begun to use DSpace for its institutional repository, and thus, created a second instance of DSpace to be the OAI Service Provider for harvesting and hosting the metadata from the many PRL OAI Data Providers. Each PRDLA member became a "community" in the DSpace paradigm (Fig. 1).

DSpace worked well when there were few records. After more PRDLA member repositories were harvested, however, and record numbers reached 300,000, the response time slowed considerably. We also found problems in OAI harvesting and updating. We then found a replacement in CDS Invenio,⁵ another open source software, developed by CERN. Invenio promises fast searching across a repository of up to 1.5 M records at one time. CDS Invenio was easier to customize for OAI harvesting and provided several out-of-the-box interfaces in different languages and scripts. Our developer also contributed to the code base by making new translations for traditional and simplified Chinese as shown in the following illustration (Fig. 2).

In order to increase the value of PRL, we changed the oai_dc schema used among the libraries, to include one more qualification on Identifier, that of an "identifier thumbnail." This identifier allows for the harvesting and storage of thumbnail images along with the usual bibliographic details.

Palmer is Scholarly Communications Head, HKU Libraries, The University of Hong Kong, Hong Kong; e-mail: dtpalmer@hku.hk.

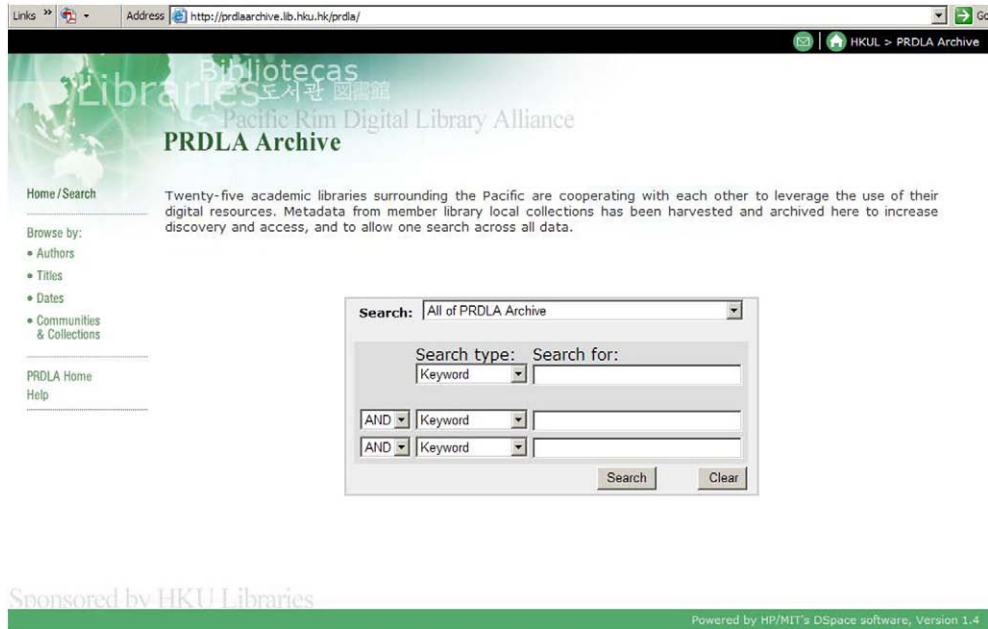


Figure 1. PRDLA archive in DSpace (2006).

80 The 2007 Berkeley PRDLA meeting pronounced this repository
 81 fit for purpose and asked that it go into “production.” We removed
 82 the password access and specifically invited external robots and
 83 crawlers inside the system by creating sitemaps for the major
 84 search engines to index content.⁷ We created records for PRL in the
 85 following major OAI-PMH registries and repositories:

- 86 • Open Archives list of registered OAI repositories: <http://www.openarchives.org/Register/BrowseSites>
- 87 • OAI registry at University of Illinois at Urbana-Champaign: <http://gita.grainger.uiuc.edu/registry/>
- 88 • Celestial OAI registry: <http://celestial.eprints.org>
- 89 • Registry of Open Access Repositories: <http://roar.eprints.org/>
- 90 • Directory of Open Access Repositories — OpenDOAR: <http://www.openarchives.org/>
- 91
- 92
- 93

The Purpose of PRL

94

Although the membership of PRDLA was generally pleased with
 95 the progress of PRL, at the 2007 meeting some members expressed
 96 uneasiness at the disparate nature of the collections hosted in PRL.
 97 Traditionally, databases are created and chosen for searching
 98 because of a selection process in the creation of these databases
 99 that strives to include as much relevant content in a given subject
 100 as possible, and which, therefore, excludes material beyond the
 101 scope of that subject. Although there are many collections in PRL
 102 that are unquestionably relevant to the Pacific area, such as the
 103 “Sea of Korea Map Collection” and the “Hawaiian Photo Album,”
 104 there are also several that do not have direct relevance to the
 105 Pacific area, such as “The Automobile Club of Southern California.”
 106



Figure 2. PRDLA archive in CDS Invenio; simplified Chinese interface.

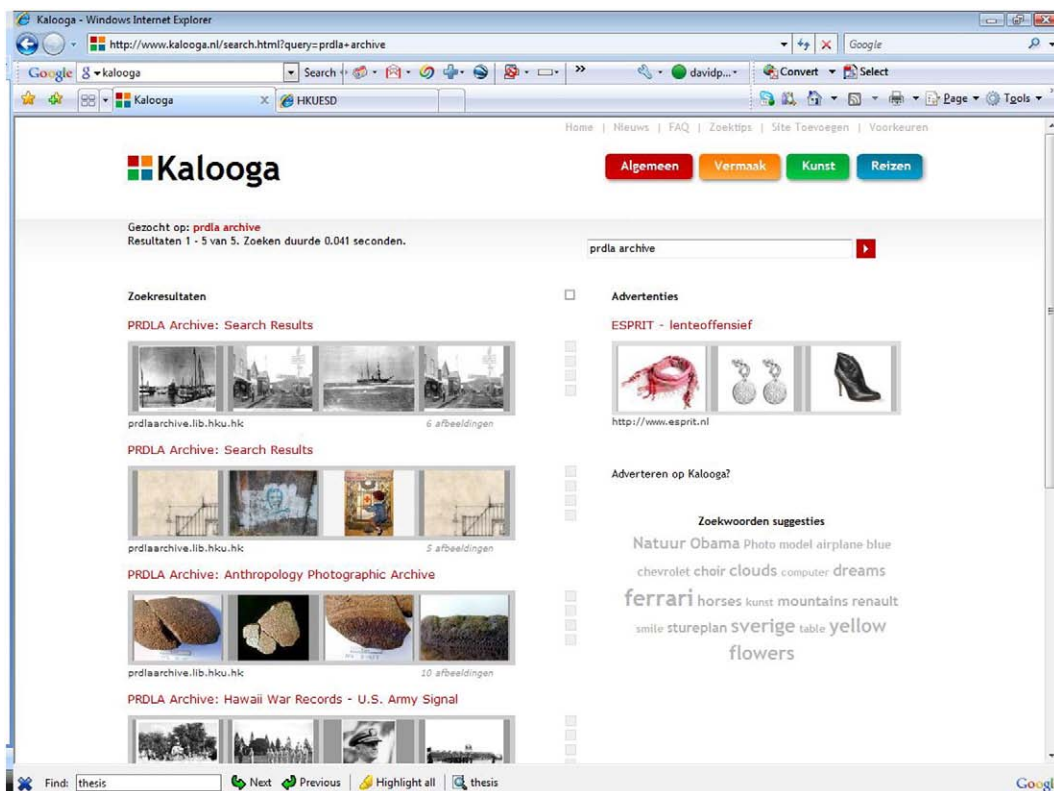


Figure 3. PRL thumbnails in Kalooga.

107 PRDLA charged a new task force to survey the membership and
 108 propose scope and direction for PRL. This task force reported its
 109 findings at the 2008 Singapore meeting and made the following
 110 recommendations:

- 111 1. Change the name to “Pacific Rim Library” with the acronym
- 112 “PRL,” pronounced as “Pearl” (as in “pearl of the orient”) to
- 113 more accurately represent the content of this repository.
- 114 2. That the contents of PRL be about the Pacific Rim areas and that
- 115 the purpose of PRL is to provide people with a way of finding
- 116 material about the peoples, cultures, history, etc. of the Pacific
- 117 Rim.
- 118 3. That PRDLA should invite others with Pacific Rim-related
- 119 content to allow harvesting into PRL, and that this invitation
- 120 be extended to anyone with relevant content, within or external
- 121 to the PRDLA membership.

122 Unexpected Pearls

123 The PRL Technical Committee also presented a report at this
 124 meeting that described its findings on the usage statistics of PRL.
 125 These statistics showed high activity by all of the major search
 126 bots, including Googlebot, Yahoo Slurp, Cuil's Twiceler, BaiduSpider,
 127 etc. One top referring site was a Wikipedia entry about James
 128 Wong (黃湛森), a recently deceased Cantopop singer and
 129 songwriter from Hong Kong. At that time, the entry hyperlinked
 130 to the record in PRL, which, in turn, hyperlinked to his full-text
 131 PhD thesis in the HKU Libraries. One top target identified in PRL
 132 was the War Poster Collection of the University of Washington.
 133 Thumbnail images provided by this collection in PRL are very
 134 attractive to Kalooga, and other major projects seeking thumbnail
 135 images (Fig. 3).

Search Engine Results

136

Comparing searches in Google and Yahoo on a known PRL title, with
 137 the original host providing the metadata, gives interesting results. A
 138 Google search in the HKU network on “Lise’s Lunchwagon” shows
 139 only two hosts, the originating collection at the University of
 140 Hawaii-Manoa and PRL (the UH-M record appears above PRL, as
 141 expected, because they are the original provider); however, a search
 142 on “Hoo Hoo House,” a record from the University of Washington,
 143

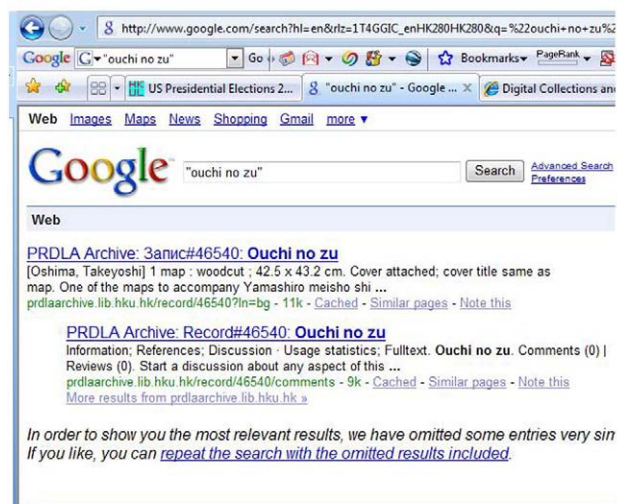


Figure 4. Google: “Ouchi no zu.”



Figure 5. Yahoo: "Ouchi no zu."

160 shows PRL first and UW second. A search on "Ouchi no zu," a record
 165 from University of British Columbia, only shows the PRL entry in
 166 Google. Yahoo search results, however, show PRL first, followed by
 167 UBC, as shown in Figs. 4 and 5.

168 During the same week that HKU performed these searches,
 169 colleagues in Honolulu and London confirmed receiving the same
 170 results on the same searches.

171 These results demonstrate that PRL is linking users to those
 172 resources that may not have been discovered otherwise. In this
 173 manner, PRL is adding value to the original locally hosted
 174 databases that comprise PRL.

175 The Deep Web

176 Indexing or revealing the many hidden Web objects in the deep
 177 Web has long been a goal toward which many commercial and
 178 academic search engines continue to strive. Several reasons
 179 explain why these Web objects remain hidden. For example, to
 180 prevent undue traffic, some repositories use protocols to specifi-
 181 cally exclude robots. Also, unless there are other links that point to
 182 these hidden Web objects, robots and crawlers will not find them.⁸

183 Most major search engines now provide procedures for
 184 repositories to follow in order to expose more of the deep Web
 185 to their robots. At one time, Google and Yahoo included OAI-PMH
 186 in these procedures; however, Google recently announced that
 187 they were retiring support for OAI-PMH in their sitemaps.⁹ This
 188 lack of support, therefore, shows the further value of PRL as a
 189 discovery source. PRL harvests OAI-PMH data from Web objects
 190 that may be hidden, in repositories that may not be completely
 191 visible to these search engines. PRL then exposes this harvested
 192 metadata to these search engines for their indexing. A recent study
 193 by the OAIster project indicated similar results.¹⁰
 245

These search results, combined with the dissimilar nature of the
 194 PRL collections, suggest that PRL's true value is as an indirect tool of
 195 discovery. PRL has enabled the discovery of this hidden Web to
 196 users of Google and Yahoo, and presumably pushed ranking of
 197 these pages higher within these and other search engines. The
 198 end-user will presumably find an item in the search engines,
 199 bounce to PRL in Hong Kong, and then link to the originating
 200 database in Hawaii, Wuhan, Singapore, or other PRDLA member
 201 location, to begin a more focused search.
 202

The Future Value of PRL

PRL has become a flagship project of PRDLA and serves to rally
 204 member support and to promote their existence to the world. In
 205 recognition, PRDLA has begun two new projects focused on PRL.
 206 PRDLA is redesigning its logo and PRL graphics to enhance its
 207 image. PRDLA has also created a collaboration for a new collection,
 208 Oceania Digital Libraries (ODiL), expected for release in 2009.¹¹
 209

In summary, the PRDLA goal for PRL is the creation of a tool for
 210 searching content about the Pacific Rim. However, the way that PRL
 211 does this is not as a traditional destination database, but one that
 212 acts as an intermediary between the search engines and the
 213 originating database. Although this is beyond the original goal, the
 214 PRDLA membership understands the value that PRL provides and
 215 will continue to use this project and its Web pages to showcase
 216 their collections.
 217

Notes

1. "About [PRDLA]," http://www.prdl.ucmercedlibrary.info/?page_id=2,
 (accessed March 30, 2009). 219-220
2. "Membership [of PRDLA]," <http://prdl.ucmercedlibrary.info/about/membership/>,
 (accessed March 30, 2009). 221-222
3. "OAI for Beginners," <http://www.oaforum.org/tutorial/>, (accessed March 30,
 2009). 223-224
4. "Dublin Core Metadata Initiative," <http://dublincore.org/usage/terms/dc/current-elements/>,
 (accessed March 30, 2009). 225-226
5. "CDS Invenio Overview," <http://cdsware.cern.ch/invenio/>, (accessed March
 30, 2009). 227-228
6. A "thank you" on this page to Dr. Ku Kam-ming of HKU, "CDS Invenio
 Overview," <http://cdsware.cern.ch/invenio/index.html>, (accessed March 30,
 2009). 229-230
7. "Sitemaps," <http://en.wikipedia.org/wiki/Sitemaps>, (accessed March 30,
 2009). 232-233
8. Frank McCown, Michael L. Nelson, and Mohammad Zubair, "Search Engine
 Coverage of the OAI-PMH Corpus," *IEEE Internet Computing* (March/April
 2006), <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01607990>.
 234-236
9. John Mueller, "Retiring support for OAI-PMH in Sitemaps," *Google Webmaster
 Central Blog* (April 23, 2008), [http://googlewebmastercentral.blogspot.com/
 2008/04/retiring-support-for-oai-pmh-in.html](http://googlewebmastercentral.blogspot.com/2008/04/retiring-support-for-oai-pmh-in.html), (accessed March 30, 2009). 237-239
10. Kat Hagedorn and Joshua Santelli, "Google Still Not Indexing Hidden Web
 URLs," *D-Lib Magazine* 13, no.7/8 (2008), [http://www.dlib.org/dlib/july08/
 hagedorn/07hagedorn.html](http://www.dlib.org/dlib/july08/hagedorn/07hagedorn.html), (accessed March 30, 2009). 240-242
11. "PRDLA," [http://prdl.ucmercedlibrary.info/2008/10/oceania-digital-
 libraries-odil-collaborative-digitization-project/](http://prdl.ucmercedlibrary.info/2008/10/oceania-digital-libraries-odil-collaborative-digitization-project/), (accessed March 30, 2009). 243-244