Accessing library materials via Google and Other Web Sites

Paper presented to ELAG (European Library Automation Group) 9 May 2007, Barcelona, Spain

by Janifer Gatenby, Strategic Research, OCLC PICA

Abstract

Two and a half years have passed since libraries first started to make available the contents of their catalogues to the major Internet search engines, Google, Yahoo and Microsoft Network (MSN). The paper examines the success of this initiative and various aspects including search engine selection policies, ranking, service evolution and statistics in terms of "click throughs" and "conversions". The benefits to libraries of exposing their collections as broadly as possible, additional sites to GYM and methods of exposure are examined. But there is becoming a serious "discovery to delivery gap"; linking seamlessly to delivery systems is a necessity if libraries are to sit proudly alongside web sites like online book stores and match them for ease of requesting materials. Recent developments in standards and in OCLC's worldcat.org are presented as steps towards improvement in the delivery area.

History

"if you can't beat 'em, join 'em" so said the author John Martin in 1939 (Dictionary of quotations 1939) [1]
It was about 3 to 4 years ago that this started to occur to librarians as the key to the evolution of their services and the revival of their collections. In 2003 OCLC produced the *Environment scan* followed 2 years later by *Perceptions of libraries and information resources*. These reports confirmed what everybody already knew, that users, young, student, post graduate alike, go first to search engines with their information needs. The library's local OPAC was and is being challenged on multiple sides, but particularly by search engines and online bookshops with refreshingly simple and innovative user interfaces providing either direct access to online text or direct online requesting.

Instead of persisting in the faith that users would turn to libraries and their OPACS for better quality results and better coverage of all resources, whether electronic or physical, some started to see that exposing hitherto "hidden treasures" to the major Internet Search engines would actually produce an increase in traffic to the library's OPAC. For a few, the fact that the entry may be directly to a record or holdings page remained problematical, but the usage statistics started to come in from those who had dipped their toes in the water.

In Dec 2004, OCLC launched the OWC program that was presented to ALA mid winter in Jan 05. This program includes the major search engines, Google, Yahoo and MSN (commonly known as GYM), online books stores, antiquarian book stores and other sites. Many other libraries and union catalogues have contracted directly with Google including union catalogues from 12 nations that have made their data available to Google Scholar as part of its union catalogue program. As at August 2006, these nations were Australia, China, *Czech Republic, *Denmark, Ireland, Israel, Hungary, Lithuania, *Netherlands, Taiwan, *United Kingdom and *United States. Negotiations with others are under way and many more contribute via WorldCat such as South Africa and Poland. The asterisks indicate significant contributions via WorldCat.

How data is contributed

Contributing to the search engines means providing the data in a format easily ingestible by them. OCLC created small XML records for each work in WorldCat. By request, a subset of these records, representing approximately 75% of all WorldCat holdings were placed in a separate server environment and made

[1] He actually paraphrased James Eli Watson, the US senator whose phrase "if you can't lick 'em, jine 'em" was his catch cry.

accessible via HTTP in tagged XML or Inktomi Data Interchange Format (IDIF) depending on harvest partner. The current composition of this subset is:

Title, Author, Publisher, Standard identifiers, electronic location, Subject, Language, Genre / Form, Document type, Person as Subject, Contents, Country holding and Holdings count. All elements but the last are mapped to MARC21 subfields.

The number of records harvested depends on the search engine. The Google main index only includes about 4.4 million records, but covering about 75% of the holdings on WorldCat. 3 million of the 4.4 million are clustered "work type" records with holdings consolidated from related manifestation records. MSN includes 4.5 million books, theses and dissertations limited to physical sciences and biomedical. This will grow to 10 million by mid 2007. Yahoo includes 3.5 million, 3 million of which are clustered records. Google Scholar accepts more, mapping its subset to 67 million clustered records and the mapping from Google Books to WorldCat is close to 100%. At OCLC additions and manifestations are available in real time but OCLC has little influence over the search engine's frequency of harvestings and the subsequent ingestion and inclusion in their indexes.

Common Problems

There have been some common problems reported by libraries and union catalogues contributing to the major search engines:

1. Coverage. The search engines do not take all the records available to them. Even Google Scholar drops the unique material, only keeping holdings for records already appearing in its database (Larsen 2007) and there are matching failures (Libraries Australia 2006).

2. Ranking. Competition is fierce to appear on one of the first three pages of search results as this represents real money to a large majority of organisations. The search engines keep their algorithms secret; perhaps to avoid accountability, fraudulent manipulation or to protect the algorithm as an essential business asset or a combination of all. This much is known; the ranking is based on pages referring, page hits and whichever pages will bring in the most advertising revenue to the search engine. Google created Google Scholar (Quint 2004) and Google Books as an answer to the conflicting demands of business and scholarship. In Google's main index it seems that more recently loaded material is ranked higher. Also it is important to note the growth of Google as a factor. The larger it gets, the greater for all is the struggle for relevance. When OCLC first started loading to Google in 2005, the results were appearing regularly in the first 3 pages of the main Google index. At that time there were 2 billion index items in Google compared with 12 billion now.

3. Matching. Google Scholar only matches books, not articles and as noted above Google's current matching algorithms are not as sophisticated as those developed in the library world.

4. The Danish load to Google Scholar was all at the manifestation level which makes the inwards links less effective as the Danish National Union Catalogue is clustered following FRBR principles. (Larsen 2007)

5. There is not enough influence after the data is harvested.

To overcome these problems, OCLC decided to create a global library site with "web presence". This site, launched in August 2006 exposes the entire bibliographic contents of WorldCat. Since launching worldcat.org, traffic for each month has doubled compared with the same month the previous year. As a result, many union catalogues from around the world, including those already loading independently to Google Scholar are now loading to WorldCat with the main motivation being exposure in worldcat.org and its partner program.

Positive Signs of Success

Despite the short comings of exposing and loading to the search engines, the results have been resoundingly impressive, particularly looking at OCLC's statistics.

WorldCat.org statistics for March 2007. Highlights:

- **Total Referrals: 14,118,777**. **Up 96% over March 2006** (7,195,770)

- **Average Daily Referrals: 455,444**

- **Links to Library Services:** OWC facilitated **834,886 clicks to library services** (e.g. OPACs, ILL, QP) from the interface in March. Clicks to libraries are **up 363% over March 2006** (180,264).

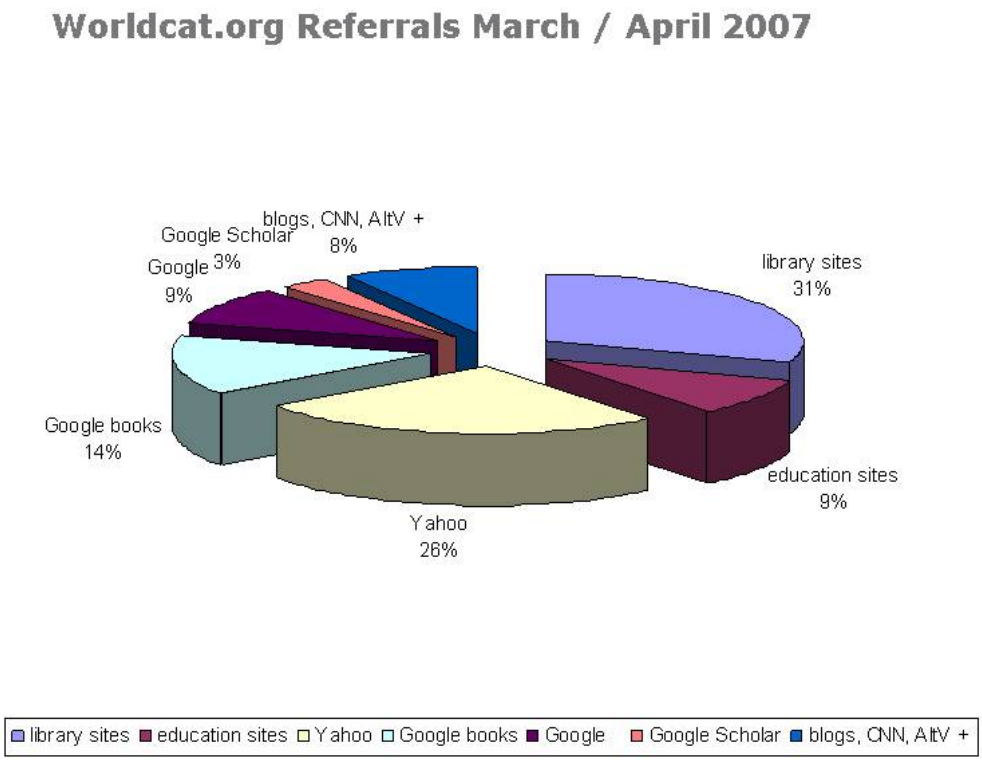
- **Library OPAC Click-through Percentage:** of the 6.9 million record views on which a library OPAC link appeared, **10.5% (~723,000) resulted in a click-through to a library catalog**.

- **Yahoo! Referrals: 1,590,841**. Up 6.8% from February (1,489,787).

- **Google Referrals (inc. Scholar): 1,647,165**. Up 9.2% from February (1,508,106). Taken separately, **Google Scholar referrals** were up 39% to 205,209, while **Google main index referrals** rose 6% to 1,441,956 (this includes Google Books referrals).

- On March 31st OWC passed **215 million total referrals** to the project since its inception.

It's difficult to get comparative statistics but the Danes report that 0.1% of overall visits are from Google Scholar, but for digital books which are covered 100% by Google Books, traffic from Google is 1.2% (Larsen 2007). OCLC releases statistics for worldcat.org, the free end user access but these are not combined with library professional accesses via cataloguing and traditional enquiry interfaces, so the OCLC figures are not direct comparisons. Nevertheless, it is clear that size and branding play an important part in user activity on the web.



What this pie chart above does ***not*** show is that this represents a little under half the traffic ("impressions") on worldcat.org in the time period. Once users find the site, they search around more and presumably, some bookmark it and return directly.

Some Observations

- There is a need for maximum exposure, not just to Google. Yahoo and Google are currently providing approximately equal traffic but in the past have alternated the top position. The MSN service is evolving and statistics from this site are expected on the same scale.

- Users value library information. The statistics of "click-throughs" defined as those who click further once arriving at a web page, are high and more importantly for "conversions" , defined as those who click through to a library or other delivery site, are significantly higher than the average for web sites as a whole (5.5% compared with 1 to 3%), ranked by Google as "best in class"

- Alternative means of exposure produce impressive results. These include a downloadable search box, a toolbar pug in for Internet Explorer and Firefox browsers and permalinks (direct URLs to each record). A surprising amount of traffic comes from small sites. For example, Squiggler.com is a general discussion blog of a courageous struggling lady. She has added the Worldcat.org search box towards the bottom of her page. In a 30 day period from March to April 2007, this site sent 4913 referrals to worldcat.org. Other examples are mysticbourgeosieblogspot.com with 13,234 referrals, websearch.cnn.com with 6,511 referrals and stumbleupon.com with 4,307 referrals.

- Users are interested in getting what they find which is evidenced by "conversions" not just to library OPACs but also to Amazon. And the "conversions" experienced by Amazon from WorldCat referrals are well above the web average, between 5.5% and 6%.

Referrals to Amazon.Com Q1 2007

		Jan-07	Feb-07	Mar-07
Ger	Referrals	0	1546	2553
Can	Referrals	1493	1311	1314
UK	Referrals	7291	6525	6461
US	Referrals	71072	82813	92631
	total	79856	92195	102959
Ger	Conversions	0	4.72%	4.07%
Can	Conversions	3.82%	1.68%	2.51%
UK	Conversions	8.30%	9.16%	7.85%
US	Conversions	4.84%	4.26%	4.05%
	av.	5.65%	6.60%	6.16%
Ger	Orders		73	104
Can	Orders	57	22	33
UK	Orders	605	598	507
US	Orders	3438	3525	3747
	total	4100	4218	4391

It is curious to note here that before Amazon links were introduced, there were links to a major library supplier that produced very disappointing traffic. This underlines again the importance of brand recognition in relation to user behaviour.

- Google Print accesses, of which there were 950,765 in March, are interesting in that they are in the most part pointing from snippets of fully digitised copyrighted material to WorldCat where the holdings may be either physical or digital.

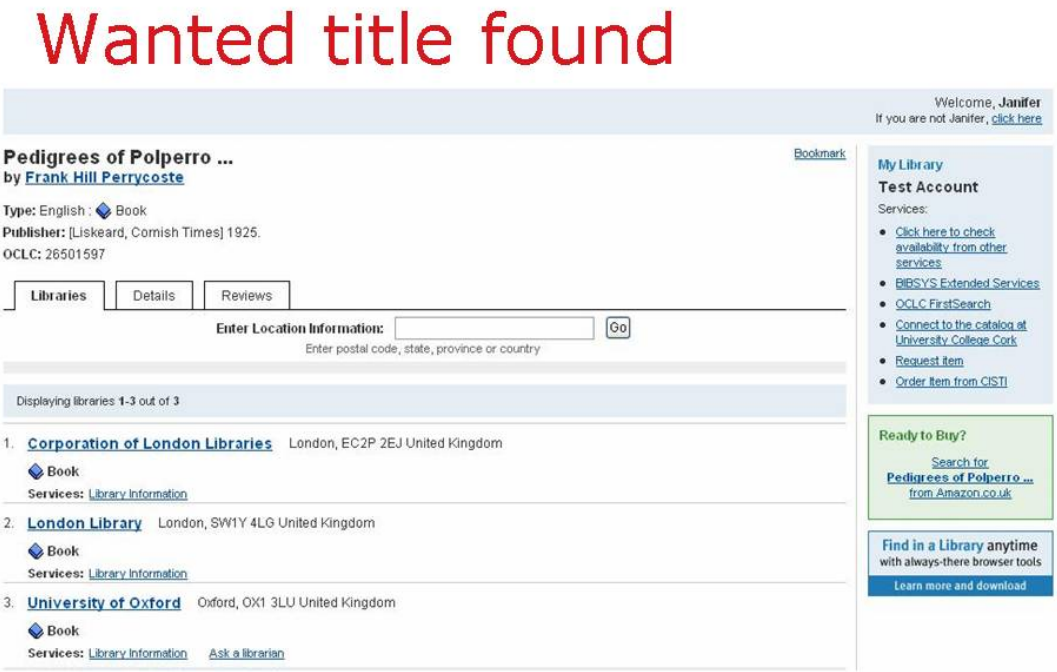
- Working with search engines and other partners is an ongoing task at OCLC with one dedicated senior staff member in at least weekly contact with the major partners. The partnership agreements are in evolution with discussion currently centring around improvements in delivery and cooperation on collection analysis.

- There is a serious discovery to delivery gap that threatens to worsen as the data is exposed more broadly in systems remote from delivery systems. A major Dutch university reports rejecting 25 requests a week from foreign end users within the first months of the Dutch Union catalogue being loaded to WorldCat.

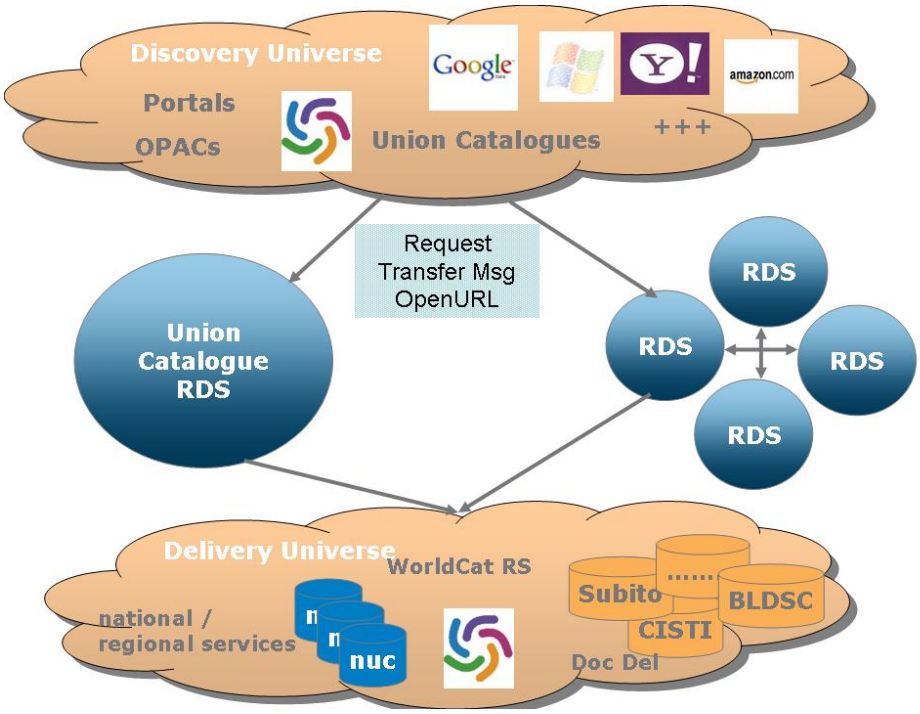
Discovery to Delivery

"The ultimate goal for using a discovery service is getting… [and libraries are becoming] great at finding but getting needs work" (Fitch 2007).

As impressive as the worldcat.org statistics are, the site is currently ranked by Alexa.com at 19,652, behind Wikipedia at 10 and Amazon at 29. Because there are two URL entries into the database (worldcat.org has superseded worldcatlibraries.org that currently ranks better at about 11,000), the actual rank is estimated to be more like 6,000. Still, 6,000 or 19,000 out of billions means that OCLC has achieved its objective of creating an international web presence for libraries, but it could be better. To improve and increase the traffic it is necessary to act on two fronts, firstly creating an attractive, easy to use and navigate interface, and more importantly, truly facilitating delivery. At the moment discovering the existence of a resource does not necessarily mean a user has any way to access it. .

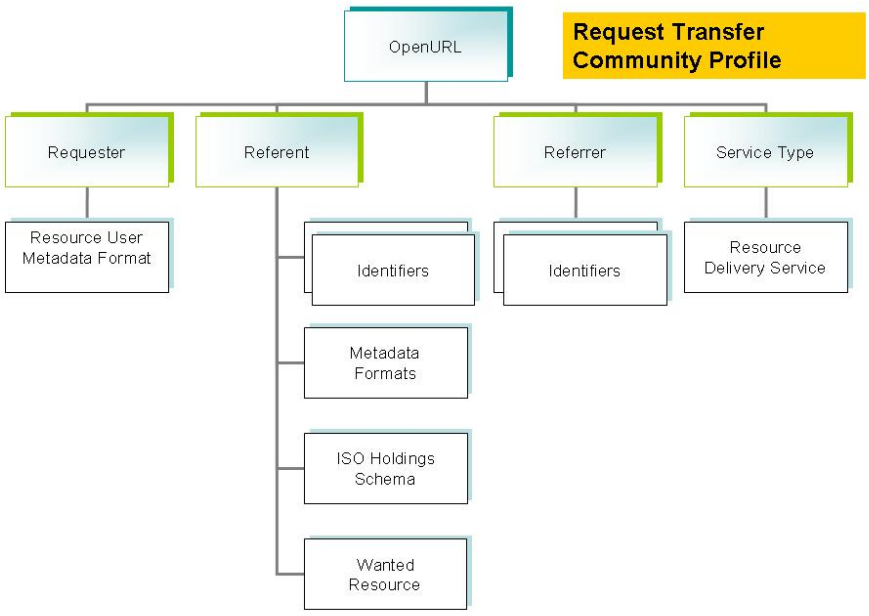


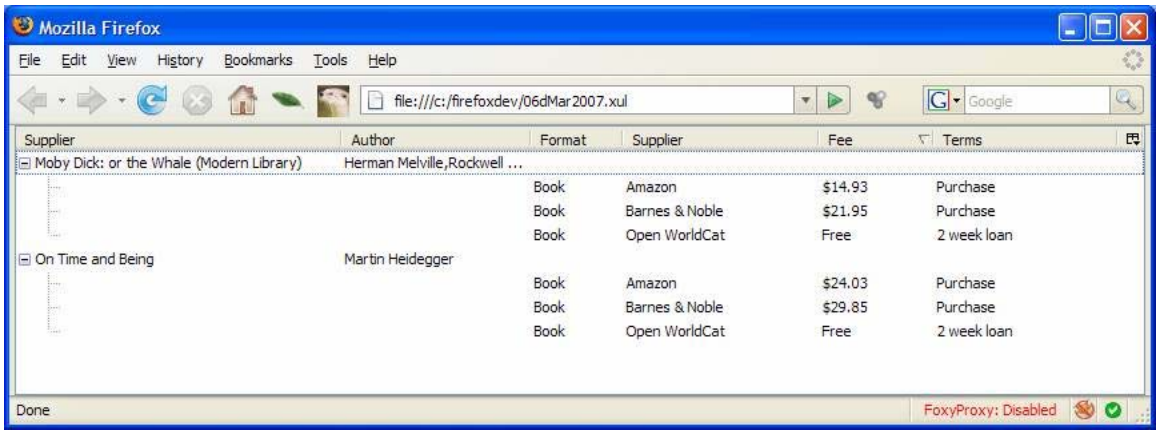
The resource listed above is rare, with only 3 copies known to WorldCat. What does a user in Amsterdam do if he or she wants to read it? The message is ambiguous and will not be allowed to persist in WorldCat.org.



The illustration above indicates that the discovery universe is becoming increasingly separated from the delivery universe with the Resource Delivery Systems (RDS) in between. There is an urgent need to develop a robust and comprehensive bridge between the two universes.

A "Get it" function will be introduced into WorldCat.org within the next few months. Behind the button will be a super resolver that determines available options depending on whether the resource is physical or digital and what can be detected about the user. This super resolver will then convey a Request Transmission Message, a community profile of the OpenURL standard (Z39.88). The Request Transmission Message includes the ISO Holdings Schema (ISO 20775) indicating possible suppliers. The Request Transmission Message is sent to the most appropriate delivery system, one that will best serve the user. The main components of the Request Transfer message are illustrated below.





OCLC's John Bodfish and others, on behalf of the Rethinking Resource Sharing Initiative, have developed a similar "Get It" function as a piece of Open Source software. This software, illustrated above, extracts bibliographic metadata from a growing number of web pages and when activated suggests delivery options indicating comparative delivery time and cost. Depending on the configuration it may create a message to transfer to worldcat.org for locations which then creates a message to transfer to the appropriate delivery system.

Alongside the "Get it" button several other pieces of "delivery architecture" are, or will be these services:

- online user registration, introducing a user to a library (free service)
- registry of digital masters to record digital content or the intent to digitise to prevent unnecessary duplication (free service)
- global library registry (free service)
- direct home delivery pilot

The National Library of Australia has also launched an ambitious program to provide an integrated national delivery service featuring direct user requesting and home delivery (Fitch 2007).

The delivery gap will not be solved just by making available new technical components. Libraries must make the necessary policy changes to cooperate on an international scale (Rethinking Resource Sharing Initiative 2006). The motivation of exposure has been to attract one's own user population to the resources of one's library. But exposure also attracts the users and potential users of other libraries as reported by the major Dutch university. By serving these external users (either directly or via their own library) a library's own users are better served by a broader and more robust international cooperative.

Conclusion

"In a pre-network world where information resources were relatively scarce and attention relatively abundant, users built their workflow around the library. In a networked world where information resources are relatively abundant and attention is relatively scarce, we cannot expect this to happen. Indeed the library needs to think about ways of building its resources around the user workflow. We cannot expect the user to come to the library web site any more." Dempsey (2006)

It doesn't matter at all if a user finds our OPAC through the "back door", i.e. by linking directly into a full record or holdings display. The more routes to the "back door", the better; that is, it is optimal to have multiple points of exposure. OCLC statistics show that once users find worldcat.org, they stay to "look around". Moreover, it isn't of paramount importance that our users appreciate our interface and learn how to do masterful searches. What ***is*** of paramount importance is that once our users actually find what they want, they should be able to get it and get it easily.

Now is the time to focus on delivery, yet most of our professional attention is still on discovery and the user interface. Attention is being placed on discovery at the expense of delivery. It's only digitisation that is currently getting the attention that it deserves and then the delivery process of digitised materials is assumed without examination. That is not to say that we shouldn't strive to improve our discovery experiences. We must be placing much more emphasis on improving the delivery experience. Just how many of the users

who are currently purchasing from Amazon would actually borrow or acquire from a library instead if it were just as easy?

References

Burnhill, Peter, Guy, Fred and Osborne, Nicola (2007) Scholarly Communication and National Union Catalogues: a Strategic Role for SUNCAT in the UK Information Environment. New Review of Information Networking 12: 1, 1-21 http://dx.doi.org/10.1080/13614570601133039

Dempsey, Lorcan (2006) The library catalogue in the new discovery environment: some thoughts *Ariadne*, 48 http://www.ariadne.ac.uk/issue48/dempsey

Fitch, Kent (2007) New paradigms for getting
http://www.nla.gov.au/initiatives/meetings/documents/New_paradigm_getting.ppt

Gatenby, Janifer (2006) Today's information consumer tapping into international library resources: making it a reality. http://www.oclcpica.org/content/1400/pdf/NVB_20061107.pdf

Larsen, Kirsten (2007) Bibliothek.dk in Google. http://conference.dbc.dk/viewpaper.php?id=10&cf=2

Libraries Australia Advisory Committee (2006) Report
http://www.nla.gov.au/librariesaustralia/documents/LAAC_Meeting_Papers.pdf

Nilges, Chip (2006) The online computer library center's own Open WorldCat program. *Library Trends* 54 (3): 430 – 447.

OCLC (2003) Environment scan; Pattern recognition http://www.oclc.org/reports/escan/default.htm

OCLC (2005) Perceptions of libraries and information resources
http://www.oclc.org/reports/2005perceptions.htm.

Pearce, Judith (2005) New Frameworks for Resource Discovery and Delivery
A paper presented in an earlier form at the Standards New Zealand / Standards Australia IT-19 Seminar, Technical Standards for Libraries and Education: Solutions and Emerging Frameworks, held at the National Library of New Zealand, Wellington, New Zealand, Wednesday 26 October 2005.

Quint, Barbara (2004) Google Scholar focuses on research – quality content.
http://newsbreaks.infotoday.com/nbreader.asp?ArticleID=16324

Request Transfer Message, a Community Profile for OpenURL.(2007) http://oclcpica.org/?id=1409&ln=uk

Respectfully quoted – a dictionary of quotations 1989 available at Bartleby.com great books online.
http://www.bartleby.com/73/257.html

Rethinking resource sharing initiative (2006) Manifesto
http://rethinkingresourcesharing.org/docs/manifesto.pdf

Acknowledgments

Chip Nilges, OCLC

Bill Brembeck, OCLC