

WorldCat Quality



An OCLC Report



OCLC[®]

The world's libraries.
Connected.[™]



WorldCat Quality

An OCLC Report

Principal contributors

Karen Calhoun, Vice President, WorldCat and Metadata Services

Glenn Patton, Director, WorldCat Quality Management

Copyright © 2011, OCLC Online Computer Library Center, Inc.
6565 Kilgour Place
Dublin, Ohio 43017-3395

ALL RIGHTS RESERVED. No part of this publication may be reproduced, stored in a retrieval system or transmitted, in any form or by any means, electronic, mechanical, photocopying or otherwise, without prior written permission of the copyright holder.

The following are trademarks and/or service marks of OCLC: Connexion, FirstSearch, OCLC, the OCLC logo, WorldCat, WorldCat.org, WorldCat Resource Sharing and “The World’s Libraries. Connected.”

Third-party product, service, business and other proprietary names are trademarks and/or service marks of their respective owners.

Printed in the United States of America

Cataloged in WorldCat on August 25, 2011
OCLC Control Number: 747819264

ISBN: 1-55653-435-3
978-1-55653-435-5

Table of Contents

	Page
Background	1
Factors Affecting WorldCat Quality Since 2008	3
Reimplementing the Duplicate Detection and Resolution Software	5
“Parallel Records”	7
Reproductions and Reprints	8
Holdings “Scatter”	9
Taking Action: GLIMIR (Global Library Manifestation Identifier)	11
Taking Action: Other Near-term Projects	14
Looking Ahead: Global and Local Data Quality	16
End Notes	18

Background

Recent OCLC research on what end users and librarians want from library online catalogs,¹ including WorldCat, suggests that both groups approach catalogs and catalog data purposefully. End users generally want to find and obtain needed information; librarians and staff generally want to carry out their work responsibilities using catalog data. The presence of real or perceived duplicate entries in the catalog—while better tolerated by end users than librarians²—impedes not only end users’ discovery and delivery of wanted information, but also efficient work practices for librarians and staff. For this reason, OCLC’s WorldCat quality program has for many years been centered on duplicates management.

Previously unpublished background materials from OCLC’s 2008 online catalogs study provide insight into OCLC members’ perceptions and satisfaction with WorldCat quality (including duplicates management) at that time.³ The survey gathered nearly 1,400 responses from academic, public and special libraries inside and outside North America.

Respondents rated their overall satisfaction with WorldCat data using a 10-point scale ranging from excellent (10) to poor (1). There were differences in satisfaction levels by region (figure 1). Three-fourths of the North American library respondents (76%) rated their satisfaction as 10, 9 or 8, compared to two-thirds of the respondents outside North America (65%).

Figure 1. Satisfaction with WorldCat data inside and outside North America, 2008

Rating	North America (n=869)	Outside North America (n=304)
10 (excellent)	20%	17%
9	24%	19%
8	32%	29%
7	15%	17%
6	5%	9%
5	2%	5%
4 or 3	2%	2%
2 or 1 (poor)	0%	0%
Don't know	0%	2%

Background

Respondents described why they were satisfied or not with WorldCat data. Those who rated their overall satisfaction as 10 or 9 commented favorably on the database's comprehensive coverage of bibliographic and holdings information; easy to use searching and access; and WorldCat's utility for meeting end-user and library needs. While those who rated their satisfaction 10 found little to criticize, 32% of those who gave 9 ratings also described problems, notably duplicates and the existence of minimal records (figure 2). These issues were a theme of the comments across all ratings.

Figure 2. Reasons for satisfaction with WorldCat		
	Percent of Respondents Who Rated 10 (Excellent) (n=106)	Percent of Respondents Who Rated 9 (n=148)
Database	72%	68%
<ul style="list-style-type: none"> -Good hit rate; find items -Quantity of bibliographic records -Holdings information -Accuracy of bibliographic records -Information 		
Access	47%	43%
<ul style="list-style-type: none"> -Easy to use -Searching -Quick -Reliable -Clear presentation 		
Used in Services	45%	32%
<ul style="list-style-type: none"> -Meets needs of librarians and patrons -Cataloging -ILL/resource sharing 		
Problems	3%	32%
<ul style="list-style-type: none"> -Searching problems -Minimal records -Duplicate records -Downtime 		

Survey results also begin to make clearer that users' perception of quality involves more than the quality of the data itself. How that data is used and presented by interface software can be just as crucial a factor in creating a positive experience for the user.

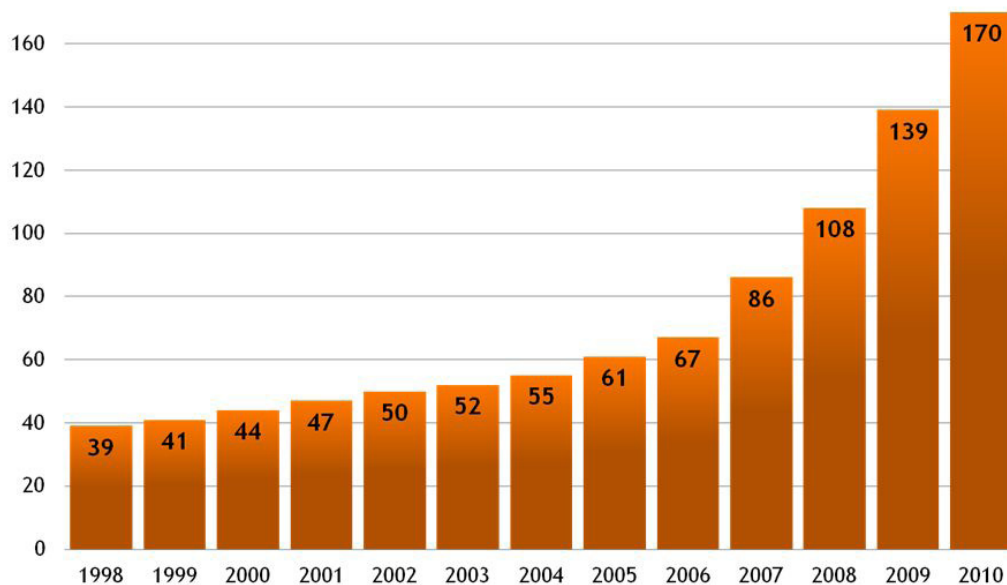
Factors Affecting WorldCat Quality Since 2008

Since 2008, the unprecedented growth of WorldCat has severely challenged OCLC's duplicates management methods and technologies (figure 3). This growth has exacerbated both the incidence and perception of duplicate entries in WorldCat for the same information content.⁴

This new metadata came into WorldCat due largely to agreements with national libraries and groups outside North America. As it entered WorldCat in large quantities, the new metadata:

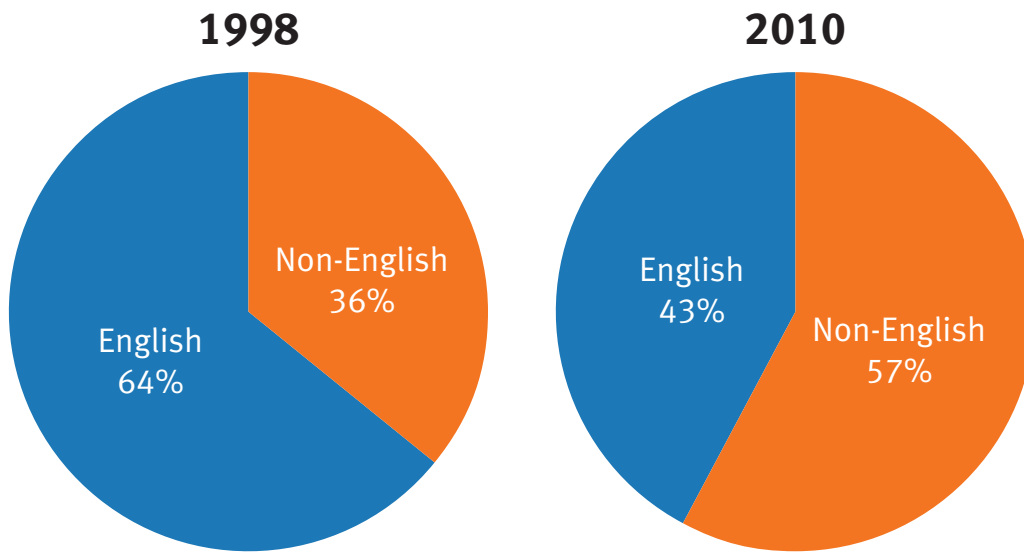
1. Overran OCLC's then-current automated tools for matching, merging and clustering records for the same work,⁵ and
2. Created new opportunities to build *multilingual* services around works, names and terminologies by substantially changing the proportion of non-English language content described in WorldCat (figure 4).

Figure 3: Growth of WorldCat bibliographic data 1998–2010 (millions of records)



Source: OCLC annual reports from FY1998 through FY2010

Figure 4: Change in non-English-language content in WorldCat, 1998 to 2010



Source: OCLC annual reports, FY1998 and FY2010

Another factor affecting WorldCat quality since 2008 is the loading of a larger number of vendor records for new and forthcoming titles (it is now estimated that 1.59% of the WorldCat bibliographic database represents vendor-supplied records). As libraries redesign their technical services workflows and make greater use of vendor metadata for selection, acquisitions and processing, OCLC has developed partnerships to load such vendor metadata into WorldCat. These brief vendor records do offer access to information about new materials much earlier in the publication process but, because they often contain very little data, these records may not be as useful to end users or librarians. They also pose problems for matching and merging processes.

A third factor in addressing issues with WorldCat quality is the growth in activity by the Expert Community. The Expert Community consists of OCLC member libraries that share in the maintenance of WorldCat through their participation in the OCLC Enhance program, and the Program for Cooperative Cataloging's BIBCO and CONSER programs, as well as their use of capabilities for upgrading and enriching bibliographic records to all cataloging users. These capabilities were significantly expanded in 2009 with the Expert Community Experiment so that cataloging users could make changes to almost all records in WorldCat. Activity by the Expert Community in FY2011 totaled just over 1 million transactions, nearly double the same activity in FY2008.⁶

Reimplementing the Duplicate Detection and Resolution Software

Beginning in 1991, OCLC used its Duplicate Detection and Resolution (DDR) software to match WorldCat bibliographic records in the books format against themselves to find and merge duplicates. By mid-2005 when WorldCat migrated to its new platform, 16 runs through WorldCat had been completed, resulting in the elimination of a total of 1.6 million duplicate records. In 2005, a project was started to reinvent the DDR software to work in WorldCat's new environment and to expand its capabilities to deal with not only records for books, but also records for continuing resources, scores, sound recordings, visual materials, maps and electronic resources.

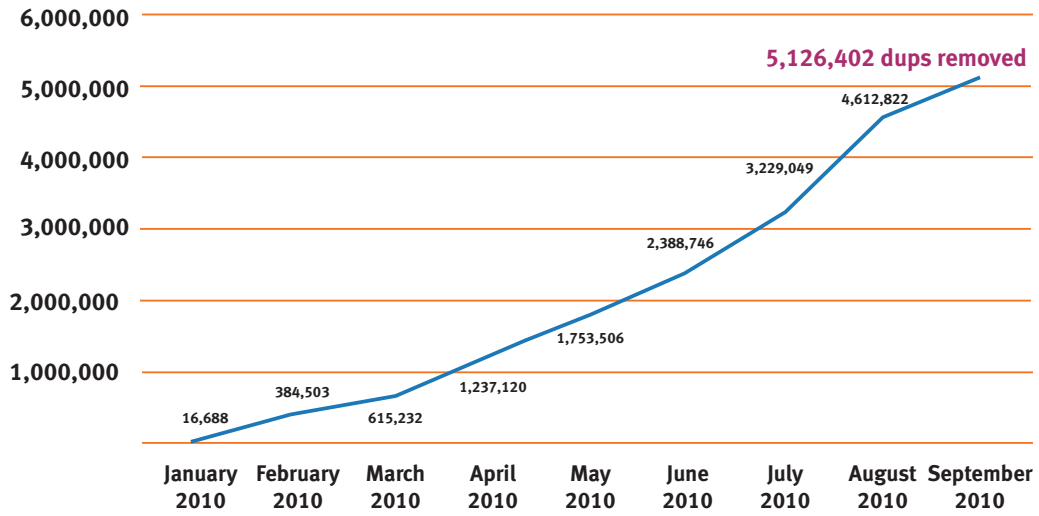
The DDR software takes a record and assembles a set of other WorldCat records that may represent the same resource. The DDR algorithms then compare this set of records (using significant portions of the data in each record) to determine if any of the records are duplicates. If the software determines that records are duplicates, it selects a record to be retained (using a hierarchy that considers the source of the record, its relative fullness and number of holdings) and selects appropriate data (such as call numbers and subject headings of types not already represented in the retained record) to merge into the retained record. All holdings (including detailed local holdings data) are moved to the retained record, and the OCLC number(s) of the duplicate records are recorded in field 019 of the retained record.

This large multiyear project reached a milestone in September 2010, when the DDR software completed a full pass through WorldCat, removing more than 5.1 million duplicates (figure 5). In addition, DDR has been incorporated into the processing of the daily journal files and batchloading. These processes continue to merge duplicates every day as new records are processed to WorldCat. As of April 30, 2011, over 7.5 million duplicates had been removed by DDR.

The new DDR software merges a large number of bibliographic records, and libraries will notice fewer duplicate records in WorldCat. Nevertheless, DDR is *not enough* to meet the challenge and complexity of duplicates management in today's WorldCat, a truly global, multilingual database that represents many languages of cataloging, many metadata traditions from within and outside libraries, and a growing array of the world's individual and union catalogs.

Reimplementing the Duplicate Detection and Resolution Software

Figure 5: DDR: Processing the WorldCat database, January–September 2010



“Parallel Records”

At the urging of OCLC members, in 2003 OCLC policy changed to allow for “parallel records” within WorldCat by language of cataloging (note: **not** the language of the content being described, but the language of cataloging used to describe that content). Previously, records for the same title, but cataloged in different languages, such as English, Spanish and French, were considered duplicate records. OCLC no longer considers these records duplicates, but considers them parallel records (figure 6). It has long been envisioned that WorldCat would need a parallel record structure to display records by language of cataloging, and with the evolution of WorldCat, that vision is now reality.

Figure 6 shows search result displays for Michael Buckland’s publication, *Redesigning Library Services*, in WorldCat.org, OCLC’s interface for end users. Intentionally, these displays provide only a few data elements to identify the title; as a result the metadata that would reveal these records’ different languages of cataloging is not apparent. However, looking at the underlying data (for example, through the OCLC interface for cataloging, Connexion) makes the differences apparent (cf. OCLC record numbers 25628636, 611064627, 464816186).

Figure 6: Example of parallel records for the same title cataloged in different languages

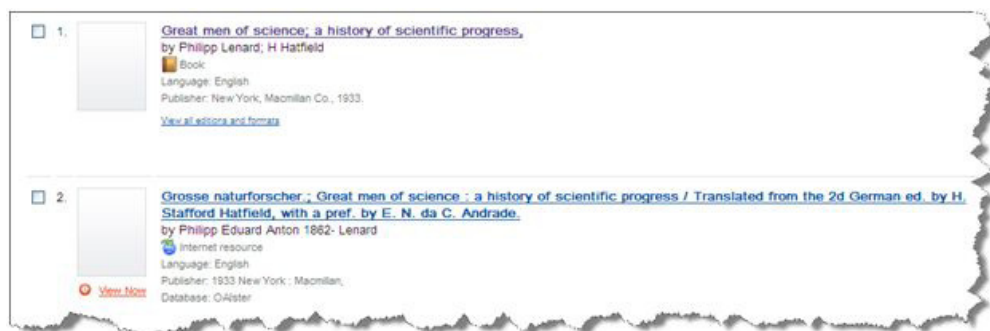


Reproductions and Reprints

Managing the appearance of duplicate entries for original works and their reproductions is another challenge for duplicates management in WorldCat, a database which must serve the searching and discovery needs of multiple constituencies among librarians and end users. These two groups may have quite different ideas of what is and is not a duplicate entry for the same title. For example, the original text of Philipp Lenard's 1933 English-language edition of *Great Men of Science*, published by Macmillan, was digitized and is now part of the HathiTrust Digital Library. The *Anglo-American Cataloguing Rules, 2nd ed.*, call for separate descriptions for the original and digital reproduction, and accordingly OCLC does not merge the records for such descriptions with the DDR software (figure 7). The situation is similar for reprints.

While the cataloging rules and their interpretations around reproductions and reprints are important for proper cataloging description and collection management, arguably the apparent duplicate entries provide at best a confusing experience for the end user of WorldCat.org—and at worst, the failure of the system to correctly guide the end user to the relevant location or appropriate link to the content.

Figure 7: Example of separate cataloging records for an original and its reproduction



Cf. OCLC record numbers 1394475, 296526894

Holdings “Scatter”

While the aggregation of WorldCat metadata describing library collections is unique, it is arguably the *holdings* (library locations) that represent the unique and inimitable value of WorldCat. It is important to be able to identify all holdings relevant to a search. For a cataloger, the ideal search result may be limited to a particular edition or format, while an end user is likely to be interested in any edition or format (regardless of language of cataloging) that is available to him or her (and there may be a preference for online content when it is available).

OCLC numbers, as well as being the foundation stones of the database, are used to tie the end user’s *discovery* experience to the *delivery* experience—in other words, within WorldCat, these numbers link from metadata *describing* particular content to information about the *location* of that content in a particular library (or on the Web). OCLC numbers are also used for direct record access into WorldCat as well as for direct linking to metadata or content in external catalog databases (when those catalogs or databases have ingested and indexed OCLC identifiers).

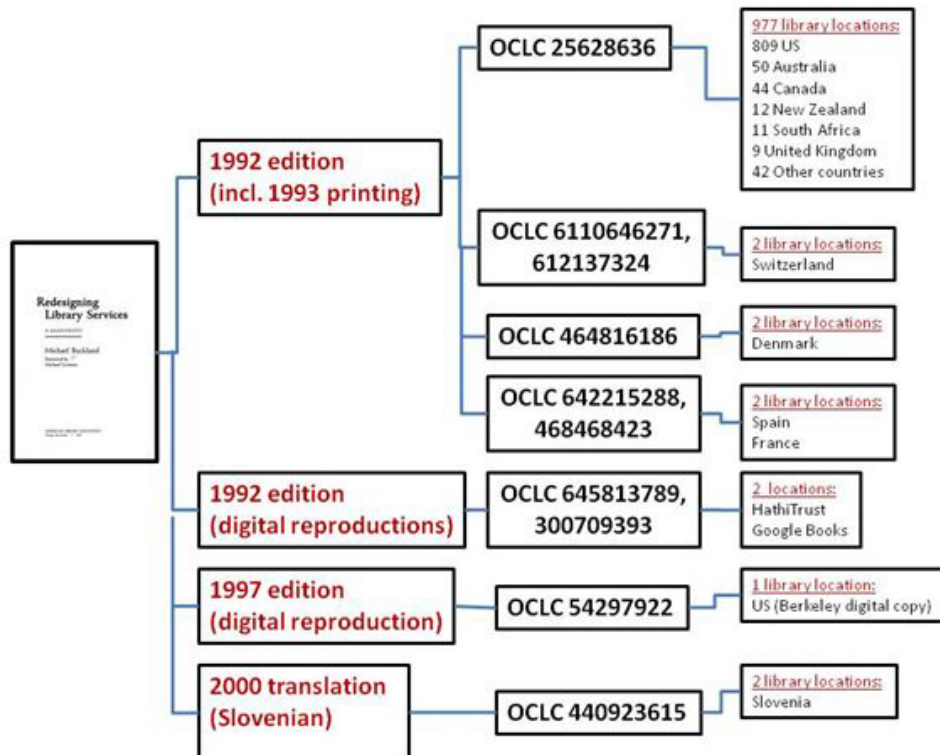
The rapid growth of WorldCat since 2008 has diluted the effectiveness of OCLC numbers for bringing coherence to holdings information, thereby supporting the visibility of, and linkages to, library assets represented in WorldCat. As multiple national bibliographies were loaded, more parallel records were introduced, and more records for reproductions (especially digital reproductions) came to be represented in WorldCat, holdings information became scattered across multiple records with different OCLC numbers.

Figure 8 provides an example of the “holdings scatter” now affecting the visibility and links to library catalogs for Michael Buckland’s *Redesigning Library Services*. As currently represented in WorldCat, holdings information for the 1992 English-language print edition is dispersed across six OCLC numbers that identify several parallel records (different languages of cataloging) plus a record for a 1993 printing. Two OCLC numbers identify digital reproductions of the 1992 edition from two sources (HathiTrust and Google Books), while another OCLC number identifies a digital reproduction of a (purported) 1997 edition. In nearly all cases, the records bearing these OCLC numbers display as separate search results in WorldCat interfaces. OCLC has implemented FRBR clustering⁷ in WorldCat.org to mitigate some of the confusion. Nevertheless, the end user’s success in finding the most relevant library holdings or the link to the desired content depends on selecting the “right” entry in the search results. For example, if a North American user selected the entry representing OCLC record number 25628636, he or she would find a wide range of library holdings. Selection of any of the other entries describing the 1992 edition could lead that user to believe that there were no nearby libraries holding the item.

Holdings “Scatter”

It should be noted that the current situation also affects “social bibliographic data.” Such data (reviews, lists and recommendation systems) is now dispersed across multiple search results. In addition, the current situation erodes the effectiveness of resource sharing and interlibrary data exchange as well as cooperative collection analysis and management.

Figure 8: Distribution of holdings in WorldCat for the title Redesigning Library Services



Taking Action: GLIMIR

(Global Library Manifestation Identifier)

As described previously, OCLC's DDR program has helped to resolve true duplicates in WorldCat. Today, however, WorldCat quality needs to be strengthened through the provision of a unique common identifier for multiple metadata records where the duplication is not only **intentional**, but also in keeping with cataloging standards and OCLC policy for parallel records. The days of reliance on just one master record in WorldCat for a given publication are now over, and it is time for a new approach to clustering metadata and holdings information.

GLIMIR is an OCLC project begun in 2009 to address the problems described for OCLC numbers and the consequential dispersion of metadata and holdings information across multiple records. In this project new GLIMIR identifiers are affording the means to cluster various records (describing linguistic, format-specific or other variants) that describe given information content into a consolidated cluster with its own GLIMIR identifier and cross-referenced OCLC numbers. GLIMIR's impact, once deployed in WorldCat's various interfaces for end users and libraries, will be to:

- Enable easier-to-use, more intuitive search result displays for end users and librarians
- Lower the processing costs of acquisitions, cataloging, selection and resource sharing by making it easier to identify and select the "correct" record
- Improve linking from WorldCat.org to library catalogs and smooth the flow of data among WorldCat and library systems locally and globally
- Increase the accuracy of FRBR clustering in WorldCat "work sets"
- Improve the visibility of significant differences among manifestations within work sets, e.g., original versus revisions, translations, audio versions
- Enable the sharing of enriched content by making it available for all records within a cluster
- Facilitate anchoring and disclosing library assets on the Web
- Provide a new identifier that can be used to cluster variant records in other systems
- Support more authoritative WorldCat statistics that would clarify the number of unique resources made available to libraries worldwide.

Figure 9 is a *pro forma* reorganization of the records in the *Redesigning Library Services* set of records from figure 8 (arranged as a work set and shown in red). Figure 9 goes one step further to call out the records related to the 1992 edition and its digital reproductions (arranged as a GLIMIR cluster and shown in blue). The better organization of data in GLIMIR clustering will support more intuitive WorldCat displays; the clustering of related holdings; data exchange, aggregation and

Taking Action: GLIMIR

syndication; and improved linking between WorldCat, library catalogs, content and other data services on the Web.

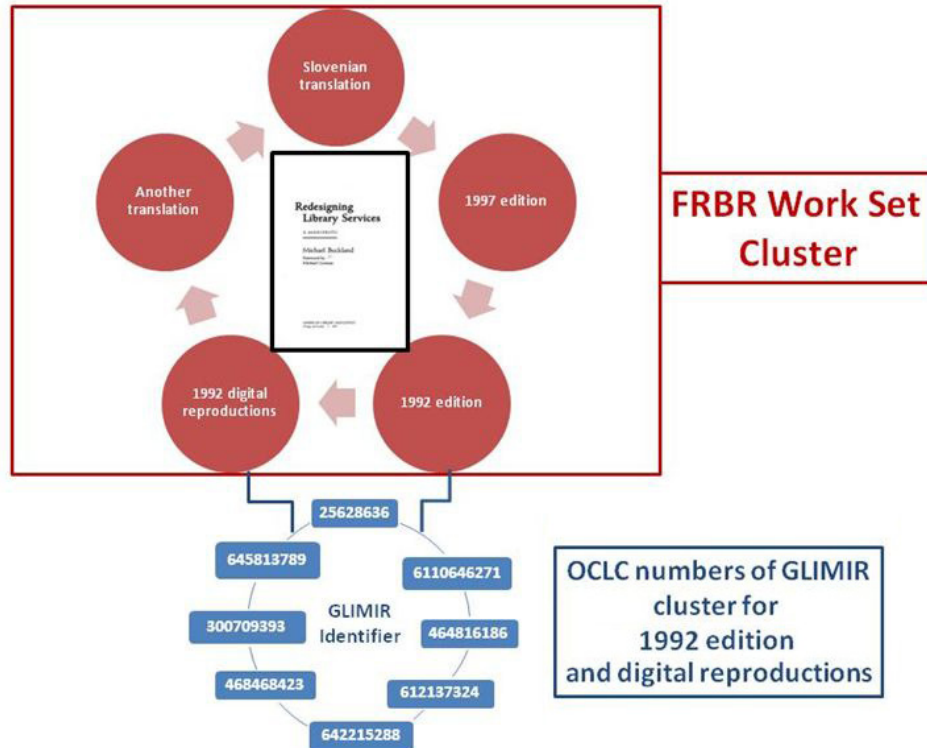
Principal Benefits of GLIMIR:

- GLIMIR clusters will improve the accuracy of FRBR work sets and reduce the appearance of exact duplicate records.
- GLIMIR clusters will improve the clustering of holdings information, supporting more reliable linking from WorldCat to library catalogs.

Similar clustering will also improve the display of other cases in which the same content is presented in multiple physical forms. Examples include sound recording of the same musical performance presented as LPs, cassettes, CDs, MP3 files and streaming audio, and the same motion picture presented in its original film version as well as on videotape, DVD and as streaming video.

As of this writing, the GLIMIR infrastructure is nearing completion. In early FY2012, WorldCat will be “GLIMIRized”; because WorldCat is so large this may take some months to accomplish. When enough GLIMIR clusters are ready, users of WorldCat Local and WorldCat.org (OCLC’s end-user interfaces) will begin to see improvements in searching, displays and linking to holdings, library catalogs and online content. Later in the fiscal year, GLIMIR clusters will be incorporated into Connexion (OCLC’s online cataloging interface) to reduce the time required for catalogers to choose the right record and to provide a better holdings picture for the manifestations of a work. This work in Connexion should also improve the clustering (if not the deduplication) of sparse vendor records, saving catalogers’ time and effort.

Figure 9: Pro forma illustration of the impact of GLIMIR on WorldCat records and holdings for Redesigning Library Services



Taking Action: Other Near-term Projects

The following table describes other WorldCat quality improvements scheduled for work in FY2012. These activities strive to strike a balance between improving the existing metadata and its arrangement and enriching the metadata and its external linkages. In addition, the FY2012 program begins to address issues with the WorldCat Registry and further expands the Expert Community.

<i>Scheduled WorldCat quality activities for FY2012</i>		
Action	Impact	Comments
Repeat WorldCat quality satisfaction survey	Validate relevance and priority of FY2012 data quality program; identify new activities and enhancements.	Compare to baseline results of 2008 survey.
Enhance DDR	DDR and batchload matching become more flexible and better able to deal with variations in incoming data caused by differences in cataloging practice, incorrect coding, etc. Fewer exact duplicate records are introduced to WorldCat.	At present DDR is cautious taking into account small differences between records. Test and enhance DDR to enable more liberal matching when warranted.
Reduce missing or broken links to library catalogs and/or union catalogs	Searching by the OCLC number associated with a set of merged records automatically links to the matching record in the external catalog.	When records are merged by DDR, the OCLC numbers for the records that are not retained are stored in the 019 field of the retained WorldCat record. The holdings from the merged records are also moved to the retained record. In most cases, the deep links from WorldCat.org to an institution's catalog for that merged record fail because the institution still has the original OCLC number in its catalog and the catalog has no "awareness" of the OCLC number of the retained record. The project will enable better linking from WorldCat to the library catalog using numbers from the 019 field.
Resolve issues with sparse vendor records	Supply improved metadata for use in WorldCat and library systems; save catalogers' time and effort.	Engage with selected vendors to improve and align OCLC/vendor metadata workflows so that better records show up sooner, replacing the very brief records that support selection and ordering, but that are inadequate for cataloging.

<i>Scheduled WorldCat quality activities for FY2012</i>		
<i>Action</i>	<i>Impact</i>	<i>Comments</i>
Link and update selected headings in WorldCat records to headings in LC name and subject authority files	Increase the reliability and accuracy of searching and displays for authors and topics in WorldCat interfaces; improve quality of headings in records exported to systems serving English-speaking communities.	Headings in records will be automatically updated, substantially reducing the need for catalogers to manually control individual headings.
Enrich WorldCat records with summaries, tables-of-contents notes, call numbers and other enhanced data	Improve end-user and librarian ability to assess the utility of items in a search result set and decide which ones merit taking the time to obtain.	Run data enrichment software on all of WorldCat and newly added/changed WorldCat records.
Expand community involvement in the maintenance of WorldCat	Increase the pool of libraries eligible to edit and correct records contributed via the Program for Cooperative Cataloging (i.e., PCC BIBCO records).	Expand the Expert Community program by allowing editing of PCC BIBCO records by OCLC NACO institutions.
Provide a tool set for quality activities in the WorldCat Registry	Improve the usefulness of institution data in the Registry.	Provide reporting to assist in maintenance of institution data.

Looking Ahead: Global and Local Data Quality

OCLC's emerging plans for continuously enhancing WorldCat quality rest on a commitment to local, regional, national and global knowledge organization for libraries across multiple material types (physical, electronic, digital). It will be necessary to reinvent OCLC's long-standing and successful, but English-language-centric approaches to metadata creation and data quality management for the realities of the increasingly multilingual, multinational OCLC cooperative. An inclusive approach that meets the data quality requirements of **both end users and librarians** will be critical to the future viability and appeal of WorldCat.

The WorldCat quality roadmap beyond the next 12 months is understandably less defined but includes the following principal elements:

- Further advances in duplicates management within and outside OCLC systems
- Further work to improve the management and display of holdings information
- Extension of GLIMIR to enable new data services for internal and external use (this work will involve further optimizing FRBR and GLIMIR and making GLIMIR identifiers available for wider use)
- WorldCat quality projects and new data services around authority and terminology files used outside North America
- Further expansion of the Expert Community (including more participation by libraries outside North America)
- Deployment of multilingual data from the 23rd edition of the Dewey Decimal Classification (DDC) and mapped terminology sets in new outward-facing Web services
- Dynamic management of record descriptions in different languages using algorithms, DDC and authority files.

Together with the growing pains associated with the rising global relevance of WorldCat since 2008, there are new opportunities. As mentioned early in this paper, the rapid growth of WorldCat has substantially changed the proportion of non-English language content and metadata represented in the database, thereby creating new options for multilingual services around works, names and terminologies.

A glance at a partial list of the linguistically diverse subject headings associated with this paper's sample title, *Redesigning Library Services*, reveals the presence of the necessary multilingual data (figure 10). Further, the development of WorldCat Identities and VIAF⁸ has captured different language forms of author and creator names. It is now a matter of deploying this linguistically diverse data in a coordinated WorldCat quality program to support localized, Web-based end-user and library services in which not only the interface, but the data itself is offered in the preferred language.

Figure 10: Linguistically diverse subject headings describing the topics of Redesigning Library Services



It will be necessary to reinvent OCLC's long-standing and successful, but English-language-centric approaches to metadata creation and data quality management for the realities of the increasingly multilingual, multinational OCLC cooperative.

Endnotes

1. Calhoun, Karen, and Diane Cellentani. 2009. *Online catalogs: what users and librarians want: an OCLC report*. Dublin, Ohio: OCLC.

2. Calhoun, Karen, and Diane Cellentani. 2009. *Online catalogs: what users and librarians want: an OCLC report: synopsis*. Dublin, Ohio: OCLC, page 14. Available from: <http://www.oclc.org/reports/onlinecatalogs/> .

3. Unpublished background materials prepared for *Online catalogs: what users and librarians want*. 2008.

4. The scope of this white paper is the database of WorldCat bibliographic records and library holdings that are for the most part contributed by OCLC members, national libraries and consortia, and some publishers and vendors. The rapidly growing number of metadata records for articles, e-books, digitized public domain content and some other digital content that is indexed for WorldCat.org discovery is not within scope for this paper. Admittedly the expansion of WorldCat.org to include such metadata has contributed to the perception that WorldCat.org is “noisy.” OCLC’s WorldCat quality staff are aware of the discovery, linking and delivery issues associated with this new metadata in WorldCat.org, but the issues are not dealt with in this paper.

5. FRBR ([Functional Requirements for Bibliographic Records](#)) is a 1998 recommendation of the International Federation of Library Associations and Institutions (IFLA) to restructure catalog databases to reflect the conceptual structure of information resources. The FRBR model brings together bibliographic records that are intellectually related as “works.” Having resources brought together under the “works” umbrella enables users to sift through the myriad resources available as variant texts, digitized copies, arrangements, translations and so on. Works clustering helps end users obtain the work, or content, that they are looking for, irrespective of the specific “container” in which the content is carried. In large databases such as WorldCat, such collocation is indispensable for discovery, navigation and cost-effective library processing. OCLC has “FRBRized” WorldCat.org using the OCLC FRBR Work-Set Algorithm.

6. More information about the Expert Community, the OCLC Enhance program and other updating and enrichment capabilities can be found at <http://www.oclc.org/us/en/worldcat/catalog/quality/>. Information about the Program for Cooperative Cataloging and the BIBCO and CONSER programs is available at <http://www.loc.gov/catdir/pcc/>.

7. The OCLC FRBR Work-Set Algorithm collects WorldCat records into groups based on author and title information from metadata describing particular titles and authors. Author names and titles are normalized to construct a key. All records with the same key are grouped together into a “FRBR work set.”

8. WorldCat Identities compiles information data-mined from WorldCat and other resources to illustrate persons and corporate bodies represented in WorldCat. Included is such information as variant names by which these persons and corporate bodies are known, a publication timeline for works by and about them, their roles in relation to the works, etc. Each person or corporate body, currently numbering about 30 million, is represented by a separate Identity page. These Identity pages are incorporated into the WorldCat.org interface and are also available at <http://worldcat.org/identities>. VIAF (the Virtual International Authority File) is a joint project with the Library of Congress, the Deutsche Nationalbibliothek and the Bibliothèque nationale de France, in cooperation with an expanding number of other national libraries and other agencies. It combines the name authority files of these participating organizations into a single name authority service (<http://viaf.org/>).



OCLC[®]

6565 Kilgour Place
Dublin, Ohio 43017-3395
1-800-848-5878 +1-614-764-6000
Fax: +1-614-764-6096
www.oclc.org

ISO 9001 Certified

ISBN: 1-55653-435-3
978-1-55653-435-5