November 2006

Reflections on the OCLC Worldcat Collection Analysis Tool: We Still Need the Next Step

Darby Orcutt

North Carolina State University, darby_orcutt@ncsu.edu

Tracey Powell

University of North Carolina-Chapel Hill, tracy.n.powell@gmail.com

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

Recommended Citation

Orcutt, Darby and Powell, Tracey (2006) "Reflections on the OCLC Worldcat Collection Analysis Tool: We Still Need the Next Step," *Against the Grain*: Vol. 18: Iss. 5, Article 10.

DOI: <https://doi.org/10.7771/2380-176X.5010>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Reflections on the OCLC WorldCat Collection Analysis Tool: We Still Need the Next Step

by **Darby Orcutt** (Senior Collection Manager for Humanities and Social Sciences, North Carolina State University)
<Darby_Orcutt@ncsu.edu>

and **Tracy Powell** (Master of Science in Library Science Student, UNC-Chapel Hill) <tracy.n.powell@gmail.com>

Introduction

In March 2005, the **Online Computer Library Center (OCLC)** announced the **WorldCat Collection Analysis Tool (WCA)** as a means for libraries to "compare their collections with those of peer libraries, and compare, as a group, the level of overlap or uniqueness of their collections."¹ At **North Carolina State University (NCSU)** we logged hundreds of hours over more than six months on a variety of analysis projects using **WCA**. Projects included comparing **North Carolina State University (NCSU) Libraries'** holdings against groups of other research universities in selected fields; comparing holdings across broad subject areas within our local **Triangle Research Library Network (TRLN)** consortium; identifying unique holdings versus the "**Google Five**"² as possibilities for digitization; and conducting (or at least attempting to conduct) a research project seeking to quantify minimum core collections needed at a university level for support of diverse disciplinary programs.

Among our other observations, we found that **WCA**, implicit to its design, privileges institution-level data, much better equipping librarians to conduct old-fashioned, single-institution analyses rather than the cross-institutional comparisons needed for the cooperative decision-making increasingly desired in practice. In this article we will examine **WCA's** abilities and limitations before turning to an investigation of the capabilities which could make it a truly useful tool for twenty-first century collection analysis.

Abilities of the Tool

WCA enables several kinds of quick data-gathering. It provides easily-accessed **WorldCat** data limited to the user's home institution (as defined by a single **OCLC** hold-

ings symbol) or with reference to a group of two to five other libraries. It collates the holdings of these comparison schools to produce data regarding overlap and uniqueness, both within the comparison group and as compared to the home institution. This function helped us to conduct quick assessments of the relative depth of collections in broad subject areas among institutions in our consortium, and allowed for the speedy production of charts and other marketing data geared toward researchers and administrators.

WCA greatly facilitates the examination of general trends, at least at larger scales where room exists for margins of error. In analyzing our own as well as **TRLN** holdings, we were able to conduct analyses of age of collections by subject area. This allowed us to generally assess the accuracy of collection assumptions and adherence to collection policies, as well as to see how collecting patterns had changed over time. Interestingly, some legendary "bad budget years" and "times of plenty" could be easily recognized when looking at size of collections by year of imprint.

The ability to quickly develop and export title lists into spreadsheet programs also proved very helpful. For example, we used **WCA** to examine our holdings in public history versus a group of four other institutions. After identifying relevant subject areas, we could readily compare the size of our collection against a calculated mean number of titles in each area. We could also produce title lists for potential acquisitions that, when compared to our catalog, identified items commonly held by other institutions but not our own. While the tool did not eliminate the work of manually assessing title lists in order to eliminate titles that were out-of-date, out-of-scope, or

held in other editions, it did quickly help narrow our focus to titles of likely interest. Sampling and manual analysis of other such lists across all subjects allowed us to identify potential areas and estimate numbers of titles of unique holdings (versus the "**Google Five**") for potential digitization projects. In each of these cases, **WCA** provided an easy, quick-and-dirty starting point, even though the work required subsequent manipulation and informed human analysis.

As noted by **Phyllis Spies**, Vice-President of **OCLC Collection Management Services**, the tool "leverages the cooperative effort of thousands of librarians around the world who have built — and are continuing to build — **WorldCat**."³ With more than one billion holdings records, **WorldCat** offers the largest and most comprehensive shared library catalog.

Limitations of the Tool

As any user of **WorldCat** knows, the data it contains is far from perfect, and even if perfect by the rules of **AACR2**, may not conform to certain needs. Thanks to differences in cataloging choices and practices, many bibliographically-like editions are represented by multiple records, each with its own **OCLC** accession number. Since **WCA** matches on accession number only, comparison studies over report uniqueness and underreport overlap. In addition, **AACR2** requires multiple records for many different editions of like titles, although in many collection situations, edition may not matter. The net effect of these two factors (cataloging practice and rules) should neither be presumed uniform nor underestimated. In the process of conducting the potential digitization project noted above (comparing to the "**Google Five**"), we discovered a much higher degree of consistency between **WCA** title lists and the libraries' **Online Public Access Catalogs (OPACs)** in the sciences than in the social sciences, and dramatically less consistency in the humanities. In fact, across the subject area of Language & Literature (broad **LC** class P), only about 20% of titles reported as unique at the title level truly were.

When comparing **WorldCat** listings for our own collection against our **Endeca**-driven **OPAC**,⁴ we found consistent underreporting among **WCA** data. While some of this is attributable to differences in the pace of cataloging and updates between the tools, the degree of difference presents cause for caution when utilizing **WorldCat** data as the basis for institutional analysis. Several subcategories in sociology, for example, show differences of more than 1,000 titles between **NCSU Libraries** and **WorldCat** records — a sizeable percentage of difference within a set of 6,000 titles. Other

continued on page 46

Changing the way Libraries ... from page 42

change in language has the possibility of producing startling results.

By approaching teaching faculty bearing lists of titles, data on usage, charts showing the decline in periodical use, and asking them to tell us what they think we should get rid of, without clearly focusing on what resources the library may already provide online or asking what students truly need to succeed at our institutions, we have inadvertently fostered a sense that the scholarly literature so valued by

academia is less useful and less worthwhile in the eyes of the increasingly digitally oriented library. In order to reverse that perception, and instead foster a culture in which teaching faculty understand that the libraries are striving to provide the best, most vital and most relevant collection of print and online periodicals possible, we must ask new questions. We must probe for honest answers and reactions rather than cursory and negative responses to the same questions we've been asking for years. Our budget management, our collections, our relationships, and most of all our relevance to our users will benefit from a new approach. 🐭

subcategories generally exhibited a similar rate of 10-20% under representation in **WCA**. As we discuss in the next section, there is reason to believe that **WorldCat**'s title numbers in other ways actually over represent titles in some listings, further exacerbating this difference between **WCA** numbers and catalog records.

Currency of **WCA** data is also an issue. The tool relies on an extract from **WorldCat** taken once per quarter (theoretically, at least; the second quarter 2006 update never occurred). For more current imprints, the frequency of updates precludes tenable comparisons. For example, in Fall 2005, we checked titles of 2005 imprints against our own **OPAC** and those of the other institutions and found in every case (across comparisons and subject areas) that the majority of unique titles reported in **WCA** were apparently no longer unique. For large and/or retrospective analyses, currency may be moot, as even tens of thousands of records likely would not radically alter the complexion of a large research library's holdings (10,000 out of 1 million records equates to 1%), and few libraries would routinely catalog exceedingly large numbers of older volumes.

Other limitations include sorting by language within an institution (for example, **NCSU**'s **Urdu** collections cannot be analyzed with this tool), and which subject divisions may be selected for closer inspection. Both of these limitations hamper the tool's ability to assess diverse and interdisciplinary collections. More (or, better yet, user-delimited) categories would better address the needs of research-level institutions.

Sampling issues

One of the largest issues we encountered in using **WCA** is its inability to accommodate sampling methods. A project examining the existence and size of core collections shared by **ARL**-ranked institutions presented the greatest difficulties for **WCA**. For this analysis, we identified for comparison the five top- and five bottom-ranked **ARL** institutions within each of the three categories of "land grant," "public," and "private" universities. We could easily review the size of shared collections using the numerical data by subject area readily accessible in **WCA**, but encountered difficulties when we attempted to sample titles in six selected subject areas.

Several things called into question the reliability of the data we had gathered. Obtaining workable data required that we gather data within certain limits. First, we needed to limit our analysis to books alone, as other format types presented difficulties, including especially high rates of multiple **OCLC** accession numbers for like items and apparently lower rates of institutional reporting to **OCLC** of video and other formats. We also had to exclude titles with imprint dates within the last two years in order to account for differences in cataloging and acquisition rates across institutions. **WCA** is therefore not the tool for comparing very recent acquisitions.

against the grain people profile

Tracy Powell

M.S.L.S Student
The University of North Carolina at Chapel Hill
<tracy.n.powell@gmail.com>

BORN & LIVED: Born: Louisville, Ohio. I've also lived in Minnesota, Massachusetts, North Carolina, Indiana, and Georgia.

EDUCATION: B.A., **Macalester College**, St. Paul, MN; M.T.S., **Harvard University**, Cambridge, MA; in progress: M.S.L.S., **The University of North Carolina at Chapel Hill**.

PROFESSIONAL CAREER AND ACTIVITIES: I've worked as a Graduate Assistant in Collection Development at **North Carolina State University** and as a Research Assistant for faculty in the School of Information and Library Sciences at **UNC-Chapel Hill**; I am currently working as a Reference Intern in Perkins Library at **Duke University**.

IN MY SPARE TIME I LIKE TO: Hike; write; read; and visit parts unknown.

GOAL I HOPE TO ACHIEVE FIVE YEARS FROM NOW:

To be working in a combination Reference/Collection Development capacity and contributing to new developments in academic librarianship.

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS:

Working out the proper balance between different technologies, paper and digital, and developing ever-greater levels of sophistication in being able to respond to user information needs via catalog & database options as well as human-to-human interaction.

A much larger problem that derailed our core collections research project consists of a "bug" in **WCA** that appears in many title lists. In one instance, the title list repeated the same titles with the same **OCLC** numbers at multiple places in the list. Fortunately, our sampling caught this occurrence and alerted us to the problem, but sampling techniques might not reveal such repetition. Others later reported the same issue via **OCLC**'s listserv for **WCA** users (OCLC-WCA@OCLC.ORG).⁵

Combined, these problems limit the usefulness of **WCA** as a research tool. The accuracy of the data and its ability to handle increasingly complex cataloging records and collection decisions in a sophisticated, user-friendly way are necessary if **OCLC** truly wishes the tool to provide the type of information which could result in more informed decision-making on the part of collection managers.

Embedded Self-Centeredness

The implicit institution-centeredness of **WCA** rears its head in several ways. The tool allows one to create a group of two to five institutions for comparison, or to compare against any one of several predefined groups. Each of our studies utilized individually-selected institutions except for the comparison with the "Google Five," which exists in **WCA** as a predefined group. The tool only allows for comparison of the user's institution to the aggregate "collection" of the group of institutions selected (or mere analysis of that aggregate). While this might be a sufficient default for simple com-

parison work, it does not allow for more complex comparisons. For example, in examining the holdings of various **ARL**-ranked institutions we sought to include **NCSU** as one of the schools in our group of five so that we could directly examine shared holdings among all of the schools, rather than a comparison between **NCSU** and the other four. **WCA** does not allow this flexibility, which necessitated some awkward calculations in order to determine overall overlap and uniqueness in ways readily provided for groups of other institutions. This need to identify one's home institution as the basis for comparison suggests a competitive rather than cooperative stance as the implied norm for **WCA** comparisons. It also hampers the potential of conducting more general collections research.

Analyzing one's own institution within a context of cooperation, such as our own **TRLN** consortium, requires a wholly separate (and additionally paid) "group" subscription to **WCA**. In group accounts, the self-focus (this time, on the group) prevents even comparison against peer consortia, let alone analyses of cooperation across consortia. Unless "group" subscriptions were to include all relevant institutions, one could not analyze a collection's place within, for example, a geographic region, which is but one alternative context which selectors might find compelling.

Not Learning from the Past

Most of the problems and unfulfilled poten-

continued on page 48

tialities of the WCA mirrored those identified previously by Lucy E. Lyons with regard to WCA's predecessor, OCLC's Automated Collection Assessment and Analysis Services (ACAS). Lyons noted the necessity of performing manual sorts in order to make ACAS's categories meaningful to local institutions.⁶ This holds true for the WCA as well; in many of our analyses we found it necessary to disregard holdings in business, medicine, and government documents in order to account for the different ways in which institutions catalog these collections. Furthermore, the WCA subject categories remain inflexible and often less than useful, such as when philosophy is conflated with religion.

At the decision-making level, however, several limitations prevent WCA from being useful as a research tool for large collection management decisions. Lyons noted that ACAS "presents a huge slab of marble from which to sculpt, with the correct tool, an impressionistic (at worst) or realist (at best) shape of monograph collections."⁷ Lyons's analysis included examinations of collection size, age, growth, title overlap, and uniqueness, all capabilities touted by the WCA, despite no significant improvement over ACAS. Our experience with WCA suggests to us that its broad release was premature, and that the tool ultimately leaves tremendous room for improvement in meeting current and emerging assessment needs.

Where to Go From Here

No librarian can deny the competitive component of some collection analyses. Institutions must often pit themselves against their peers in

obtaining grants and other funding, attracting faculty, building a reputation in a field, and so on. Yet, a narrative of cooperation undergirds modern librarianship, and grows increasingly vital in these times, when ratios of collection budgets to potential acquisitions seem particularly vexing. Even aside from organized cooperation, collection managers must consider the greater contexts within which they are developing collections.

The inadequacies pointed out about this tool, as well as those previously reported with regard to ACAS, need to be taken seriously. The sorts of analyses librarians need for print monograph collections should already be within the reach of a tool such as WCA. Furthermore, as electronic collections grow, newer issues even now arise. As WorldCat now stands, no database built upon its foundation would likely answer many of the questions posed regarding networked resources, including issues of "ownership" versus access, relationship to physical versions, and how to quantify electronic content for comparison in meaningful ways (neither "file size" nor "number of volumes" is sufficient). AACR2 itself lacks a vocabulary for addressing the licensing terms vital to understanding e-collections.

The particular challenges of electronic resources aside, we need a tool robust enough to help selectors consider the multiple concentric and/or overlapping circles of ownership and access. Especially in these times of lower print materials budgets and relatively greater bibliographic visibility, the question of "what do we own?" often precedes such questions as "what does my consortium provide?" and "what is available in my region?" Hopefully, librarians do not in vain anticipate

more flexible and dynamic tools to evaluate collections across institutional boundaries. Building superior research collections in today's world, let alone tomorrow's, already demands as much.

Endnotes

1. "WorldCat Collection Analysis Tool Now Available," 11 March 2005. Retrieved 13 July 2006 from <http://www.oclc.org/news/releases/20054.htm>.
2. The "Google Five" are the first universities which agreed to have the holdings of their libraries scanned for Google's Book Search project. They include Harvard University, the University of Michigan, the New York Public Library, Stanford University, and Oxford University. Since this research was conducted, the University of California has also joined this project. "Google Book Search Library Partners." Retrieved 5 September 2006 from <http://books.google.com/googlebooks/partners.html>.
3. "WorldCat Collection Analysis Tool Now Available," 11 March 2005. Retrieved 13 July 2006 from <http://www.oclc.org/news/releases/20054.htm>.
4. "Endeca at the NCSU Libraries," 13 January 2006. Retrieved 3 August 2006 from <http://www.lib.ncsu.edu/endeca/>.
5. Subscription information for this LISTSERV is available at: <https://www3.oclc.org/app/listserv/>. Since this writing, OCLC has reported via the list that the three "bugs" causing this problem have been fixed, but this has not yet been tested.
6. Lucy E. Lyons, "A Critical Examination of the Assessment Analysis Capabilities of OCLC ACAS," *The Journal of Academic Librarianship* (2005), 31(6), 506-516.
7. Ibid., p. 516.

Legacy Government Documents Collections in the Digital Age

by Brian W. Rossmann (Government Information Specialist, Montana State University)
<brossmann@montana.edu>

In the summer of 2004 the Dean of Libraries at Montana State University formed a Government Information Futures Task Force to thoughtfully consider the future of government information in the Renne Library, a selective depository. The overarching working assumption was that government information is and will continue to be a valuable and critical information resource. However, given the dramatic changes taking place with how government information is distributed by the government to both libraries and the public, along with physical space constraints in the Renne Library, and the evolving expectations of our students and faculty, it was felt that a serious assessment of our documents collection was in order to evaluate how best to meet the needs of our current

and future patrons. This paper details some of the Task Force's findings and recommendations that may be useful to other libraries engaging in a similar exercise.

Collections of government documents in libraries have always had unique characteristics which distinguish them from everything else in a library's collection. They are typically shelved together, separate from the rest of the collection. Usually they are classified according to the Superintendent of Documents Classification System (SuDocs), which classifies by issuing government agency rather than by subject. Additionally, many libraries have separated documents into their own departments within the library, complete with unique reference desks and specialized documents librarians, and in some cases, their own circulation desks. All too often, large portions of documents collections are not represented in a library's OPAC,

or if they are, they may only be cataloged at the series level. Consequently, assessing a government documents collection presents its own unique set of challenges. This situation is complicated by the fact that currently most government information is being made available to libraries only in a digital format, and very few tangible items are being added to legacy collections. This proliferation of electronic government documents, which are stored on disparate servers throughout the government, means that a library's tangible document collection itself is likely to cease growing. Many of the titles taking up precious space on the shelves could be replaced by simply providing access to copies freely available online: not an easy decision because one always wonders if the online copy will continue to remain readily available. It is, and will continue to be, impos-

continued on page 50