Some findings about selection in WorldCat Local

Arnold Arcolio
User Researcher
User Experience Group
Global Engineering Division
OCLC

I'm Arnold Arcolio, a User Researcher in the User Experience Group in the Global Engineering Division of OCLC. It's a new group, about a year old, and it includes OCLC's long-established usability lab. I'm involved with product development, and in particular with WorldCat Local.

Over the past 2 years or so, through the OCLC Usability lab, my colleague Mike Prasse and I have conducted ten tests of WorldCat Local.
Some of these were large tests, some small; some conducted on site, some remotely. Most were with Academic libraries. Tests in Illinois public libraries are going on now. Tests in Illinois academic libraries are planned for July. We do these tests in collaboration with staff at pilot sites who are often very involved in planning the tests and analyzing the findings with us. Also, we pay attention to tests others have conducted—including, recently, tests at MIT and Western Washington University.

University of Washington
University of California
Berkeley
Davis
Irvine
The Ohio State University
Peninsula Library System, San Mateo, California
We've studied a number of different kinds of users: mostly academic users, undergraduates, graduate students, and faculty, some public library patrons. Today I'll talk about academic library users, students and scholars, whose goals and values are somewhat different than those of most public library users.

We've done these tests in a number of ways: often task-based sessions with prototype or production systems followed by questionnaires and interviews, sometimes contextual interviews, where the test participant trains the observer.

Often our tests are summative: the goal is to validate designs or changes we've made. Sometimes our tests are formative: the goal is to provoke or correct our thinking about an emerging approach or design at a preliminary stage.

Tests like these have some advantages. We see what people actually do (in an artificial setting) rather than hear what they say they would do. We can probe interactively, and get participants to reflect. We can observe behavior across different institutions and over time.

These tests have some limitations too. Our samples are in some ways diverse, they come from different institutions, but they are not reliably representative. They're just loosely aligned with our developing personas and profiles for WorldCat Local by rank, discipline, age, and gender. Test participants are predominantly from the social sciences and humanities; a few have been from the sciences, engineering, and nursing.
Here’s what I believe tests like these can do:

They can show us, for example, that our View Online button is in the wrong place (it was). Or confirm that we’ve moved it to a better place (we did). This sort of page design issue is not especially interesting on a day like today, I think.

They can stop us from asking some questions over and over. Is relevance ranking really the right thing in an academic setting? When? Is bundling together different editions with FRBR really the right thing? Are we simplifying what’s most distracting, or hiding what’s most important?

They can realign our sense of what improvements will matter most. Are the things we might be inclined to think about and improve (because we can, because they matter to us) the things that will deliver most value to users? Sometimes we are startled by how different our users expectations are from ours. Sometimes we just begin to suspect it.

Staff at pilot institutions involved in tests often have questions about ways WorldCat Local might differ from what they’ve been offering:

- combining local, group, global holdings
- including journal article content within the library catalog
- simplifying searching
- simplifying bibliographic displays
- enriching bibliographic displays
- showing facets with search results
- ranking in search results with our approach to relevance
- gathering different editions together and showing just one in search results
Today I want to talk about three of those things:

- search result ranking
- what bibliographic details we show
- editions and FRBRization

These are all things that inform selection. When a student or scholar is looking at a WorldCat Local search result and evaluating what's worth taking time over, how do they decide? I think we've seen enough behavior to give us some confidence about some patterns, and also to make us feel some suspicions.
“Most of the time I come with a specific item in mind—about three quarters of the time. The other quarter of the time is when I’m starting out a research project and I just want to find one book. Or I might try to put together a list of them.”

We sometimes hear that discovery happens elsewhere, that the role of library catalog as a starting point or a gateway is declining, faster in some disciplines than others, faster for scholars and students than for librarians. I’m thinking, for example, of Ithaka’s 2006 Studies of Key Stakeholders in the Digital Transformation in Higher Education, released last summer.
"When we ask these questions, we go to Google. So when we go to a library’s website, the priority is to see if the source is here, in the library."

Still, it was startling to hear every one of our 14 scholarly users in a recent test report that they rarely or never used the library catalog for "looking up a person's work," finding all the editions of a work, or exploring in their fields of expertise.
Instead, graduate students, and faculty report using
- listservs,
- publisher Web sites,
- the New York Review of Books
- other journals that review books
- scanning core journals
- footnotes and bibliographies
- departmental library new book displays
- Google
- Google Scholar
- JSTOR
- and a range of licensed resources (these very often reached through library subject pages or A-Z lists)

Undergraduates often mentioned Google and Wikipedia (Wikipedia always diffidently) as places to go to get some orientation to a topic, and to get terms to bring back to the library catalog. And they frequently mentioned author’s Web sites. For undergraduates, these were the places to start, not the places to end.

Many of these are places where, as one test participant said, every step is interesting.

“If I wanted to know something about the book—let’s say I wanted to look up the date of the latest edition—I wouldn’t go to Melvyl for that. I would just go to Google. I would never even thing of going to Melvyl for that. But if I wanted to borrow the book, I would definitely go to Melvyl.”
All this is not to suggest that discovery isn't happening in library catalogs. In addition to using the catalog to obtain known items discovered elsewhere—the case that predominated for our scholarly test participants—our test participants reported and demonstrated that scholars working outside their area of expertise would look for an introduction, or for a few good items.

Students—undergraduates—would look a few good books, not yet known, on a subject after they've oriented themselves to issues, authors, and vocabulary elsewhere. A couple of the best things, not an exhaustive survey.

A representative Art History undergraduate articulated his criterion for selecting items from a search result as anything that was "not a waste of time." A representative Art History graduate student, searching on a topic she knew well, said "I'll be very disappointed if I find something I haven't found."

In our tests, we could not create the circumstances where scholars would explore a subject in depth, explore a large search result, or create an exhaustive list.

When students and scholars look for known items or search more broadly in WorldCat Local, what is their basis for selection?
One basis is search result ranking.

Our questions:
• What should the default order be in an academic setting?
• Does it differ between advanced researchers and undergraduates?
• What supports finding known items?
• What supports topical searching?

We found that for both for students and for scholars, in both the "known item" case and the topical search case, the expected and preferred order, is "relevance." Nearly all participants expected it; a large majority preferred it. A small minority preferred date descending, and those participants were able to change the sort order. Most test participants recognized the order of search results in WorldCat Local as "relevance."

When searching for known items, test participants expected the item to be on page one of the search result. When it wasn’t, the most common behavior was to search again, using more words or quotation marks.
“What came up first is the ‘Bible’ of this stuff, so that’s totally expected.”

“It makes sense that the most important, seminal works come up first.”

“Landmark books are expected to be among the first.”

For topical searches, test participants—advanced researchers, faculty and graduate students—stated that the “best” items will be at the top of the list. This is “relevance” in a special sense, one that includes renown.
Participants demonstrated more trust than curiosity about how this might be accomplished. Most advanced researchers—faculty and graduate students—could suggest several factors—relevance, popularity, and location—that might be involved. But attention to that was something they regarded as librarians’ expertise, not theirs. They trust their libraries to have made the right choices about how to do this.

We learned that at first we had not made the right choices.

“That’s why we have smart librarians.”

“Somebody put it there [at the top of the search result] for a reason.”
At first, ranking was too complicated. Local holdings were listed first, going from most to least relevant, followed by group holdings going from most to least relevant, followed by worldwide holdings going from most to least relevant. Test participants didn't anticipate, understand, or approve of this, when it was explained.

In response, we created a sort option for relevance without location, still weighted by holdings but without local, group, global grouping, and made this order settable by an institution as the default.
We added a more overt control for location.

Completeness of these search results and what was going on at the tail of the ranking were not a concern for our test participants. Although we had participants do topical searches in their areas of expertise, we could not get them to look past the first two pages of search results.

For both topical and known item searches, what our test participants were looking at on those first few pages—students and scholars alike—were the words they searched with. Participants tended to disregard titles in which their search terms weren't visible.

The extent of this surprised me. Again and again, even advanced scholars didn't dig down to see why an item was at the top of the result. It had better be clear why an item is at the top of that list. I don't think this means that the evidence has to be in the title, necessarily. Matching snippets other than the title could be valuable. But evidence has to be visible on that page. This complicates any effort we make to enhance or enrich relevance ranking.

Beyond those title words, what else do we see test participants rely on, on search result pages and item details pages?
Availability is a basis for selection, especially an indication of electronic availability—for example an indication that the full text was available online, which test participants mistakenly understood us to be giving them, at one point.

Authors—"the living and the dead," as Nancy said today—were an important basis for selection for advanced scholars. Faculty and graduate students indicated that recognizing a reputable author in their field was something they used to decide whether an item in a search result should be looked into further. None of the undergraduate students in our contextual interviews indicated this was a factor in their selection process.

"About the authors" was highly valued by graduate students when we began to direct participants' attention to it. Several participants said they'd value a biography and list of publications, without our saying that's what the link would lead to.
When participants examined item details pages, summaries, abstracts, and tables of contents and were highly valued. Participants often selected them with the mouse pointer, and sometimes read them aloud. This isn't surprising.

Subject headings sometimes played a similar role, supporting evaluation rather than discovery of related items.
Here's what I think is an interesting failure. In several different tests, when a book and book reviews from journals appeared together in search results, most participants became confused, and many remained confused. Eight of ten undergraduates were confused about the difference between books and book reviews; half were able to resolve this confusion. Likewise, among advanced researchers the majority of participants (ten of fourteen, in one test) were derailed by book reviews, and six of these never found the book they were looking for.

Some didn't go past the search results page (more about that later).
Most of those who did weren’t helped by the item details page. Those who went on to view the full text of an article understood as soon as they saw it that it was a book review.

Now, there are some extenuating circumstances: books and book reviews are mixed together, perhaps unexpectedly; for some participants, this was a new system they hadn’t seen before; a stranger was sitting beside them, observing; some test participants were working outside their areas of expertise. But still, these were mostly humanities scholars and social science scholars. Book reviews are a basic part of their world. How mysterious can it be?

In response, we changed the sort algorithm to move books above book reviews in search results. That will improve matters when someone is looking for a book. But it isn’t a very deep approach to the problem. When I look back over all these studies in the past 18 months or so, two things haunt me. This is one. I think it ought to strike us as peculiar. I fear that bibliographic data that is so evocative for me, a former cataloger, might be very much thinner and less expressive than I suppose, even for advanced scholars.

“I’m starting to get more and more confused. Three—four—different authors.”

“A little bit surprising because it’s the exact same title.”

“Funny, it has the same exact title. I just assumed it...”
The Details section contained little of value to any test participants, far less than they expected when they saw the "More details" link.

For items they hadn't already examined, undergraduates reported that they would need to download an article or pick up a book to know how valuable the search result is.

In addition to abstracts, summaries, tables of contents, and excerpts, one way to address this thinness in bibliographic data is with thickeners based on what we think of as social or community input—user-contributed reviews, recommendations, ratings, tags, and lists.
User-contributed reviews got mixed reactions from Graduate students. We tested an early design of the item details page including various social features with six graduate students from a variety of disciplines and one undergraduate. Response to reading reviews was mostly positive. Six of seven participants would read one. Only one of seven would write one. It matters who does write them. We saw a strong preference what participants called "editorial" reviews over user-contributed reviews in several tests.
"Readers recommend" (based, at the time, on recommendations from WeRead) got less mixed—and largely negative—reactions. One participant had a positive reaction because, he said, these recommendations might be based on connections he could not otherwise have made. Five participants had negative reactions that ranged from doubt to a sense of violation. Negative reactions had more to do with who the other readers might be than with the idea of recommendations (which were praised in MedLine) or with the specific items recommended (which were ignored).

To be valued, recommendations need to come from a meaningful source. What we often heard mattered was not just level—faculty recommendations often meant more than student ones to test participants—but also institution and department. And not only—or even mainly—the searcher’s own institution, but rather an institution the participant recognizes as highly regarded in his or her particular discipline.
Tags

“a democratization of the library monopoly on classification”

Graduate students’ reactions to tags as a way of discovering related items were mixed but mostly positive. Five out of seven participants would use others’ tags as a way to search. (Three participants began with a negative view and moved to a positive one.) Test participants said they valued the way tags "go beyond official subjects." They had different ideas about what this way might be, some participants expecting tags to be broader, others narrower, or more current, than library terms. Two of seven participants would not use tags provided by others; both questioned the expertise or viewpoint of those unknown and too various others. One participant said she would add tags of her own, one said she would not. The most critical users—scholars—saw more promise in tags. The one test participant with a strong preference for the "official" subjects provided by librarians was an undergraduate.
Ratings

“Because we don’t know who’s behind them, it ends up being just a frustration.”

“A bit of a confusion between a library site and a commercial site.”

We tested star ratings. Graduate student reactions to ratings were almost all strongly negative. Six participants would not rely on a rating. Five participants would not rate an item.

The main objection here was not only the by now familiar one about who is speaking but also one about purpose. Participants pointed out that they were not reading for leisure (a context in which they might rely on ratings) but rather constructing an argument or developing a position. The fact that something is disagreeable or mistaken would not make it useless. Public library users, whose goals include leisure, might have very different reactions to ratings.
Lists

“User lists would be interesting—a good thing from more commercial sites. Incredibly useful.”

Graduate students valued lists. Again, they'd be valued more highly if the author were known or author's institutional affiliation were recognized and respected.
“I wouldn’t spend the time to do it.”

In a different text—contextual interviews at Ohio State (with undergraduates, graduate students, and faculty)—it was clear that the use of WorldCat Local as a social networking tool for these interviewees was not a high priority. Test participants reported that they did not feel they had the time to create reviews, tables of contents, or add notes to a record.

As a result, we made a number of changes for academic users. We removed ratings. We replaced WeRead recommendations with recommendations derived from searches for author and subject. We retained reviews, but less prominently. And we conducted some focus groups with librarians in a quest for more authoritative reviews and recommendations.

Looking at my last cluster if issues for selecting from search results, what should the granularity of search results be?

Our questions about editions and FRBR:
- Is gathering different editions together through FRBR and representing them with just one entry standing for all of them on search results the right thing to do?
- Are details from just one edition misleading on search result pages? Are F/G/U aware there are others, or are they eclipsed?
- Can test participants find them?
- Do they care?
- What’s the right default edition?
“As a historian, I usually prefer to read a first edition of something if it is primary; if it is secondary, the most recent.”

“Editions really matter because of differences in translations.”

We found editions sometimes matter to scholars—the last, the first, the last during the author’s lifetime. Historians expressed a preference for the first edition of primary sources and the most recent edition of secondary sources. Undergraduates reported in post-test questionnaires that they seldom (seven participants) or never (three participants) looked for a specific edition of a book.

When participants did have a specific edition in mind and had information (publisher, date, editor) to identify it, they generally searched with that information and saw the corresponding edition in the search result, since we show the matching edition. Nearly all our test participants, both undergraduates and graduate students or faculty, have been successful at this sort of task.

But in our first tests no undergraduates or graduate students were successful at finding the latest edition of a work if they didn’t already know the date. This task is one instance of identifying a particular edition without knowing in advance any words that could be used to find it by searching. Performance was almost uniformly poor for faculty, too.
We made some changes. We added a "View all editions and formats" link, first at the tap of the item details page, and then also on the search results page. After that, undergraduates generally found the latest edition without knowing in advance the year of that edition (even when, as often, that edition is not the one shown in brief search results).

Most advanced researchers expressed a desire to have all available editions listed on the Editions tab in reverse chronological order without regard to location. We changed sort of all editions to remove grouping by local, then group, then global holdings.
We found that the default edition matters. Advanced researchers accepted the default when not directed to find a specific edition, without recognizing that it was not the edition they later described as preferring, for example, the earliest or latest edition.
Here's another of my fears. Information that supports selection—abstract, summary, information about access—would be highly valued on search result pages. But the experience I'm used to imagining and that I often take for granted is one where a user (1) sees a search result, then (2) chooses an item, then (3) sees a detailed record, then (4) assesses the item, learns about availability, and makes a choice about how to get the item.

In contrast, one of our faculty participants—smart, attentive, engaged, not atypical—said, while looking at a search result page, about what she'd do next, "Then I'd click on the title and look at the article to see whether it's a useful one." "Would you do that on this page?" I asked. "Is there another page?" she asked. Not for her. Advanced researchers remarked that a link to full-text on the initial search result page (as opposed to the item detail page) would be ideal, and would allow them to bypass the item detail page. Half of participants in another test chose to click on brief descriptions and view item details pages when making a selection from the search results page. And nearly all of these identified the relevant item. But nearly half of participants (six of fourteen) based their selection on what was visible on the brief display only. Two of these reported their next step would be clicking on something to view the full text of the article itself. Only one of these identified the desired item. Now I feel a little uneasy every time I hear us say the words "detailed record," words we use internally for the item details page.
Based on these tests, I think there are some things to feel confident about. I think we can have some confidence that relevance is the right default for search result ranking for scholarly users, although it needs to be more effective and more obvious. Even scholarly users now expect "relevance" to include prominence or importance as a factor. Scholars expect us to promote what will be important to them, students to too. When users like our test participants want to change the order of search results, as they occasionally will, most often it will be to reverse chronological order, and they will be able to. Work rather than edition is the right granularity for search results (although how we show those relationships when people need to see them still needs adjustment).

Among academic users, there will be more interest in the take than in the give of social features, and there will be a preference for recognized experts.

I think there are perhaps some things to worry less about for when we design academic users. Efforts we invest here might not be aligned with the primary concerns of users like our test participants when they're searching in WorldCat Local, since discovery often happens elsewhere, especially in areas of expertise. Navigational tools for large result handling, getting back from page 30 to page 15, ratings, contribution of user reviews might be in this category.

Some important areas for further work might include making the relevance of each item obvious; enriching search results with summaries, abstracts, tables of contents, or excerpts; making the first things searchers see more meaningful; making sure we choose the default edition displayed in search results wisely; obtaining reviews and recommendations from sources that are recognizable and credible, tying opinions to meaningful identities; working to shorten the traverse from search to item.

Those are some things I think we've learned and what I think we've glimpsed in 24 months of studying WorldCat Local in the field.