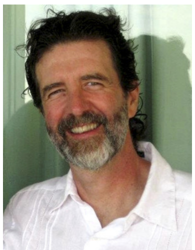
Data-driven deselection for monographs: a rules-based approach to weeding, storage, and shared print decisions

The value of local print book collections is changing. Even as stacks fill and library traffic grows, circulation continues to decline. Across the 'collective collection', millions of unused books occupy prime central campus space. Meanwhile, users want more collaborative study space and online resources. Libraries want room for information commons, teaching and learning centers and cafes. Done properly, removing unused books can free space for these and other purposes, with little impact on users.

Many low-use titles are securely archived, accessible digitally, and widely held in print. Surplus copies can be removed without endangering the scholarly record. But identifying candidates for deselection is time-consuming. Batch-oriented tools that incorporate both archival and service values are needed. This article describes the characteristics of a decision-support system that assembles deselection metadata and enables library-defined rules to generate lists of titles eligible for withdrawal, storage, or inclusion in shared print programs.

The Problem

Book stacks in many libraries are filled beyond the 75% capacity recommended for efficient operation[1]. As library buildings evolve from a 'book-centered' paradigm to a 'learning-centered' paradigm[2], shelving claims a disproportionate share of library space. To quote library space planning expert Scott Bennett: "The crowding out of readers by reading material is one of the most common and disturbing ironies in library space planning"[3]. It is time to rethink the role of print collections. Volume count has receded in importance as a measure of library quality and access to resources[4]. Circulation in academic libraries declined more than 32% between 2004 and 2010[5]; the rate is even more precipitous when growth in enrollment is factored in[6]. In late 2010, Cornell University reported that 55% of its books have not circulated since 1990[7]. Also in 2010, Paul Courant and Buzzy Nielsen estimated that holding books in open stacks costs $4.26 per volume per year; $.86 in high-density storage[8]. The inescapable conclusion: low-use, high-cost print monographs occupy space that is wanted for and by users. This represents a clear call to action.



RICK LUGG
Partner
Sustainable
Collection Services
LLC

Many, if not all, of these same books are widely held and readily available elsewhere. For example, 5.4 million monographs have been digitized and securely archived in the Hathi Trust Digital Library. Of these, 28% are in the public domain, freely accessible to any user[9]. At least 24% of those 5.4 million titles are also held in print by more than 100 libraries, many of them in facilities with climate and access controls[10]. They are available via inter-library loan or direct borrowing arrangements. If these titles follow typical patterns, 40–50% have never circulated. Collectively, they represent millions of surplus copies, most of which are unnecessary to support user demand and archival security.

"… low-use, high-cost print monographs occupy space that is wanted for and by users. This represents a clear call to action."

It is clearly time to reduce the level of redundancy in print collections. This will release space for other purposes – those in higher demand from users and librarians alike. Each 100,000 print volumes removed will yield up to 20,000 square feet[11]. Constance Malpas of OCLC Research estimates that "the median space savings that could be achieved at an ARL library if a robust shared print offer were in place today amounts to approximately 36,000 linear feet or the equivalent of more than 45,000 assignable square feet[12]. There are enormous opportunities here, which can be realized with negligible risk, by removing some excess copies from the collective collection. Addressing this is simply another form of good stewardship, and should be pursued whether or not a library needs more space.

"It is clearly time to reduce the level of redundancy in print collections."

The argument can certainly be developed further, but there is already enough evidence to justify action. It is vital that this action be co-ordinated, to assure that no content disappears, and that many libraries are not removing the same titles. But the combination of factors outlined above suggests that controlled deselection on a large scale can be done safely and cost-effectively. Deselection is defined broadly here, to include withdrawal, storage, and inclusion in shared print management programs. The goals for such a drawdown would include:

- reduce the overall number of low-use surplus copies
- minimize the number of low-use copies held in open stacks
- assure that deselected content is securely archived in both print and digital form
- co-ordinate deselection regionally and nationally to assure sufficient copies are retained
- assure that deselected content remains accessible to users (in the unlikely event that it is wanted).

This is a rational case with rational objectives, but reason – even when supported with good data – may not be enough. Unlike journals, the prospect of removing books from shelves arouses strong emotions. Perceiving a valuable and venerable institution at risk, faculty members, librarians and even undergraduate students are moved to protest. Strong arguments surface about browsing, serendipity, and a disproportionate effect on humanities scholars[13,14,15]. Electronic books have not yet supplanted print to the extent that e-journals have done. These issues warrant discussion and adjustment, but need to be balanced with use data and economic realities. They should give us pause, but they should not stop us entirely.

"Unlike journals, the prospect of removing books from shelves arouses strong emotions."

Archival values and service values

In addressing concerns about deselection, it is important to remember they are multi-faceted. Michael Buckland, in his 1992 *Redesigning Library Services: A Manifesto,* describes two key roles of library collections: 'preservation' and 'dispensing'[16]. These embody the core values of librarianship, and both must be honored in any collection strategy. For each deselection candidate title, we must answer two questions:

Has the content been securely archived? As a community, we must make certain that no content is lost, and that all books are preserved in both digital and printed form. Archived titles do not necessarily need to be immediately available to every library's users, but archiving status does need to be known for every title under consideration for deselection, and especially for those that will be withdrawn. In the event that content has not been archived, the library can retain its copy as a contribution to the collective collection.

Does the content remain accessible to users? This is a very different question, relating to discoverability, convenience and delivery. User needs might be met with a copy from a shared print partner, with digital access through Hathi Trust or a commercial e-book supplier, by inter-library loan or purchase of a used copy. Knowing where and how a withdrawn title can be quickly re-obtained if wanted mitigates risk and lends confidence to the deselection decision.

While both questions are vital, it is important to consider them separately, as archived content and 'serviceable' content have different characteristics. Both are needed to assure low-risk deselection decisions.

In a recent discussion about 'archive copies' and 'service copies', I argued that responsibility for archiving belongs at the national or even the network level, while responsibility for 'servicing' operates more effectively at the regional level[17]. Highly-specialized or local-interest content is best handled locally, whether through formal special collections units or informal areas of subject or geographic concentration. For circulating monographs, locally-held collections will increasingly consist of high-use titles, along with *only* those low-use titles that are part of that library's commitment to a regional distributed print collection. Management of collections in this manner requires better data than most libraries currently have.

> "Good information about collections, circulation and holdings can clarify strategic and practical choices, give weight to priorities, and help estimate the impact of decisions."

Data and deselection

Good information about collections, circulation and holdings can clarify strategic and practical choices, give weight to priorities, and help estimate the impact of decisions. To answer questions about archival status, serviceability and local interest, a number of data points about each title must be assembled, including information about usage, redundancy, archiving, value and alternative access. This data is usually available, but is often dispersed across a mix of library and third-party sources. Identifying, aggregating and normalizing such data are the first steps in a rules-based deselection process.

Library data such as bibliographic records, item and holdings records and circulation history provide a good starting point for collection analysis. The structure and extent of information depend on the library management system in use, and on past decisions, e.g., was circulation data carried over at the last system migration? Library policies also affect what data may be available. Are reshelving counts in place to capture in-library use? Are inter-library loan statistics incorporated in circulation history? Can the date of acquisition be retrieved from the acquisitions module or derived from item creation date? For deselection decisions, key elements include publication date, OCLC control number, the number of total check-outs, and location (to screen for special collections, reference, government documents and other titles with different use patterns). It is useful to know date of accession and date of last circulation, extent of in-library use, call number and barcode number. Identifying titles of local importance often relies on notes in the MARC 590 field; examples include works by faculty authors, gifts from specified donors, or historical collections of importance to the institution.

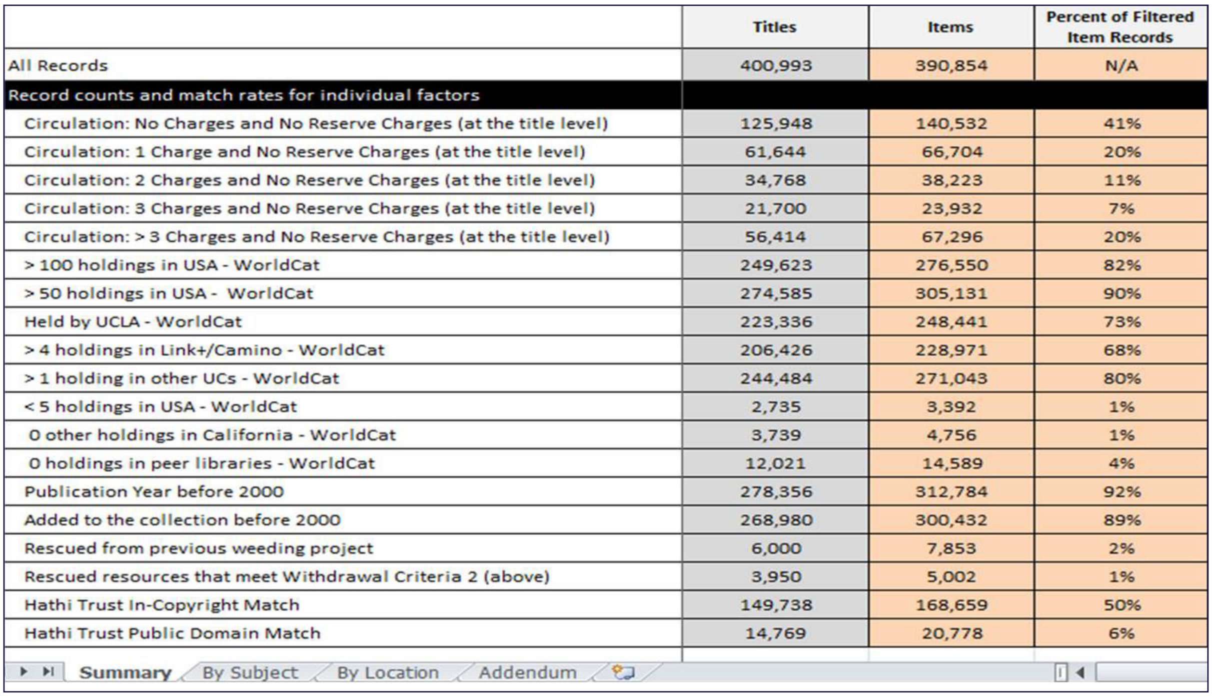
All these elements of the library's data must be gathered, validated and normalized, both for comparison to external sources and, for shared print projects, for aggregation with data from partner libraries. Normalization seeks to improve match rates with external sources, and to assure that locally-usable lists can be produced once analysis is complete. One useful by-product of data normalization work is that corrections and additions can be returned to the library to update its local catalog. Examples of such data remediation include new or adjusted control numbers and holdings updates to WorldCat. Improved library data can in turn improve subsequent analyses. Most libraries can extract the necessary bibliographic, item and circulation data with one to two days of effort by a systems librarian. Data normalization may require more time, but can also be addressed through a holdings reclamation project or use of a vendor, which can help save staff time and control costs.

But library data alone is usually not enough for deselection decisions. It must be augmented with information about holdings in other libraries, authoritative title lists, archiving status and alternative availability. This data exists outside the library catalog. It can be obtained by searching sources such as WorldCat and Hathi Trust, but given the scale of most collection analysis and deselection work, batch matching and retrieval is really the only practical solution. Here again, both archival and service values must be considered. To assure that a title is securely archived, it is important to know if it appears in the Hathi Trust Digital Library, and if so, whether it is in the public domain or in copyright. This can be determined directly though the Hathi Trust data API. To assure that a title is reasonably secure in print form, it is important to know how widely held it is in the collective collection, and whether any of those holdings include a secure print archive. The OCLC WorldCat Search API provides detailed answers to these questions.

WorldCat holdings information is also valuable in gauging 'serviceability'. Holdings data can be sorted by country, state/province and by relevant groups of peers: direct borrowing partners, shared print partners and libraries of similar type and size. For deselection and shared print, the proximity of a copy (and a corresponding regional shared print commitment) can assure timely access for future users. Other data points and sources may be important as well, to identify titles that have particular importance to an individual library. Titles that appear on authoritative lists (e.g., CHOICE Outstanding Academic Titles, prize-winners, accreditation checklists) offer one example. Titles of local interest (geographical or topical focus, works by faculty authors) are another. To the degree that these attributes are reflected in the library's data or obtainable via a license to third-party data, they can be flagged and possibly protected from deselection.

Any library can on its own gather and use data from WorldCat, Hathi Trust or authoritative lists, to supplement information from its own catalog. But developing the necessary batch processes, matching routines, and validation can be time-consuming. There are always other things to do. As with other tasks, such as book purchasing or systems maintenance, vendor services can supplement staff efforts. Our firm, Sustainable Collection Services, was founded specifically to assemble and manipulate library-provided and external data in support of deselection decisions. But whether assisted or unassisted, every library needs this augmented data to inform deselection decisions.

The next step in 'data-driven deselection' is to bring together the library's data with that obtained from external sources. One of the most interesting moments in a collection analysis project is the first composite view of data drawn from multiple sources. A collection summary (see Figure 1) provides a high-level view of the augmented data, with each attribute quantified. This summary can serve as a management tool, suggesting where deselection efforts will yield most benefit. More granular views, by subject or location, can help refine project strategy. Titles that are held scarcely elsewhere can become candidates for preservation. The summary becomes still more powerful when librarians can interact with



	Titles	Items	Percent of Filtered Item Records
All Records	400,993	390,854	N/A
Record counts and match rates for individual factors			
Circulation: No Charges and No Reserve Charges (at the title level)	125,948	140,532	41%
Circulation: 1 Charge and No Reserve Charges (at the title level)	61,644	66,704	20%
Circulation: 2 Charges and No Reserve Charges (at the title level)	34,768	38,223	11%
Circulation: 3 Charges and No Reserve Charges (at the title level)	21,700	23,932	7%
Circulation: > 3 Charges and No Reserve Charges (at the title level)	56,414	67,296	20%
> 100 holdings in USA - WorldCat	249,623	276,550	82%
> 50 holdings in USA - WorldCat	274,585	305,131	90%
Held by UCLA - WorldCat	223,336	248,441	73%
> 4 holdings in Link+/Camino - WorldCat	206,426	228,971	68%
> 1 holding in other UCs - WorldCat	244,484	271,043	80%
< 5 holdings in USA - WorldCat	2,735	3,392	1%
0 other holdings in California - WorldCat	3,739	4,756	1%
0 holdings in peer libraries - WorldCat	12,021	14,589	4%
Publication Year before 2000	278,356	312,784	92%
Added to the collection before 2000	268,980	300,432	89%
Rescued from previous weeding project	6,000	7,853	2%
Rescued resources that meet Withdrawal Criteria 2 (above)	3,950	5,002	1%
Hathi Trust In-Copyright Match	149,738	168,659	50%
Hathi Trust Public Domain Match	14,769	20,778	6%

Figure 1: SCS Collection Summary, with data drawn from library and third-party sources

the data, using different criteria to model a range of withdrawal scenarios, and ultimately to define a set of deselection rules that reflect local imperatives and local values.

A collection summary shows the potential effect of individual parameters, and can also help gauge what happens when they are combined. To the degree supported by the augmented data, different deselection scenarios can be composed, tested and modified. Sometimes many iterations are needed to evolve rules that satisfy both local users and the need to reclaim space. Sometimes different rules must be developed for different subjects. At their simplest, withdrawal rules might look like these examples:

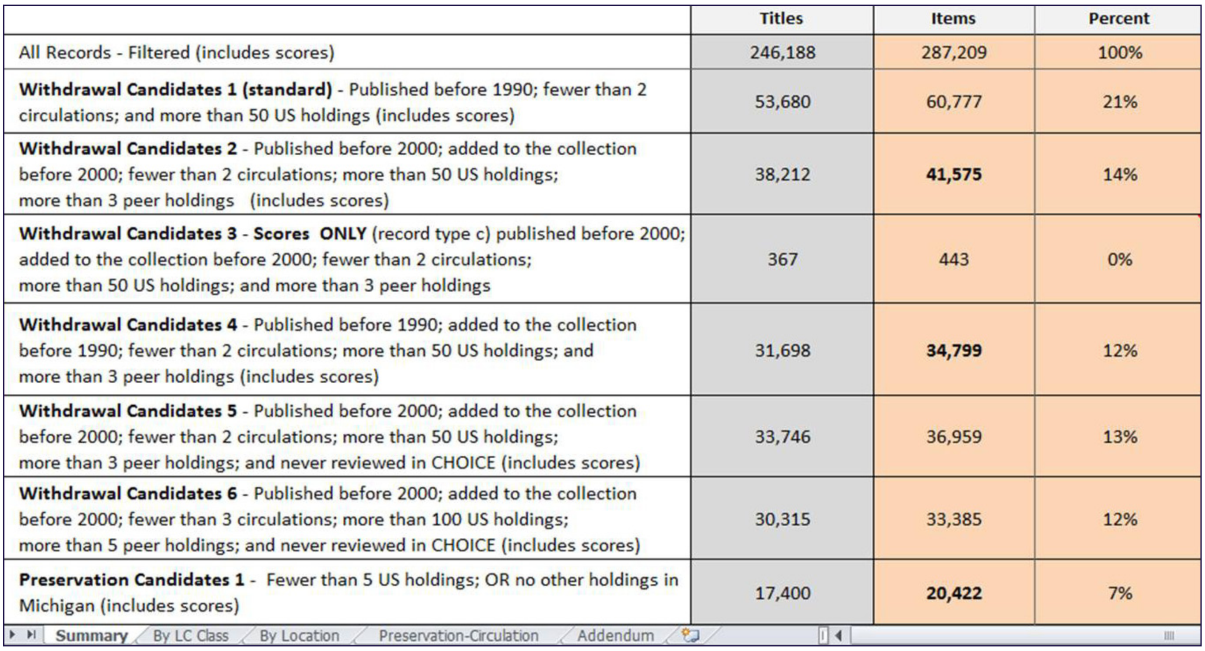
- 0 checkouts, AND published before 2000, AND never reviewed in CHOICE, AND held by more than 4 peer libraries.
- published before 2002, AND fewer than 3 checkouts, AND more than 3 peer holdings, AND more than 100 US holdings
- fewer than 5 total checkouts, AND more than 50 US holdings, AND in Hathi Trust.

And preservation rules might look like this:

- fewer than 5 US holdings, OR no other holdings in my state
- fewer than 10 US holdings, AND not in Hathi Trust
- published before 1875, AND no peer holdings.

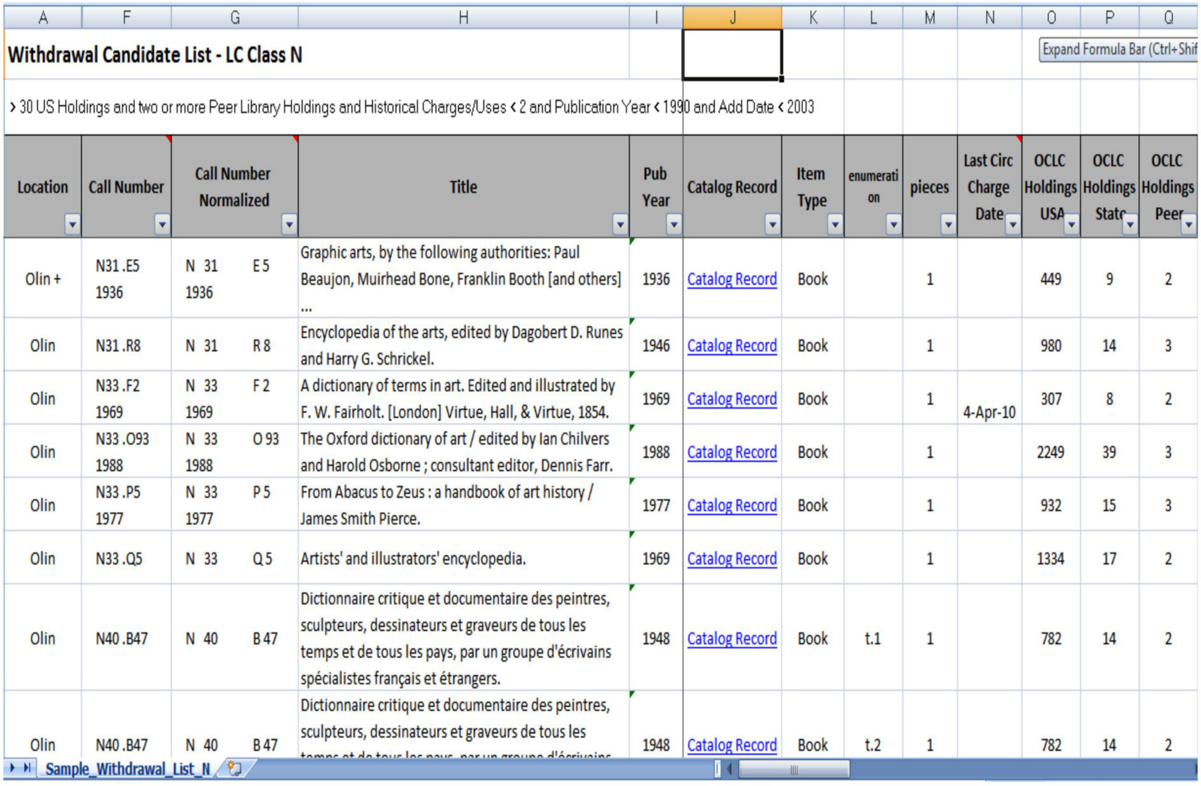
Rules can vary by subject or location, and can involve as few or as many elements as the data supports. Figure 2 shows some additional examples, and quantifies the effect of each scenario.

The basics of data-driven deselection, then, are simple. Gather data from library and external sources. Look for patterns. Develop criteria that both secure the collection and assure future accessibility. Model those rules against the data to estimate impact. Adjust and iterate until the desired balance can be achieved. Then rely on those rules, rather than title-by-title examination, to generate lists of candidates for withdrawal and preservation. Finally, as shown in Figure 3, present candidate lists that provide the deselection metadata for each title, to enable spot-checking and to maintain confidence in the rules.



	Titles	Items	Percent
All Records - Filtered (includes scores)	246,188	287,209	100%
Withdrawal Candidates 1 (standard) - Published before 1990; fewer than 2 circulations; and more than 50 US holdings (includes scores)	53,680	60,777	21%
Withdrawal Candidates 2 - Published before 2000; added to the collection before 2000; fewer than 2 circulations; more than 50 US holdings; more than 3 peer holdings (includes scores)	38,212	**41,575**	14%
Withdrawal Candidates 3 - **Scores ONLY** (record type c) published before 2000; added to the collection before 2000; fewer than 2 circulations; more than 50 US holdings; and more than 3 peer holdings	367	443	0%
Withdrawal Candidates 4 - Published before 1990; added to the collection before 1990; fewer than 2 circulations; more than 50 US holdings; and more than 3 peer holdings (includes scores)	31,698	**34,799**	12%
Withdrawal Candidates 5 - Published before 2000; added to the collection before 2000; fewer than 2 circulations; more than 50 US holdings; more than 3 peer holdings; and never reviewed in CHOICE (includes scores)	33,746	36,959	13%
Withdrawal Candidates 6 - Published before 2000; added to the collection before 2000; fewer than 3 circulations; more than 100 US holdings; more than 5 peer holdings; and never reviewed in CHOICE (includes scores)	30,315	33,385	12%
Preservation Candidates 1 - Fewer than 5 US holdings; OR no other holdings in Michigan (includes scores)	17,400	**20,422**	7%

Figure 2. Withdrawal and preservation scenarios for an undergraduate library



Withdrawal Candidate List - LC Class N

> 30 US Holdings and two or more Peer Library Holdings and Historical Charges/Uses < 2 and Publication Year < 1990 and Add Date < 2003

Location	Call Number	Call Number Normalized	Title	Pub Year	Catalog Record	Item Type	enumeration	pieces	Last Circ Charge Date	OCLC Holdings USA	OCLC Holdings State	OCLC Holdings Peer
Olin +	N31 .E5 1936	N 31 E 5 1936	Graphic arts, by the following authorities: Paul Beaujon, Muirhead Bone, Franklin Booth [and others] ...	1936	Catalog Record	Book		1		449	9	2
Olin	N31 .R8	N 31 R 8	Encyclopedia of the arts, edited by Dagobert D. Runes and Harry G. Schrickel.	1946	Catalog Record	Book		1		980	14	3
Olin	N33 .F2 1969	N 33 F 2 1969	A dictionary of terms in art. Edited and illustrated by F. W. Fairholt. [London] Virtue, Hall, & Virtue, 1854.	1969	Catalog Record	Book		1	4-Apr-10	307	8	2
Olin	N33 .O93 1988	N 33 O 93 1988	The Oxford dictionary of art / edited by Ian Chilvers and Harold Osborne ; consultant editor, Dennis Farr.	1988	Catalog Record	Book		1		2249	39	3
Olin	N33 .P5 1977	N 33 P 5 1977	From Abacus to Zeus : a handbook of art history / James Smith Pierce.	1977	Catalog Record	Book		1		932	15	3
Olin	N33 .Q5	N 33 Q 5	Artists' and illustrators' encyclopedia.	1969	Catalog Record	Book		1		1334	17	2
Olin	N40 .B47	N 40 B 47	Dictionnaire critique et documentaire des peintres, sculpteurs, dessinateurs et graveurs de tous les temps et de tous les pays, par un groupe d'écrivains spécialistes français et étrangers.	1948	Catalog Record	Book	t.1	1		782	14	2
Olin	N40 .B47	N 40 B 47	Dictionnaire critique et documentaire des peintres, sculpteurs, dessinateurs et graveurs de tous les	1948	Catalog Record	Book	t.2	1		782	14	2

Figure 3: Sample withdrawal candidate list with selected data elements

Taking action

Even the best decision-support systems cannot actually make decisions. Nor can they do the work that results from a decision. The library will incur costs for data extraction, selector review, record maintenance and disposition of physical items. But by building decisions around data and rules, and by supplying information in context, a rules-based approach can save time and improve consistency in deselection. This same approach can enable batch-level handling of data remediation, record maintenance, discovery improvements, and transfers and withdrawals. In a shared print context, it can enable equitable allocation of withdrawal opportunities and retention commitments. In a recent project involving seven academic libraries, the Michigan Shared Print initiative identified 534,000 withdrawal candidates, while retaining two copies of every title in the group collection.[18]

> " … by building decisions around data and rules, and by supplying information in context, a rules-based approach can save time and improve consistency in deselection."

There are limits, of course. Data-driven deselection can only be as good as the underlying data. Accuracy depends on how recently and how well inventories and reclamation projects have been done. The quality of bibliographic, item and circulation data determines the effectiveness of matching with external sources. The completeness of holdings information in WorldCat or among consortial partners governs those results. But in the end, data offers the safest and most efficient way forward. In the Michigan Shared Print Initiative, it cost $.29 to identify each withdrawal candidate, based on criteria defined by the group. (This includes only the data work, not development of criteria.) In a 2011 blog post, I estimated the total cost of deselection at $3-4 per volume[19]. But deselection is a one-time cost, and must be judged against the $4.26 *per year* cost of retaining books in open stacks. These numbers suggest that a strong business case can be made. By improving the data, by working with rules and batches rather than individual titles and items, libraries can assure that archiving and accessibility are assured, while realizing space gains, avoiding costs, and better aligning resource use with institutional priorities.

References

1. Libris Design Project, *Library Stacks and Shelving*. This material has been created by Earl Siems and Linda Demmers, supported by the US Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian [undated]:
http://www.librisdesign.org/docs/ShelvingforLibraries.pdf (accessed 1 May 2012).

2. Bennett, S, *Libraries and Learning: A History of Paradigm Change,* 2008, Scott Bennett, Library Space Planning:
http://www.libraryspaceplanning.com/assets/resource/Libraries-and-learning.pdf (accessed 1 May, 2012).

3. Bennett, S, *Libraries Designed for Learning,* November 2003, Washington, DC, Council on Library and Information Resources:
http://www.libraryspaceplanning.com/assets/resource/libraries-designed-for-learning.pdf (accessed 1 May 2012).

4. Kyrillidou, M, *Reshaping ARL Statistics to Capture the New Environment,* ARL: A Bimonthly Report, no. 256 (February 2008), 9–11:
http://www.arl.org/bm~doc/arl-br-256-stats.pdf (accessed 1 May 2012).

5. Phan, T, Hardesty, L, Hug, J and Sheckells, C, *Academic Libraries: 2010* (NCES 2012-365). US Department of Education, Washington, DC: National Center for Education Statistics:
http://nces.ed.gov/pubsearch (accessed 1 May 2012).

6. Kurt, W, *The end of academic library circulation?* 1 February 2012, ACRL TechConnect blog:
http://acrl.ala.org/techconnect/?p=233 (accessed 1 May 2012).

7. Cornell University Library, *Report of the Collection Development Executive Committee Task Force on Print Collection Usage*, October 22, 2010 (revised 22 November 2010):
http://staffweb.library.cornell.edu/system/files/CollectionUsageTF_ReportFinal11-22-10.pdf (accessed 1 May 2012).

8. Courant, P and Nielsen, M, On the Cost of Keeping a Book. In: *The Idea of Order*: *Transforming Research Collections for 21st Century Scholarship*, June 2010, Washington, DC, Council on Library and Information Resources:
http://www.clir.org/pubs/reports/pub147/pub147.pdf (accessed 1 May 2012).

9. Hathi Trust Digital Library:
http://www.hathitrust.org/statistics_info (accessed 1 May 2012).

10. Malpas, C, *Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment,* 2011, Dublin, Ohio, OCLC Research:
http://www.oclc.org/research/publications/library/2011/2011-01.pdf (accessed 1 May 2012).

11. Lawrence, S, Connaway, L and Brigham, K, Life Cycle Costs of Library Collections: Creation of Effective Performance and Cost Metrics for Library Resources, *College and Research Libraries,* November 2001, 62(6), 541–553:
http://crl.acrl.org/content/62/6/541.full.pdf (accessed 1 May 2012).

12. Malpas, C, ref. 10.

13. Lugg, R, 5 July 2011 [blog post] Disturbing Dust and Data, Sample & Hold blog:
http://sampleandhold-r2.blogspot.com/2011/07/disturbing-data.html (accessed 2 May 2012).

14. Lugg, R, 17 January 2012 [blog post] Browsing Now, Sample & Hold blog:
http://sampleandhold-r2.blogspot.com/2012/01/browsing-now.html (accessed 2 May 2012).

15. Lugg, R 25 January 2012 [blog post] Browsing Now (2), Sample & Hold blog:
http://sampleandhold-r2.blogspot.com/2012/01/browsing-now-2.html (accessed 2 May 2012).

16. Buckland, M. *Redesigning Library Services: A Manifesto,* 1992, Chicago, American Library Association:
http://sunsite.berkeley.edu/Literature/Library/Redesigning/html.html (accessed 1 May 2012).

17. Lugg, R, Library Logistics: Archiving and Servicing Shared Print Monographs, *Against the Grain*, 2012, 24(3).

18. Lugg, R, 6 March 2012 [blog post] MCLS and SCS, Sample & Hold blog:
http://sampleandhold-r2.blogspot.com/2012/03/mcls-and-scs.html (accessed 21 May 2012)

19. Lugg, R, 16 May 2011 [blog post] The Cost of Deselection (10): Summing Up, Sample & Hold blog:
http://sampleandhold-r2.blogspot.com/2011/05/cost-of-deselection-10-summing-up.html (accessed 21 May 2012).

Article © Rick Lugg

Rick Lugg, Partner
Sustainable Collection Services LLC, 63 Woodwell's Garrison, Contoocook, NH 03229 USA
Email: rick@sustainablecollections.com | web: http://sustainablecollections.com

To cite this article:

Lugg, R, Data-driven deselection for monographs: a rules-based approach to weeding, storage, and shared print decisions, *Insights*, 2012, 25(2), 198–204, doi: 10.1629/2048-7754.25.2.198

To link to this article:

http://dx.doi.org/10.1629/2048-7754.25.2.198