# The Emergence of the Collective Collection: Analyzing Aggregate Print Library Holdings

## By Lorcan Dempsey

OCLC Vice President, Research, and Chief Strategist

**OCLC**®
Research

December 2013

OCLC Research

Dublin, Ohio 43017 USA

www.oclc.org

Please direct correspondence to:

Lorcan Dempsey

OCLC Vice President, Research, and Chief Strategist

dempsey@oclc.org

# As the network continues to reconfigure personal, business and institutional relationships, it is natural that we also continue to see changes in how library collections are managed: changes in focus, boundaries and value.

One important trend is that libraries and the organizations that provide services to them will devote more attention to system-wide organization of collections—whether the "system" is a consortium, a region or a country.

We have become used to this in the digital environment, where the scale advantage of such consolidation is apparent. Think of shared approaches to preservation of licensed material, as in CLOCKSS or Portico, for example. Or think of the emergence of JSTOR and HathiTrust, under very different business models, as shared digital hubs that concentrate capacity—a natural trend as materials are digitized and aggregated at the network level. Or think of the interest in aggregating metadata across institutional digital collections, as in Europeana,[1] WorldCat or the Digital Public Library of America.[2]

Recently, print collections have also been the subject of such shared attention. Libraries are beginning to evolve arrangements that will facilitate long-term shared management of the print literature as individual libraries begin to manage down their local capacity. Examples of initiatives here are the WEST Project[3] and the print management activities of the HathiTrust. Initially, attention was focused on journal runs, but it is now spreading to monographs, as well. Of course, libraries have long worked with print repositories, individually or in shared settings. However, a more systemic perspective is now emerging and we have been using the phrase "collective collection" to evoke this more focused attention on collective development, management and disclosure of collections across groups of libraries at different levels. In a major shift, a shared approach to print management is on the rise, and we anticipate that a large part of existing print collections, distributed across many libraries, will move into coordinated or shared management within a few years. This may involve physical consolidation, or a more distributed approach where individual libraries declare commitments around parts of their collections. In this way, some attention shifts from the institution to supra-institutional structures as the venue for print collection management. Policy, organizational and service arrangements are now emerging around this trend.

The collective collection has been a major interest of OCLC Research. This is to be expected given the data we have in WorldCat about collective library holdings and OCLC's goal to make

shared working among libraries more efficient. As interest in coordinated management of the collective print collection grows, we thought it was a useful time to pull together some of our writings on this topic in a single volume. This short piece provides some environmental introduction for the contributions that follow.

Interest in shared print strategies has had several drivers.

- **Google Books**. Google's December 2004 announcement of its intention to collaborate with five major research libraries to digitize their print collections and make them available for searching galvanized discussion about the collective print collection. Notably, it suddenly became possible to imagine the digitization of a large part of that collection, providing a significant alternative entry point to the print literature. The establishment of HathiTrust has provided community focus for curation of the digitized book corpus, even as the rate of digitization has slowed; it has moved curation to the network level. These developments have raised major questions about stewardship and permissible behaviors, questions that have resulted in legal actions. Institutions are beginning to plan their local collections in the context of the collective collection. For example, local decisions about print will increasingly be influenced by the emerging shared print management apparatus and by HathiTrust, as well as by the growing availability of e-books and more on-demand lending or acquisition practices.

- **The digital turn: changing patterns of research and learning**. While the print literature remains important to learning and research, overall use has declined to the extent that in some cases a misalignment is seen between current levels of investment in acquisition and management of print and the research and learning demands being placed on the library. The real cost of managing print has also become more apparent, at the same time as the use of digital materials increases. While print remains important in some contexts, there is a general move to digital resources, e-books and on-demand models.

- **Opportunity costs and space**. The opportunity cost of using space for the management of print collections is also becoming more apparent. Space is often required for higher value activities than storage of print collections, which are seen to be progressively releasing less value in actual use by students and faculty. Space is being reconfigured around broader education and research needs, and less around the management of print collections. It supports social interaction around learning and research, and access to specialist expertise, equipment or communication facilities, as well as more exhibitions and interaction. In this context, many libraries are now actively managing down print.

- **Efficient access to print**. If fewer print materials are available in close proximity to users, it becomes important to ensure convenient discovery and delivery of those materials within new arrangements. Adequately supporting humanities scholars or other groups for whom print is important may depend on efficient delivery from within a system-wide apparatus of provision. For example, there are early discussions within the Committee on Institutional Cooperation (Big 10 institutions) about distributing print repositories to members specialized by subject, and ensuring rapid delivery across the system as a whole. For local service and political reasons it may be difficult to move in preferred directions without assurances about such broader provision.

- **A general move to collaboration**. Stronger models of collaboration are emerging, such as 2CUL[4] (Columbia and Cornell University Libraries) and the Orbis Cascade Alliance[5] (academic libraries in Oregon and Washington), which rebalance collections (as well as services, expertise and systems) in larger units.

Against this background it is interesting to consider this excerpt from a recent vision statement at the University of Arizona:

Our goal is to be a primarily digital library. Simply put, it is no longer possible to sustain the massive print collections of the past. Our current physical plant is virtually full, and campus realities dictate that new buildings for the foreseeable future will be devoted to STEM initiatives with revenue generation potential. By shifting our focus from large print collections to electronic resources that are available anytime anywhere, the Libraries have moved from a "just in case" strategy to a "just in time" approach involving on-demand purchasing and large investments in speedy interlibrary loan. More than 94% of our serials and 20% of all our books are now electronic. In FY2012, our electronic book purchases exceeded 50% of total monographs purchased. In FY1999, we began to remove print materials duplicated by eresources to provide more out-of-classroom learning space for the campus. We have removed more than 750,000 print volumes to date. When print materials are desired or required, however, we are able to offer our quick and efficient interlibrary loan service or on-demand acquisition. We make collections decisions based on customer feedback, which has grown consistently more positive over the years.

In 2005, we became the nation's first all-electronic Federal Government Depository Library. In 2011, we initiated on-demand, patron-driven access to electronic and print books. Listings for more than 60,000 scholarly books have been added to the library catalog; as these books get used by customers, we buy them and add them to our permanent collection. Research shows that patron-selected items get used more often,

so this buying method maximizes the Libraries' purchasing dollars while giving users access to more information resources. (University of Arizona Libraries 2013, 4)

A system-wide perspective signals a real shift in emphasis. For most of its history, the library model was largely one of managing locally assembled collections. And the "goodness" of a library was strongly associated with its size, because more resources were available to its user base. This was natural in a print environment, where physical distribution in multiple collections in a just-in-case model was the most efficient way of meeting institutional needs. Responding on a case-by-case basis to satisfy student or faculty information needs as they were expressed would give rise to intolerable transaction costs.

This local assembly model was augmented by cooperation at the margins, whether through interlibrary loan or some cooperative approaches to collection development. As more of these print collections move into collective management, some core characteristics of the model change in interesting ways.

- **"Collection intelligence"** has to scale to the level of the system rather than the institution. In line with the local focus, libraries are used to treating their local catalogs as the definitive record of their collections, and shared cataloging and other approaches support this local management. As we think more about shared collections, this model flips. It becomes important to understand the characteristics of the collective collection so that local decisions can be made accordingly. The nature and quality of collective data becomes important to facilitate decision-making and effective processing. For example, while libraries now share data about the titles in their collections, they typically do not share data about individual copies. This becomes more important as libraries are interested in how many copies are in the system. OCLC's WorldCat and other union catalogs have become important tools in thinking about the characteristics of system-wide provision.

- **Preserving the scholarly record**. In the print world, preservation was a benign consequence of the redundancy inherent in the physical distribution model. Lots of copies, as they say, keep stuff safe. As this redundancy is reduced, a more planned or interventionist approach becomes important.

- **A balance of responsibilities**. Perceived responsibility for stewardship, provision or funding will vary across libraries in any evolving arrangement. Many research and national libraries recognize a mission-driven responsibility of stewardship to the scholarly and cultural record, and will undertake to work together and individually to discharge it as the environment changes. Other libraries may have specific regional or subject interests. However, many libraries may prefer to be consumers rather than providers of shared collections, and may wish to participate more selectively, on a fee

or membership basis, relying on collaborative or third-party arrangements to manage print collections. Others again may feel no need to make such a contribution. An important part of the shared print initiatives underway is to develop sustainability models that recognize the various interests at play within the system, and to put in place incentives to try to assure appropriate levels of participation.

- **Ownership.** It has been usual for libraries to think that they "own" the books in their collections. Google Books and HathiTrust have underlined that libraries actually have a bundle of rights—they can do some things but not others. At the same time, libraries are beginning to think about shared models of ownership and curation.

This is the context in which OCLC Research has developed a stronger interest in the contours of the collective collection. From our early work on the characteristics of the collections of the first libraries to participate in the Google Books program, there has been keen interest in knowing more about the composition of individual collections and about overlap and distinctiveness in the context of the aggregate library collection. We have looked at a variety of questions here. Our main resource has been WorldCat. At the time of writing, WorldCat contained approximately 300 million bibliographic records representing approximately 2 billion library holdings around the world.[6] While its coverage varies by type of library and region, WorldCat is the most complete record of global library holdings available.

We have three broad interests, which cluster around better understanding the existing collective collection and supporting the optimal evolution of reconfigured collections:

1. Understanding the characteristics of the collective print collection: how it is distributed across libraries and regions; its composition in terms of age, subject, copyright status and so on; levels of overlap, rarity and distinctiveness.

2. Supporting policy and service decision-making with good intelligence based on WorldCat and other data resources.

3. Understanding patterns or trends within the scholarly and cultural record. This is akin to "culturomics" (Michel et al. 2011) or "distant reading" agendas (Moretti 2013), which apply data-mining techniques to large aggregations of digitized text and metadata. It is a relatively new interest, and is not strongly represented in this volume. It is an area where we would like to encourage others to use WorldCat as a scholarly resource.

The report, *Understanding the Collective Collection: Towards a System-wide Perspective on Library Print Collections* (Dempsey et al. 2013) contains the following contributions:

Lavoie, Brian F. and Roger C. Schonfeld. 2006. "Books without Boundaries: A Brief Tour of the System-wide Print Book Collection." *Journal of Electronic Publishing* 9,2 (Summer). DOI: http://dx.doi.org/10.3998/3336451.0009.208.

*Based on an analysis of WorldCat, this article discusses the characteristics of the North American print book "collective collection."*

Dempsey, Lorcan. 2006. "Libraries and the Long Tail: Some Thoughts about Libraries in a Network Age." *D-Lib Magazine*, 12,4 (April). http://www.dlib.org/dlib/april06/dempsey/04dempsey.html.

*The long tail proposition is about how well supply and demand are matched in a network environment. This article considers library collections from this point of view and asks whether the current situation is the optimal system-wide arrangement of collections.*

Lavoie, Brian F., Lynn Silipigni Connaway, and Lorcan Dempsey. 2005. "Anatomy of aggregate collections the example of Google Print for libraries." *D-Lib Magazine*, 11,9 (September). http://www.dlib.org/dlib/september05/lavoie/09lavoie.html.

*The initial Google digitization initiative galvanized interest in the composition and overlap of book collections. This important study looked at the overlap between collections of the original five library participants. It found that "rareness is common."*

Lavoie, Brian, and Lorcan Dempsey. 2009. "Beyond 1923: Characteristics of Potentially In-copyright Print Books in Library Collections." *D-Lib Magazine*, 15,11/12 (November/December). http://www.dlib.org/dlib/november09/lavoie/11lavoie.html.

*Rights and allowable uses became a major area of discussion and contention around the emerging digitized corpus. This article aimed to provide some empirical basis for those discussions by exploring the characteristics of print books published in the US after 1923.*

Malpas, Constance. 2011. *Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment.* Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2011/2011-01.pdf.

*The objective of the project was to examine the feasibility of outsourcing management of low-use print books held in academic libraries to shared service providers, including large-scale print and digital repositories. It helped set the agenda around the emerging shared print discussions.*

Lavoie, Brian, and Günter Waibel. 2008. An Art Resource in New York: The Collective Collection of the NYARC Art Museum Libraries. Dublin, Ohio: OCLC Programs & Research. http://www.oclc.org/content/dam/research/publications/library/2008/2008-02.pdf.

*This report examines the collective collection of four New York City-area art museum libraries, highlighting areas of distinctiveness and overlap that suggest opportunities for collaboration and collective action.*

Lavoie, Brian, Constance Malpas and JD Shipengrover. 2012. *Print Management at "Mega-scale": A Regional Perspective on Print Book Collections in North America.* Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2012/2012-05.pdf.

*This report maps North American collections against mega-regions, areas which concentrate economic and social activity. It provides a new framework within which to think about organizational patterns of access and management, within a new geography of collections.*

Malpas, Constance. 2013. "Subsidence and Uplift—the Library Landscape." OCLC Research Hangingtogether.org Blog on 18 April. http://hangingtogether.org/?p=2680.

*This short piece uses data about collection distinctiveness (in terms of subjects and names) to consider how HathiTrust is emerging as a significant center. This is one example of how ongoing reconfiguration will result in a rebalancing in how the collective collection is distributed across libraries, and shared print and digital repositories.*

Taken together, this work has helped shape the service and policy discussion as library collections are reconfigured by mass digitization and shared management initiatives. This theme is an important focus for us and we look forward to working with colleagues as the collective collection evolves in coming years.

## Notes

1. See http://www.europeana.eu/portal/aboutus.html.
2. See http://dp.la/info/.
3. See http://www.cdlib.org/west/.
4. See http://2cul.org/node/17.
5. See http://www.orbiscascade.org/index/mission-statement.
6. See http://www.oclc.org/en-US/worldcat/catalog.html.

## References

Dempsey, Lorcan, Brian Lavoie, Constance Malpas, Lynn Silipigni Connaway, Roger C. Schonfeld, JD Shipengrover, and Günter Waibel. 2013. *Understanding the Collective Collection: Towards a System-wide Perspective on Library Print Collections*. Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2013/2013-09.pdf.

Michel, Jean-Babtist, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, John Orwant, Steven Pinker, Martin A. Nowak, Erez Lieberman Aiden. 2011. *Science.* 331 (6014): 176-182. http://www.sciencemag.org/content/331/6014/176.abstract#aff-10.

Moretti, Franco. 2013. *Distant Reading.* London: Verso. http://www.worldcat.org/title/distant-reading/oclc/813931586&referer=brief_results.

University of Arizona Libraries. 2013. *Everywhere You Are: The Library.* Arizona: University of Arizona Libraries. http://www.library.arizona.edu/sites/default/files/users/blakisto/vision.pdf.