Too Much Data? Never Enough! Cost-Efficient Collections Acquisitions Decision Making Through Data Analysis

Jaimie Miller
University of British Columbia, jaimie.miller@ubc.ca

Kat McGrath
University of British Columbia, kat.mcgrath@ubc.ca

Eva Gavaris
YBP Library Services, egavaris@ybp.ca

Jaimie Miller, Kat McGrath, and Eva Gavaris, "Too Much Data? Never Enough! Cost-Efficient Collections Acquisitions Decision Making Through Data Analysis" (2013). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284315302

Too Much Data? Never Enough! Cost-Efficient Collections Acquisitions Decision Making Through Data Analysis

Jaimie Miller, Monograph Acquisitions Coordinator, University of British Columbia Library
Kat McGrath, Librarian, University of British Columbia Library
Eva Gavaris, Collection Development Manager, Western Canada, YBP Library Services

Abstract

Libraries are increasingly called upon to efficiently use collection dollars in creative ways. Content needs are ever increasing, and, with the growing range of format and delivery options, finding means to identify resources that provide unique, or added, value is essential.

Libraries regularly receive offers of sale pricing, or reduced pricing, for the subscription or purchase of multititle collections. Most often, these packages are for online content that the library may, or may not, have already acquired in one of the multiple formats available.

In an environment of multiple formats, ISBNs and/or ISSNs per title, variable titles, and alternate imprint or copublishing, identifying the unique or duplicated holdings of library collections becomes a major challenge.

The knowledge bases supporting booksellers, serials agents, and discovery tool providers strive to do a good job of linking content available in different formats and on different platforms.

Although these vendors robustly provide alternate format, title, provider, and imprint data on a title-by-title basis, none of their administrative tools provide the library customer with the ability to easily compare aggregate data held in the knowledge base with data extracted from a title package list.

This paper presents a description of library data needs and bookseller data provision goals, followed by a review of the power and functional limitations of current marketplace tools. Practical examples are provided of how these tools may be used to guide collection development and make wise acquisitions decisions.

Library Data Needs

As a top 20 ARL institution, the University of British Columbia Library is offered, or needs to seek, the purchase of monographic works in large batches.

Some examples include a publisher's entire output as backlist or subject collection of hundreds to thousands of titles. As vendors frequently place time limits on offers, there is pressure to investigate the suitability and cost effectiveness of an offer quickly. Often, evidence must be provided to show judicious spending for large purchases, including lack of duplication of ordering, inclusion of titles from specific publishers, subject areas, or other criteria. Due to forward budget uncertainties, one-time-only purchases have become preferred over continuing subscriptions.

There are few tools available at this time to identify duplicate content between different monograph platforms. The tools that exist should have low barriers to usage, but the authors find that this is not the case.

The challenge of identifying appropriate content and preventing overlap applies to a publisher's catalog, but also to any list. Examples include:

- holdings of a peer institution,
- aggregated collections,
- titles or publishers most frequently requested via interlibrary loan,
- titles to which an institution's faculty have contributed,
- prize nominees and winners,

http://dx.doi.org/10.5703/1288284315302

- titles reviewed in prestigious or notable publications,
- any bibliography, and
- any multiplatform or multiformat list comparison.

When considering the purchase of large monograph packages within an active collections program that combines print and electronic materials, the data needs are complex. If accepted, the mission becomes to design a package purchase that includes one copy of each title that is format agnostic from a catalog or offering of tens to thousands of titles. Format agnostic refers to a broad range of content availability options including cloth, paper, e-form publisher, aggregator, DDA, leased or perpetual purchased, individually or within a package.

Content aggregators of scholarly presses (e.g., Project MUSE and JSTOR) are unable to confirm what, where, or how much of the content is included in platform aggregators (e.g., ebrary, myiLibrary, EBL, EBSCO). Although this discussion centers on monograph acquisitions, there are parallels to the serials world, especially when considering the stability of access to leased content in subscribed or aggregated collections. There are concerns for constantly repurchasing content—perhaps first in print, then in various aggregators and/or subscribed on publisher platforms, and ultimately via perpetual online access from purchased archive collections that further incur hosting fees.

On researching titles, it is found that data elements used in key identifiers are variable and, thus, unreliable for comparison purposes. Imprint can refer to press or publishing house interchangeably. Year is inconsistently recorded as date of publication from title page version, or e-publish date. Unanalyzed monographic series may lack access points for matching. Even page count is recorded variably when pagination is identical.

Standards, Data Points and Tools

In the following paragraphs, we outline standards and access points that should be expected to provide the potential for appropriate analysis but do not. The shortcomings of each access point are explained.

FRBR

The Functional Requirements for Bibliographic Records (FRBR) uses the entity-relationship model, also used for abstract descriptions of a database, to describe four levels of representation of information objects. The manifestation is the physical embodiment of an expression of a work, such as a print book or a digital book. As an entity, manifestation represents all the physical objects that bear the same characteristics, in respect to both intellectual content and physical form. The same content in hardcover or paperback would be different manifestations, as would a PDF version. So long as the content and the physical form are the same, two objects would be the same manifestation. That is to say, two PDF versions on different platforms are the same manifestation.

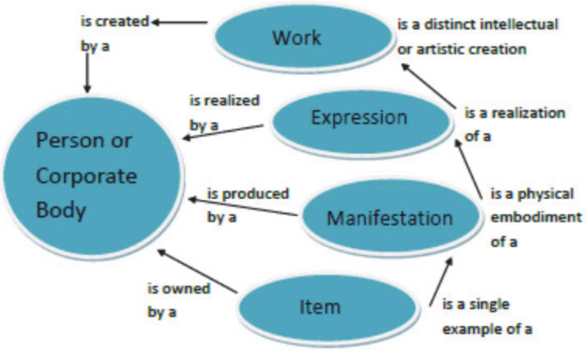

Figure 1.

Unique Manifestation Identifiers

The international standard book number (ISBN) was conceived in the 1960s as an identifier of a unique product (unique edition of a work = manifestation) and was codified in 1970 as ISO 2108. Interpretation has changed through the years as a matter of need.

> The purpose of this International Standard is to establish the specifications for the International Standard Book Number (ISBN) as a unique international identification system for each product form or edition of a monographic publication published or produced by a specific publisher. (ISO, 2005)

The ISBN Standard, ISO 2108:2005:

> Each different format of an electronic publication (e.g., .lit, .pdf, .html, .pdb) that is published and made separately available shall be given a separate ISBN. (Green, 2009)

This has been variously interpreted to imply that each platform offering a publication may choose their own ISBN, although, increasingly, multiple aggregators will all use the same eISBN to identify an identical work that each of them offers independently.

Date of online publication for a print monograph may be years, if not decades, apart. E-monograph imprint date may suggest that new content is available and be included in front list title packages, when the offer is actually legacy content.

So, use of the same ISBN for the same manifestation regardless of platform would provide an effective means of identifying, and therefore deduplicating, content. However, ISO 2108 allows for the assignment of unique ISBN if "this level of detail is required by the publisher for sales reporting" (Book Industry Communication, 2009).

Unwittingly, commercial interests have trumped the usefulness of the standard to provide a single unique identifier for each manifestation. Although the eISBN is not anticipated by ISO 2108, this clause effectively condones the practice.

OCLC and Sustainable Collections

The xISBN service offered by OCLC uses an algorithm to FRBR-ize bibliographic numbers. While this is invaluable for linking manifestations of similar works, it does not serve the library's need for linking identical works. Furthermore, this is only an intermediary step. Libraries still need to match OCLC numbers to the objects in their collection, and mismatches can easily contribute to inflated counts. (OCLC, 2014).

A more recent arrival to the library collection management marketplace is Sustainable Collection Services. This vendor offers a number of tools and services to help libraries with their print deselection processes. Using APIs that build on OCLC numbers and MARC field linking, the tools are aimed at libraries focused on developing last copy retention strategies, as opposed to collection building. (Sustainable Collection Services, 2014).

Serials Solutions and Q

Linking of manifestations is said to be programmatic with adjustments made by Serials Solutions as requested by clients. In practical use of the Serials Solutions KnowledgeWorks, there is a reasonable identification of similar manifestation. However, the KnowledgeWorks interface is designed for single-title queries. At this time there is no interface to extract the underlying metadata representing titles. In addition, while title and holdings analysis tools are provided on the platform, they are optimized for research into serial titles and holdings and provide no capability for monograph researches.

We anticipate that the forthcoming launch of Intota Assessment will bring with it new capabilities in monograph collection analysis.

ProQuest offers a mediated analysis tool called "Titles Matching Fast." However, it is intended for use as a ProQuest sales support tool only.

YBP Library Services

YBP Library Services manually reviews and catalogs books from over 1,400 publishers every year, and, during this process, links any related manifestations. Some titles are already linked when the title feeds are received from the publishers while others are not. YBP will not change the ISBN assigned to the title from the title feeds unless the ISBN has an error or seems to be assigned to the wrong manifestation, such as a print ISBN being assigned to an e-book. If the same e-book from three aggregators has three different eISBNs, YBP will leave them as is. The linking in YBP Library Services's database, GOBI, is fairly consistent and reliable, although it is still subject to human error. Errors are usually found and reported by the librarians that use the database. Libraries can view their "library history" in their own account in GOBI. In each title record, it will be noted if they own that manifestation or any linked manifestation. Libraries can use this to review titles and avoid duplication between print and electronic. Although this information can always be reviewed in GOBI for title-by-title analysis, it is not very helpful for analysis of a large number of titles.

In the case where the library might want to check if the titles included in a large package are already owned, they could load the ISBNs into a search for viewing in GOBI, but they cannot then export the same data including the library history portion. This data need has come up a lot more frequently with more libraries trying to make this sort of collection decision. As a result, YBP has received requests from libraries to run a query in the backend to add this library history data to a list of titles. YBP has often been willing to do this for customers in hopes that they consider this a unique service and benefit and continue to get content, whether it is title by title or in packages, through YBP when available. In this scenario, the vendor must have faith that use of these data will not undermine the sales relationship with the library.

This analysis of what the library owns is only as good as the data available from both the library and YBP. If a library buys a lot of their content from other sources and does not choose to load these other holdings in the GOBI database, then this search for holdings is not very useful. So far, book vendors continue to offer a reasonably good service of providing the necessary data for informing this type of collection decision and managing duplication. However it would be more ideal if customers could access the data themselves.

YBP is working on a new product development for academic librarians to support their collection decisions called GOBI analytics. More information will be released about this at a later date, but the intention would be that the library could load and manage their holdings information in the tool and use it for the analysis and management of their collection. The tool could be useful for making selection as well as weeding decisions.

Conclusion

As libraries increasingly move to ordering monograph titles in large batches, sets, series, or packages, the need for an adequate analysis tool is paramount. Libraries of any size can benefit from optimizing acquisitions spending. The lack of low barrier (in time, dollars, data, or computational resources) analysis tools effectively inhibits judicious decision making on large package purchases. The University of British Columbia Library values having the assistance of YBP bibliographic data linking to inform purchase decisions. We remain interested in other analysis solutions that benefit the library community and their partners in the vendor and publishing world.

References

Book Industry Communication. (2009). *Code of practice for the identification of e-books and digital content*. Retrieved from http://www.isbn.org/sites/default/files/images/BIC_Code_of_Conduct_e-books.pdf

Green, B. (2009). *E-books and ISBNS: Requirements for separate identification of different e-book versions*. Retrieved from http://www.niso.org/international/sc9/isbn_ebook_requirements_survey.pdf

ISO. (2005). *ISO 2108: 2005 Information and documentation—International standard book number (ISBN).* Retrieved from http://www.iso.org/iso/catalogue_detail?csnumber=36563

OCLC. (2014). *xISBN*. Retrieved from http://www.oclc.org/en-americalatina/worldshare/platform/web-services.html

Sustainable Collection Services. (2014). *Sustainable collection services: Data-driven deselection*. Retrieved from http://sustainablecollections.com/