From the Winter of Messy Data into the Spring of Standardization: E-Book Vendor Data Reenvisioned

Bonita Pollock
University of South Florida Libraries, pollockb1@usf.edu

Brian Falato
University of South Florida Libraries, bfalato@usf.edu

Xiying Mi
University of South Florida Libraries, xmi@usf.edu

Author ORCID Identifier: https://orcid.org/0000-5500-6862

Bonita Pollock, Brian Falato, and Xiying Mi, "From the Winter of Messy Data into the Spring of Standardization: E-Book Vendor Data Reenvisioned" (2018). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/https://doi.org/10.5703/1288284317015

From the Winter of Messy Data into the Spring of Standardization: E-Book Vendor Data Reenvisioned

Bonita Pollock, University of South Florida Libraries, bpollock1@usf.edu

Brian Falato, University of South Florida Libraries, bfalato@usf.edu

Xiying Mi, University of South Florida Libraries, xmi@usf.edu

Introduction

This paper will give a brief history of USF Libraries' EBA/PDA programs and the Ebooks for the Classroom Plus (EB+) database project. The benefit of standardized data and the various uses of vendor-supplied e-book data in the library projects will be discussed. Specific metadata issues related to EBA/PDA programs will be addressed along with standardization issues involving the EB+ database. Data standardization issues and data cleanup workflows will be shared along with suggestions for providing more customizable vendor metadata. Finally, a future plan is proposed to further standardize the data and employ linked data technology to improve the functionality and increase the usage of the database.

Background

The University of South Florida (USF) has been a leader in offering innovative library services to patrons and the general public. Among the innovations are creation of information portals, open access publishing, and, most recently, textbook affordability initiatives. The USF Libraries are part of a university that has been a trendsetter from its birth, becoming the first new public university "to be conceived, planned, and built in the United States in the 20th century" (Allen, 1966, p. 153). Opening in Tampa in 1960 with a student population of 1997 (Cooper & Fisher, 1982, p. xi), its enrollment now exceeds 50,000 (University of South Florida, 2018, p. 8) and includes students at campuses in St. Petersburg and Sarasota, as well as a health sciences complex on the Tampa campus.

Although the USF Libraries now hold 2,649,476 print volumes (University of South Florida, 2018, p. 20), the system's focus in recent years has been on electronic resources that can be made available to all campuses and reach students in a multitude of ways. A recent search in USF's online public access catalog showed 1,157,398 e-books were available.

The first large collection of e-books acquired came from a consortial purchase by the state universities of Florida in 2008, with titles coming from NetLibrary (now EBSCOhost eBook Collection). A patron-driven acquisitions (PDA) program for e-books was begun in 2009 from what is now ProQuest Ebook Central. As publishers began to offer evidence-based acquisitions models (EBA), which gave librarians more control over the ultimate purchase of e-books, USF moved in that direction. There are now seven e-book EBA programs at USF Libraries, in addition to the PDA program.

USF was thus well positioned to use e-books as part of a textbook affordability initiative. The Florida legislature decreed in 2008 that state colleges and universities should implement "policies, procedures and guidelines . . . that further efforts to minimize the cost of textbooks for students" (Florida Statutes, 2008). In 2016, the legislature amended the statute to require the colleges and universities to submit an annual report stating "specific initiatives of the institution designed to reduce the costs of textbooks and instructional materials" (Florida Statutes, 2016). USF Libraries decided to increase efforts to promote the use of e-books for adoption as classroom texts and also to provide material assistance in filling out the required reports.

Winter of Messy Data

One question was how best to inform faculty of the e-books available to them. Traditionally, the library's public access catalog, or OPAC, was the repository to search for what the library owned. More recently, discovery layers have become prominent. USF Libraries has always loaded all its e-books into both the OPAC and the discovery layer. But loading of the records has had several problems that compromise the integrity of the catalog.

MARC records for the e-books come in batches, either from OCLC Collection Manager or directly from the vendor. However, there is a lag time in

https://doi.org/10.5703/1288284317015

record delivery, which affects the currency of the catalog. A vendor's website may show that a particular e-book is available to USF patrons, but the title will not show up in the OPAC or discovery layer. Thus, a patron cannot get an accurate view from the catalog of what USF has available in e-books.

Getting access to the records when they are available also can be a major problem. Some vendors make it easy by allowing date range selection, so that only the records created since the last download can be retrieved and loaded in the local system. But at other vendor sites, it becomes much more complicated and the librarian has to follow a multistep process to get the needed records. The process sometimes is so complicated that a librarian may be uncertain whether all pertinent records were gathered for loading into the catalog.

After gathering the records, there can be problems with their content, particularly in the case of URLs. E-books issued in monographic series sometimes have links with the wrong volume numbers. Using the made-up series *Adventures in Metadata* as an example, a record describing volume 26 in the series would have a URL that leads the patron to volume 24 of that series.

Similarities in titles also cause problems. A record describing the title *Getting Around in Charleston, South Carolina,* for instance, may have a link to the full text for *Getting Around in Charleston, West Virginia.*

A third situation causing problems with URLs comes with records that describe multivolume sets, especially when records were received from OCLC Collection Manager. The e-books in the Loeb Classical Library provide a good example of this problem. A particular record may describe a set containing eight volumes, but there is only one URL on the record, which links to one particular volume in the set. The other volumes are represented by seven additional records, each with the same description on the record, but a different URL. When the URL is not labeled to indicate which volume it represents, it causes confusion and requires manual intervention to label the links with volume information, or, alternatively, move all the labeled links to a single record to represent all the volumes in the set.

There are also problems more particular to USF and other public universities in Florida. Since USF gets e-book packages from so many vendors, it is inevitable that some titles will be available from multiple vendors. The goal is to represent all vendors available for a particular title in the catalog, since different vendors have different use policies and a title may be represented by the PDA program, an EBA program, and a collection purchased outright. For records loaded from OCLC Collection Manager, this can pose a problem, however. After a title is loaded from one vendor, USF's holdings for the title are set in OCLC. Since records indicate USF already owns the title, a record from Collection Manager that comes from a different vendor won't necessarily be received. Librarians and staff then have to figure out what vendors are missing.

The conditions under which libraries at public universities in Florida operate can cause records that have been revised or augmented not to be loaded into a library's catalog. The state university libraries in Florida use what is called a shared bib cataloging system. Each university that has a particular title will have its individual holdings and URLs on the same bibliographic record, instead of each university having its own instance of the record in its catalog. The system is administered by the Florida Academic Library Services Cooperative (FALSC), which has coordinated library automation at Florida's public universities since they first went to online catalogs in the late 1980s.

FALSC developed the loading software that is used by the university libraries when they are batch loading records. The software is designed so that when a second university loads a record for the same title into the catalog, the existing record does not have the descriptive cataloging material overwritten. This prevents another library from erasing content in a record, but means that updated information, such as tables of contents or subject headings, will not appear in the record used by the state universities. Individual URLs for a university can be added to the shared bib record, but not other material.

Given the numerous problems that can occur when relying on the OPAC or discovery layer to determine if a particular e-book is available, librarians at USF decided on another approach. What was wanted was a database that a professor could browse for particular titles or subjects, then adopt available titles as classroom texts, thereby saving students money that would otherwise go toward purchasing a copy of the desired texts. The database developed was called Ebooks for the Classroom Plus.

Ebooks for the Classroom Plus

The Ebooks for the Classroom Plus database is based on the eTextbook Database designed by the University of North Carolina (UNC) at Charlotte's J. Murrey Atkins Library. This institution graciously shared the code for the database with USF. USF Libraries adapted and enhanced the design to meet the needs of its patrons. The USF EB+ database went live in March 2017. The database currently contains approximately 650,000 titles with unlimited simultaneous users. This includes records for 7 EBA/PDA programs and 23 other e-book vendors, along with several open access collections. In addition, some single-title-purchase e-books not owned by the library are also loaded. This allows faculty access to a wide selection of e-book titles to choose from for adoption in the classroom.

The Web interface for the e-book database is designed to be keyword searchable and displays the following fields: Title, Author, Publication Date, Platform/Publisher, Digital Rights Management (DRM), Subjects, and ISBN. Faculty are given two options for adopting an e-book for the classroom. "Access Now" signifies this title is either owned already by the library or is in one of the EBA programs and is available for use now. The "Request Purchase" button allows the faculty member to request the purchase of either an unowned title or a PDA title.

The current metadata collection process for the e-book database entails several steps. E-book records are loaded per vendor into the database using spreadsheets. First, a master spreadsheet is created with standardized headings for the various data fields. This template is used to ensure consistency and accuracy of information across vendors and to allow the data loading process to be automated. Next, entitlement lists, title lists, and KBART files are gathered from the vendor website for all the collections the library has with that vendor. This information is then compiled onto the master spreadsheet using matching formulas. Finally, the metadata in the spreadsheet fields is standardized and the spreadsheet is loaded into the database. The entire process for one vendor can take 2–5 hours of work, depending on the complexity of the information. Overall, 100 hours of work goes into updating the database each semester.

Spring of Standardization

Since the data for the e-book database comes from a variety of sources, consolidating it all onto one spreadsheet is a difficult task. The vendor title list might have the title, author, and the URL, while the entitlement list might contain the publication date, and the KBART file might have the subject headings. These three lists have to be matched on a common identifier using vlookup formulas. First, it can be difficult to find a common identifier to match the spreadsheets. Second, since the data on the master spreadsheet comes from a variety of other spreadsheets, the fields are not always formatted in the same way. For example, dates might be year only or dd/mm/yyyyy or even yyyy/dd/mm in the same master list This requires multiple cleanup procedures to standardize.

Author names are also frequently in a variety of formats, including last name only, first and last, or last, then first. Sometimes even with all the matching, some fields cannot be found and must be left blank. Usually, the most difficult fields to find are subject headings and price. All of these inconsistencies make cleaning up the metadata in the spreadsheet fields very important.

Standardizing the metadata in the database increases the reliability of the search results and makes it easier for the faculty to find relevant e-books for their courses. The e-book database team decided on several standardizations for the database. The first was creating controlled vocabularies for the Platform, Digital Rights, and Owned Status fields. The Platform field information is taken from the vendor website and standardized for all titles in that collection. The Digital Rights fields have controlled vocabulary to let users know if they have unlimited, by chapter, or by page rights to print and download. The Owned Status field allows librarians to track the e-book collections in the database. The controlled vocabulary for this field is Purchased, Subscription, EBA, PDA, Open Access (OA), or Not Owned.

Next ISBN numbers are standardized and separated into types. All ISBNs (electronic, online, hardback, and paperback) are formatted as numbers without dashes to improve searchability. Each type of ISBN has its own field in the database. This allows staff to search the database for a print book by ISBN number and find the e-book equivalent in the database.

Titles sometimes are broken out into title and subtitle columns in the original vendor-provided data. They are concatenated into one field with a semicolon delimiter between the title and subtitle. Additionally, the author fields are combined into one

column when the author name had been split into two columns. These changes improve the indexing and increase the accuracy of title and author searches. Finally, the Publication Date and Online Date fields are standardized and reformatted to show year only. This eliminates many of the date formatting issues and improves the user interface by making all the dates consistent.

Future Solutions

The e-book database team has outlined plans for future enhancements to the e-book database. The first enhancement would be to create a controlled vocabulary for subject headings. This will be an extremely difficult task since each vendor uses its own vocabulary for subject headings. Therefore, it would require cross-walking each vocabulary into Library of Congress Subject Headings.

The team would also like to create a separate field of discipline that would match with the courses offered at USF. These two enhancements would make it easier for professors to find e-books in the subject area being taught in the course.

The second enhancement deals with improving the title and author standardizations. Titles typically have additional information included in the title such as edition or volume number. The team plans to break this information out into separate edition and volume fields. This will make it easier to identify the volume or edition of an e-book in the database. Authors frequently have institutional affiliations attached to their names, which needs to be deleted to improve the author search accuracy.

The e-book database team is considering adding a vendor field to the database due to the fact that some platform names like Ebook Central do not include the vendor, ProQuest, in the name. Having a separate vendor field would make the database searchable by vendor as well as platform. This would be extremely helpful for vendors such as Oxford who maintain several different e-book platforms for their various collections.

The e-book database team is also looking for ways to improve the process of updating the database. One way to do this would be to create a separate electronic resource management (ERM) database for the purchased, subscription, and open access titles. These titles are more stable because they don't change as frequently as the EBA collections. Once the metadata in the ERM database is standardized and loaded, it would only need to be updated when new material is added to the collection. This would be a big improvement on the current process, which completely reloads all the collections each semester in order to catch all the EBA title changes.

The EB+ database then would only contain the EBA/PDA programs, which are updated frequently, sometimes even weekly. The time saved not reloading the purchased records could be used to keep the EBA titles more up-to-date. In addition, because the EBA records are fluid and not a permanent part of the library collection, metadata standardization on these records would not be as critical. The Web interface would get feeds from both databases and make one consolidated e-book search display. This would be the most beneficial and cost-efficient enhancement for the EB+ database.

USF Libraries is currently exploring options in linked open data (LOD). The e-book database team is investigating the possibility of creating the e-book ERM using RDF triples. This would allow the team to store the data in a triplestore database and create SPARQL queries to interact with the database. The premise is that LOD would improve the search results accuracy of the database. The team plans to conduct user ability studies before and after the implementation of the LOD technology. These studies will assess how search functionality has been affected.

In conclusion, the USF Libraries have received some very positive feedback from faculty and staff on the EB+ database. Approximately 350 e-books have been adopted for the classroom since the database premiered. E-book usage in the library has increased since the database has gone live, and the Textbook Affordability Team uses the database daily to find resources and make suggestions to faculty about textbook alternatives. By creating a separate ERM database and maintaining it with quality metadata, the Ebooks for the Classroom Plus database will be further enhanced and be an even greater success.

References

Allen, J. S. (1966). The University of South Florida. In M. G. Ross (Ed.), *New universities in the modern world*. New York: St. Martin's Press.

Cooper, R. M., & Fisher, M. B. (1982). *The vision of a contemporary university*. Tampa: University Presses of Florida.

Florida statutes. (2008).

Florida statutes. (2016).

University of South Florida. (2018). *USF system facts, 2018/19*. Tampa: University of South Florida.