Purdue University

Purdue e-Pubs

OARS: Toward Automating the Ongoing Subscription Review

Geoffrey P. Timms
Mercer University, timmsgp@gmail.com

Jonathan H. Harwell
Georgia Southern University, jharwell@georgiasouthern.edu

OARS: TOWARD AUTOMATING THE ONGOING SUBSCRIPTION REVIEW

Geoffrey P. Timms (*timms_gp@mercer.edu*) *Electronic Resources & Web Services Librarian, Mercer University*

Jonathan H. Harwell (*jharwell@georgiasouthern.edu*) *Coordinator of Content Management, Georgia Southern University*

BACKGROUND

Thorough assessments of subscriptions are unwieldy and time-consuming to perform every year. A metric has been developed for standardizing the process. This session will share details and engage the audience in refining the metric. The goal is to build an open-source **O**ngoing **A**utomated **R**eview **S**ystem (OARS) for subscription reviews.

We begin with some context about library budgeting challenges. Then we share a project that is being developed to help streamline the subscription review process.

Jonathan Harwell came to Georgia Southern University in August 2007 as the Collection Development and Assessment Librarian. Classified by the Carnegie Foundation as a "Doctoral-Research University" (DRU), GSU has over 19,000 students, including over 2600 graduate students (about 14%). They now have their first PhD program in Logistics and Supply Chain Management, along with five other doctoral programs in education, public health, nursing practice, and psychology; 31 master's programs; and a total of 121 degrees offered.

In FY09, the library collection budget was cut by $300,000, and further in FY10 by $470,000. These were year-end distributions of funds which had been sustaining the library budget as annual supplements. These funds had allowed ongoing subscriptions that were not affordable otherwise. By FY11, the library was only expecting its long-standing base collection budget of $1.2 million. While the library budget represents a significant percentage of the institutional operating budget, the per-FTE spending is rather low in comparison with peer libraries. This is due to the State of Georgia's low level of funding for the university itself.

With only three librarians and 12.5 support staff at the time, the library's Collection & Resource Services Department devoted significant time and effort, in collaboration with other librarians and academic departments across campus, to execute a full-scale subscription review in 2009. The goal was to balance the FY10 library budget without depending upon a year-end distribution of funds from the university administration.

All ongoing resource costs were analyzed, including databases, serials, and standing orders. The Serials Librarian and Assistant had produced spreadsheets of all items, along with pricing data. Three spreadsheets were produced: one for databases; and two with other subscriptions sorted by title or by call number. Pricing data was not supplied for databases due to confidentiality in some licenses; but the costs of all other subscriptions were listed. Harwell, in collaboration with the Dean of the Library and the Subject Specialist Librarians, communicated with faculty campus-wide to explain and facilitate the review process. Each department was asked to provide a rating for each title, with three options: E for essential, D for desirable, or N for not needed. They were also encouraged to specify the format preference with P or O, e.g., EO, DP.

Harwell collated all of the feedback and determined that the total cost of the "not needed" titles was less than the necessary budget cut. He also analyzed the format preference and found that the cost of subscribing to online-only format was in many cases only a few dollars cheaper than print plus online. However, due to the number of titles preferred in online format by the faculty, even small amounts added up significantly. It was determined that if $1 or more could be saved by switching those titles to online only, then it would be beneficial to do so. The budget was still too high, so the library faculty met and determined that some "desirable" titles should be listed for cancellation, according to cost, Eigenfactor, and professional judgment. A version of the spreadsheet with proposed cancellations highlighted in yellow, and proposed switches to online-only highlighted in green, was distributed campus-wide for feedback before the final cancellation decisions were made. Faculty requested that several titles be reconsidered, and these were removed from the cancellation list. A final cancellation spreadsheet was then distributed, using the same color coding. The database cancellation list showed alternative resources recommended by the librarians.

The entire process was time-intensive, especially for a few specific library employees. Preparation began in mid-November 2008, and the lists were released to the faculty in mid-February 2009. Most of the feedback analysis occurred during the summer, when Harwell had to forgo other responsibilities, such as serving at the Learning Commons desk. Final cancellations were announced in late August 2009.

Along with the time demands, there were other disadvantages to the process. It was not ideal to rely solely upon faculty recommendations for most titles, with cost and Eigenfactor only being analyzed for those rated as "desirable," and cost per use not being factored in except when analyzing databases or certain e-journal packages.

Harwell and Geoffrey Timms, the Electronic Resources & Web Services Librarian at Mercer University in Macon, Georgia (two hours' drive apart), discussed the process and brainstormed about ways to improve it. Timms, as a self-taught computer programmer, was familiar with some coding that might facilitate a process which could automate the selection of titles to be considered for cancellation. (Another colleague, who prefers anonymity, actually had the initial idea and participated in the brainstorming.) In 2009, Timms began work on an open-source program that could analyze multiple data points and calculate them as factors, thus automating the most time-intensive portion of the budget review. Timms and Harwell conferred via phone, Facebook chat, and in-person meetings, and in 2010 Timms produced an online demonstration module of OARS, the Ongoing Automated Review System.

OBJECTIVES

OARS will automate the selection of titles and data for review, which will result in a semi-automated process for annual renewal decisions. OARS will utilize multiple, customizable variables with adjustable scales, as well as a cumulative weighted scale of all variables which the system will use to recommend a renewal decision.   OARS will use automated processes as much as possible, and will also provide data entry forms for easy input. There will also be an interface for stakeholders to view the data and respond to the review.

THE TECHNICAL SIDE

Developed by a self-taught novice programmer working in a L.A.M.P. (Linux, Apache, MySQL, and PHP) environment, the OARS platform is coded primarily in PHP with some JavaScript added to improve functionality. At the outset of the project, the programmer did not know enough of either of these languages to complete the interface, and spent much time learning

techniques to achieve specific goals within the project. At the back end of OARS are two tables in a MySQL database. PHP and MySQL interact using the PEAR MDB2 database abstraction layer.

While records in the database contain much human readable information, four data items are used for the ranking of serial resources. These data items are:

- Cost - which is used to calculate *cost as a percentage of the average cost* of all serials represented in the OARS database
- Usage - which is used to calculate *usage as a percentage of the average usage* of all serials represented in the OARS database
- Rating - which is used to assign a percentage value to a simple expression of necessity (essential=100%, desirable=50%, or not needed=0%)
- Eigenfactor Article Influence Percentile - which represents the average influence per article for a particular journal.

The calculated variables are then converted to scores on a scale of 100. As lower subscription cost is preferable to higher, cost is scored with the formula:
Cost Score = (-0.5 x Cost as a percentage of average cost) + 100

Thus, if the cost as a percentage of average cost is 100%, the score is 50; if it is 50%, the score is 75; and if it is 150%, the score is 25. Conversely, usage is judged to be preferable the higher it is, so usage is scored with the formula:

Usage Score = 0.5 x Usage as a percentage of average usage

Thus, if the usage as a percentage of average usage is 100%, the score is 50; if it is 50%, the score is 25; and if it is 150%, the score is 75. In each case, the scores are capped with a minimum of 0 and a maximum of 100.

The variable scores are then weighted and combined to create the final OARS Score for each serial. The weightings are managed in a form (see below) and permit the user to adjust the impact of each variable score in the calculation of the final OARS score. By virtue of the relative nature of the cost and usage scores, OARS scores are not static for any given resource. An adjustment of cost in one resource, for example, changes the average cost, thus changing the cost as a percentage of average cost for all serials. An OARS Score, then, is valid only in the context of the report in which it was created. This ensures that each serial is re-evaluated in each new report, relative to all of the other serials represented in the database.

THE INTERFACE

Data entry is achieved either by completing a form to add a single record to the database or by uploading numerous records in a Comma Separated Value (CSV) file. In the case of the form, data is validated upon submission with JavaScript and is modified upon successful validation to ensure homogeneity. The upload of data in a CSV file also undergoes a validation process and invalid data is rejected and made available for download in a new CSV file. Successful data uploads are also dependent upon the absence of duplicate records for a given title. Duplication is tested for a given year by comparing the unique control number of the record with those in the database. The control number system may be locally established or a standard system, such as OCLC number, may be used. Duplicate records and successfully uploaded records are reported on-screen in tables and are also made available for download in CSV files.

Data retrieval and management is made possible by a search interface. Most of the fields in the database are made available for searching in a three-tier, Boolean-enabled search interface with sorting capability. Text fields are searched using MySQL Natural Language searching while the Library of Congress Call Number (LC) field is searched using MySQL Boolean searching. This facilitates the use of the asterisk wild card. Upon conducting a search, most fields from each record are displayed, including icons to view, edit, or delete the record. By selecting any of these icons, further transactions are completed in a new window which disappears when a transaction is properly completed. The search results are then refreshed automatically to account for the changes made.

OARS can be configured for local preferences when generating a report. Each of the four variables from which the final OARS Score is calculated can be weighted according to how much emphasis is desired. Weighting, which is stored in a MySQL table, is adjusted in a form and uses percentages in increments of five percent, with the total weighting of all four variables compulsorily, using JavaScript, equaling 100%.

The OARS report presents records in tabular form from lowest to highest scoring. The purpose of the report is to assist with serial cancellation decisions, and to this end, two methods are offered for defining the scope of the report. Users can specify a bottom percentile of the collection for potential cancellation or can specify a dollar amount to cut from the budget. In each case, OARS will list records in the report until the percentile or dollar amount has been reached. The report is made available for download in a CSV file.

PROBLEMS/CHALLENGES

Aside from the challenge of learning multiple coding techniques on the fly to achieve intermediate goals, several challenges stood out as significant. LC Call Numbers and MySQL Boolean searching did not work particularly well together because the spaces and periods which characterize LC Call Numbers act as word breaks in MySQL. In addition, MySQL was unable to accurately sort LC Call Numbers. The solution to this challenge was to normalize the LC Call Numbers to a standard format for storing in the database and for retrieval using MySQL searches. Thanks to Bill Dueber's excellent work on LC Call Number Normalization in Perl, the workload in adapting the process and implementing normalization and reverse normalization processes for OARS was substantially reduced. Mastering regular expressions was perhaps the most time-consuming element.

Another problem is the minimum length of words imposed by MySQL for natural language searching. A four-letter minimum means that some of the acronyms like ACM or ACS found as publisher names cannot be located. While it is relatively simple to adjust one's MySQL configuration to account for this, it would necessitate each institution using OARS to make such a configuration change. The issue currently remains unaddressed.

CRITERIA FOR RENEWAL RECOMMENDATIONS

All of the variables and the draft requirements for OARS were shared during the session. The variables, as explained above, consist of faculty ratings, relative cost, relative usage, and Eigenfactor percentiles. Analysis of test data with actual statistics from Georgia Southern University shows that some titles were rated as "excellent," but had low usage, and others were "not needed," but one of those was actually the most used title in the test data. Basing

cancellation decisions on all of the above factors is more justifiable, in the context of data-driven decision making, than cancelling all "not needed" titles and renewing all "essential" titles solely based upon faculty recommendations. Because the variable weights can be adjusted easily on the fly by each library, more priority can be given to one or more factors according to the local situation.

FUTURE STEPS

Harwell and Timms presented OARS at three national conferences in 2010: the Acquisitions Institute at Timberline Lodge, the LITA National Forum, and the Annual Charleston Conference. They were encouraged by the feedback, with several colleagues lamenting that no vendor had created such a product and that so many libraries have to invent their own time-intensive processes when faced with budget cuts, which have recently become the norm in library funding.

Harwell plans to convene an online focus group with several interested colleagues around the country, some of whom are eager to use OARS at their own institutions. We will also make contact with librarians at another university who delivered a somewhat similar presentation at the Charleston Conference, to invite collaboration on a single, open-source, final product. In 2011, pending further feedback and possible modifications, the OARS code might be ready to be released to the public.