Cite as: Hutchinson, Alvin. 2013. “Showcasing Scientific Research Output: New Audiences for Science Libraries.” In How to STEM: Science, Technology, Engineering and Math Education in Libraries, 151–60.

Showcasing Scientific Research Output: a new audience for science libraries

by Alvin Hutchinson

New audiences: As academic library services move to a self-service model, science librarians face a clientele who require less and less assistance in identifying and obtaining scholarly material for research. Scientists at most universities today can search literature indexes and access articles from their office or lab and can find and purchase affordable books via a well developed online market. For these and other reasons scientists visit their library (and perceive a need for their librarian) less and less each year (Niu et.al, 2010). University and other research librarians must develop new services and find new audiences if they want to avoid irrelevance.

The Smithsonian Libraries began a new set of services in 2006 in anticipation of this growth in self service. The Smithsonian Research Online (SRO) program documents the published research output of Smithsonian scholars and fills a longstanding but inconsistently managed need for a master list of books, chapters and articles resulting from Institutional research. Standardization of digital publishing and the availability of bibliographic management tools have made compiling a list of publications by a given research institution possible with relatively little manual data entry required. Administrators, webmasters, public affairs and other groups at the Institution have expressed enthusiasm for this non-traditional service which grew in part out of the recent emphasis on evidence-based research assessment.

Research assessment: One of the more realistic strategies for 21st century science librarians is participation in the research process outside of simply acquiring and storing reading material in anticipation of future user needs. It is clear that scientists have more work to do outside of simply collecting and analyzing data and some of this work is suitable for assistance by librarians. The management of a scientist’s bibliographic data, advising on data management planning and scholarly publishing alternatives are among the non-traditional services which librarians are increasingly providing (Kroll and Forsman, 2010).

But many research organizations have needs which science librarians can fill outside of direct researcher support. One is in assisting with the research evaluation exercise that every scholar faces during their professional career. Because the publication record of a scientist is a key component of his/her evaluation, the compilation of an institution’s publication output is a highly valued service and one that librarians can provide easily and reliably thanks to automation in scientific publishing.

Identify: One primary appeal of this kind of service is the low initial costs. The availability of bibliographic data in a standard format for easy collection makes starting a research publication registry something that can be undertaken immediately. The first step is to identify the publication data from among the vast body of literature available online. There are many licensed and freely available sources which allow users to search and find publications emanating from a particular organization. Commercial publisher websites along with sites like Google Scholar, BioOne and Highwire Press allow the capture of bibliographic data without charge while subscription resources such as Web of Science provide a single site with comprehensive coverage and more elaborate search features. Most useful are those databases which allow a field-specific query to restrict the search term to an author affiliation or author address field. In this way the university or institution, department, research center or even zip code can be used to filter out

"noise" from general keyword searches.

Where the affiliation or address fields are not searchable themselves, a general keyword search is still useful although with the understanding that the results will contain some false hits. For example, a general keyword search on "Ohio State University" in Google Scholar may return extraneous results but whose bibliography includes a reference to a book published by the Ohio State University Press. Additional false hits may occur when the university name appears in the acknowledgement section of the paper.[1]

After an accurate search strategy is developed, most online bibliographic services allow users to create and save alerts so that email messages are automatically sent as notification when new publications are added that meet the specified search criteria. Some also allow for the creation of RSS feeds as an alternative system of notification.

It should be noted that most of the online sources listed here cover journal literature extensively, but books and chapters are frequently overlooked. However in recent years commercial publishers are giving books and chapters a DOI and their own web page which includes exportable bibliographic data. This trend suggests that the automated identification of books and chapters in the sciences will become easier in the near future.

Capture

The proactive search detailed above is the first step in automating the process of data collection and fortunately there are several free and low-cost methods to get the data from the publisher or aggregator website to a database on the librarian's workstation. Most web resources which allow search and alert services also publish their data in a format which conforms to standard bibliographic management software. Since the 1970s, there have been few options in this class of software but in the past 5-10 years, the options have expanded greatly. Both Zotero and Mendeley, for example are freely available resources which capture, format and permit sharing of publication data. Both allow data to be stored or mirrored in the cloud which permits simultaneous access so that multiple users can edit records. The ingest of citations via web browser plugins makes capture very easy.

There are also several commercial products for managing bibliographic citations but because multiple library staff members work on editing the data, the SRO requires a networked solution. One commercial option, RefWorks™ is web-based and therefore allows access from different workstations. We use RefWorks to compile, edit, review and de-duplicate records both because it is web-based and because it allows library staff to globally edit author names, journal titles and descriptors via an easy-to-use interface. Fortunately the transfer of data between and among these systems is standardized today so that citations captured from web sites via Zotero or Mendeley are easily exported to RefWorks which serves as a master copy of the database.

Once the data is reviewed and edited in RefWorks, we use a customized output format to migrate new and modified records into an SQL database which is mounted on an institutional web server. It is impractical to export the entire collection when refreshing the website database therefore the date of last modification for all records is critical to maintain as it helps in identifying

[1] This type of data can still be leveraged by librarians, however since some offices on campus (e.g. public affairs) may want to be notified of and collect acknowledgements by outside authors.

which records to move from RefWorks to the web server. This website (research.si.edu) allows anyone to search and filter publications authored by Smithsonian staff.

Pro-active Collection
Collecting publication data for a research institution would not be possible if it were left up to authors to contribute and enter the data themselves. The recent experience of institutional repositories for example, shows a general unwillingness by authors to enter their publications into a repository. But for the SRO, the automated search, filter and capture of this data results in approximately 75% of science publications being added to the database without any effort from or notification by the scientist.

However publisher/aggregator websites alone will not provide a comprehensive list of publications for an institution--even in the sciences. Inevitably there has to be some manual data entry for publications which are not online or which are not formatted for reference management software capture/import. The SRO includes a web form to manually capture the approximately 25% of publications which are not collected via regular stored searches. Using web scripting language, the data elements entered into this form are sent to library staff in a standard format called RIS which reference management software can read. The RIS format is importable by all major bibliographic software packages and at the same time easily understood by library staff when viewed as a text file. Data appears as one field per line, each using a standard two-letter prefix identifying the data type. For example PY is publication year and VL is volume, etc. Formatting manually-entered data in this way is not difficult as the RIS data fields and their prefixes are documented on the web.

False hits Although some online sources retrieve false hits as noted above, this is not always useless information. The university press and/or the public affairs office among others, frequently like to hear of the university, its labs or museum being cited or mentioned in an acknowledgement. And since the information has already been retrieved, it takes little effort to forward the citation to the appropriate staff person thereby cultivating support from another potential library user group.

Duplication: One burden of freely available bibliographic sources is that there is often duplication. HighWire press, Google Scholar and other aggregators cover a wide swath of literature and often pick up the same citation, sometimes weeks apart. The Thomson Reuters product, Web of Science often includes items after they have been captured from a publisher website or picked up by a Google Scholar email alert. As with all standard library practice, every citation should be checked for duplication prior to being added to the database. And while not always practical for large institutions, having a single person oversee this kind of service allows him/her to become individually familiar with the stream of papers and oten identify suspected duplicates from memory.

Pre-press or Online First: Automatic capture of this data, while a relief to authors presents some policy issues which influence workflow. Aside from duplication as noted above, alerts provided by many bibliographic systems often identify publications as soon as they appear online--not necessarily when they are issued in print. It should be decided whether an “online-first” version is collected and whether the item should be followed up with enumeration and pagination after appearing in print. On the one hand, a DOI provides an un-breakable link to a paper so that volume/issue/pages information can always be retrieved when needed. But many

scientists prefer to have as complete a citation as possible and this means a subsequent identification and cleanup of these “incomplete” records should be built into the workflow. The SRO includes items immediately after appearing online but has built in a stored script to identify them for follow up so that enumeration and pagination can be added later.

Editing: Once collected, citations can be tagged in a number of ways depending on the desired reports to be generated. Among the most obvious is by the department with which the author is affiliated. The SRO database is tagged with museum and/or department name (e.g. National Museum of Natural History). Additionally, tags could be assigned for the status of all institutional co-authors (e.g. post-doc, fellows). They might also be tagged with grant funding source (e.g. federal v. private), collaborating organizations, a particular lab name or other relevant data. Unfortunately identification of other aspects of authors, co-authors and sponsors often is unknown by library staff.

While it is helpful for a single person to review all citations coming into the database, at some level of publication output it is impractical and the work needs to be distributed among a larger pool of staff. An ideal situation would be for the subject specialist, selector or reference librarian who works directly with certain departments to assume oversight of the data, tagging as appropriate and monitoring publication trends among their direct library user group. This also has the effect of embedding librarians further into the research lifecycle beyond simply acquiring reading material for researchers. Serendipitous identification of duplicates would also likely be retained assuming that a subject librarian would review all publications authored by the same set of scientists and develop some familiarity with their work.

Usage/Reporting

Beyond collection and reporting for research evaluation purposes, the SRO service is leveraged to serve additional audiences without much effort. Because it is stored on a SQL database on an Institution web server, Smithsonian webmasters can reuse the data to create dynamic publications lists on web pages for individual scientists or their labs. The data is displayed in real-time, this relieving both web content editors and scientists from the chore of remembering to and actually updating these pages which may have been in static HTML format before. This kind of resource has strengthened the relationship between library staff and webmasters along with the central IT office--a group which traditionally has been a regular user of library services.

While this is a positive development, it should be noted that for those scientists or labs where their publication data is redisplayed with the integration of this database, there will likely be a higher level of corrections and feedback given to the librarian(s) who compile the data. Higher visibility means greater scrutiny and library staff should bear this in mind.

Of course, many scientists would like their web pages to display the entire corpus of their publications beginning when they were in graduate school and including those written while at other institutions where they worked prior to the Smithsonian. For this reason the SRO began accepting publications issued prior to a scientist’s arrival at the Institution if they wanted to include the data. But for purposes of metrics and to avoid artificially inflating the publication statistics for past years, a certain keyword was assigned to these which identified it as being published prior to affiliation with the Smithsonian. That way they can be retrieved/displayed and metrics reports for publication totals can exclude these records.

As mentioned earlier, public affairs staff of many scientific organizations are interested in recent activities of the research units and they use the SRO to keep abreast of recent activity. Where in the past they may have had to actively seek content for press releases, the publication registry now serves as a source of regular reports on the latest papers reflecting recent research. Public affairs is yet another user group that has largely been overlooked by research libraries and this new relationship may help to offset the reduced contact with traditional users doing research.

Research Metrics. Bibliometrics is a primary component of research evaluation and an master list of publications is essential for this kind of review and analysis. But the research evaluation process extends beyond publications to include elements such as grants and awards and it is hoped that the SRO could one day be integrated with a database listing grants or funding for scientists. This takes on a greater importance in light of the recently growing trend of some funding bodies to require open access to the publications and/or the data sets resulting from awards. Collecting an institutional publications list is the first step in creating a unified research registry which could include information about publications, grants, scientific data sets and more.

Alt-metrics: The SRO data is useful for collecting bibliometric data for the Institution's scholarly output and while the impact factor and h-index are available for most publications, the field of alternative metrics helps to measure readership and access to SRO publications without waiting for formal citation counts. Alt-metrics includes a variety of measures including a count of the number of times an item has been downloaded or viewed; mentioned in popular news, blogs or twitter; bookmarked and/or included in other social media, all of which might offer an additional perspective on the individual papers of a scholar. The SRO program's popularity has cultivated enough support to hire an outside vendor which specializes in altmetrics to analyze the publication data in order to demonstrate readership and usage of publications beyond standard citation rates.

Repository Metadata Generation

Like many research libraries, the Smithsonian Libraries manages an institutional repository (IR) of scholarly reprints authored by Institution staff. Learning from the experience at other institutions, we knew that asking authors to enter and upload reprints to the repository would not be met with a high rate of participation. We decided to implement the more realistic mediated deposit strategy whereby library staff would collect and upload reprints on scientists' behalf. And the SRO has provided a base for the service. With metadata collection already captured and "databased," it is possible to generate XML-formatted records for batch import into our repository (currently using the DSpace platform) and circumvent manual data entry that might otherwise have been required. A web server script takes the SRO data and returns it in a format which DSpace can accept for bulk upload. The only additional step is for library staff to match the digital reprints (provided by scientists) to the appropriate record in the database. This emphasis on batch processing of repository content has resulted in an above average number of items in the SRO repository compared with many other library-managed repositories. By creating a custom metadata field in DSpace which corresponds to the unique identifier for each record in the SRO, the data as displayed on institutional websites contains links (where applicable) to reprints in the repository.

System Architecture: Ideally, SRO data would reside on a single server (backed up of course) rather than as different versions (e.g. RefWorks, SQL web server, DSpace). Multiple copies of any database easily become un-synchronized and edits made in one place must be made to the

others. The primary obstacle to establishing a single, unified database instance is the absence of bibliographic management software with a user-friendly interface which would allow a wider group of staff to simultaneously log in and edit/update records, including batch updates to multiple records, standardizing author and journal names, etc.

It has been difficult to find such a bibliographic management system as most reference management software is not created for this scale of work. For example, while some allow simultaneous login and edit, they do not allow a user to perform any global operations. There are repository platforms which offer a metadata (bibliographic) management capability which may be suitable for the SRO and some include the ability to create researcher pages highlighting individual scholars. But because the SRO program began as a separate effort from the repository, the two systems were well-established and SRO services dependent on the SQL database before the opportunity to easily integrate them had passed. The consolidation of the data onto a single platform is something which will have to be more carefully coordinated sometime in the future.

SRO program staff are currently evaluating options for a bibliographic management system which can accommodate multiple users and accounts so that the work of editing data can be shared more widely among Smithsonian staff. Were a more robust system identified and implemented, this data review and editing could be crowdsourced so that authors themselves (or their support staff) could tag records with appropriate descriptors identifying funding sources, sponsorship, departments, collaborators and other information that the Smithsonian research community would find useful but that library staff are not in a position to provide.

Staff

The SRO has approximately 2.5 FTEs collecting, editing and reviewing over 2000 items each year. In addition to current publications capture, there are legacy projects to add books, chapters and articles which have appeared in printed bibliographies from past Smithsonian annual reports and other historical publications lists. Participation in the program has grown from one professional staff person in the Digital Services division to now include several others including para-professional staff who work in what has traditionally been bibliographic control of the library's online catalog records.

Many of the suggested improvements below would require wider intervention in the review and edit of the data by either authors themselves, their designated departmental appointees, and/or a broader group of SIL staff. As mentioned earlier, this is dependent on the development of a more robust database management system.

Future Improvements

Implementation of researcher, institutional or fund identifiers. There are several movements to create unique identifiers for scholars, their institutions and the grants which they have been awarded. These are analogous to the DOIs used to identify electronic publications. The use of these identifiers may help to monitor Smithsonian scholarly activities more consistently and comprehensively.

Calculation of open access membership fees: Many open access (OA) publishers offer discounted article processing charges (APC) for institutions who pay a membership fee. A count the number of SRO publications in those journals for which publishers offer institutional

memberships may inform a decision on the cost effectiveness of a Smithsonian membership. This membership threshold and our level of authorship in relation could be reported annually.

Harvest SI-authored papers from OA publishers: A number of Smithsonian-authored papers appear in open access journals but for one reason or another, the SRO has failed to include the digital reprint in the Repository (despite being freely available). We could mine sites like PubMed Central, PLoS or BioMed Central for articles where we lack a reprint and download/add to the Repository.

Text-mine the Digital Repository: with over 16,000 reprints in the Repository, this body of text could be mined for a variety of data such as place names, personal names, corporate names, longitude and latitude, museum specimen numbers, species names, etc. In turn, this data could be mapped using a public facility such as Google maps or used to generate a legacy list of funding sources, institutions with whom the SI has collaborated, links to museum specimen databases, etc.

Conclusion

The SRO program is extremely popular primarily because it serves audiences which have previously been somewhat overlooked at the Institution compared to traditional researcher/readers. The primary drivers of the program--Institutional administration and those involved in science policy--do not typically use library resources as regularly as scientists themselves and the SRO is a way to remind them of the value of library services.

Works Cited:

Kroll, Susan, and Rick Forsman. 2010. A slice of research life: information support for research in the United States. Dublin, Ohio: OCLC Research. http://www.oclc.org/research/publications/library/2010/2010-15.pdf.

Niu, Xi, Bradley M. Hemminger, Cory Lown, Stephanie Adams, Cecelia Brown, Allison Level, Merinda McLure, Audrey Powers, Michele R. Tennant, and Tara Cataldo. “National Study of Information Seeking Behavior of Academic Researchers in the United States.” *Journal of the American Society for Information Science and Technology* 61, no. 5 (2010): 869–890.