Research at Smithsonian – presented by Smithsonian Libraries

Suzanne Pilsk, Head, Metadata Department, Digital Programs and Initiatives, Smithsonian Libraries. PilskS@si.edu

Smithsonian Overview

Smithsonian Libraries (SIL) [library.si.edu] established Smithsonian Research Online (SRO) in 2006 [research.si.edu] as a direct response to a specific need from the Smithsonian’s Undersecretary for Science for information about scholarly publications of the Institution. After an initial phase focused on the science output of the Institution, the wider potential for the entire span of Smithsonian [Smithsonian.si.edu] research was apparent. The Smithsonian is composed of 21 museums and research centers and the National Zoo. These facilities are located in Washington, DC, as well as museums/research centers in New York, the Chesapeake Bay, Cambridge, MA, and the Republic of Panama. Additional research stations are located across the United States and internationally. The disciplines and topics covered by the Smithsonian range from A (Arts) to Z (Zoology) with global geographic coverage.[1]

Smithsonian Research Online (SRO)

Simultaneously with the creation of a bibliography of research articles, book chapters, proceedings, and monographs, SIL launched a digital repository hosting pdf, text (.txt), and standard Microsoft products (Word, Excel) in an open source software platform (DSpace). The conjunction of providing citations to publications and a place to store the publications solved an Institution-wide need.

An identified need, however, does not bring willing participation. The Library staff knew from the wider repository community that the citations and publication deposit would not easily and painlessly be contributed by the Smithsonian research community. There was immediate buy-in for SRO from a few researchers willing and pleased to have someone manage their publication output and provide a persistent link to the full text of publications. Other researchers had systems in place--though not necessarily sanctioned by the Institution--which minimally served their needs and for which they saw no need for change. This
disparate and siloed solutions were unable to provide metrics and statistics for Smithsonian administrators. The growing need to track the Institution’s research output and associated impact metrics were a driving force for consolidating and supporting SRO at the senior management level.

Smithsonian Libraries staff worked with existing subscription databases and indexing services for automated collection of, at a minimum, the citation to the research output of the Smithsonian curatorial and scholarly community. Using standard search phrases, SIL staff were successful in filtering searches on various forms of the Smithsonian addresses, zip codes, and abbreviations. These automated queries provided far from 100% coverage, but provided a learning experience in the way publishers and researchers use a variety of abbreviations and acronyms.

Libraries staff also understood that, though not without problems, the sciences were easier to tackle than the humanities. Science oriented indexes and databases cover their topics more thoroughly, the publication cycle is more frequent, and Smithsonian research units were more easily identifiable. Putting aside the humanities research at the Institution for a moment, the success of the Smithsonian Research Online was immediately evident when Libraries staff began providing monthly updates to administrators. These updates included citations and full text links for the major Smithsonian science units and were recognized as valuable by senior management as well as public affairs units of the Smithsonian.

The underlying mechanics and inefficiencies of the SRO platform and workflow are evidence of the growth from an initial need to a full program. Still a fairly labor intensive operation, there have been progress towards making some of the workflow more automated with scripts and bulk loads.

As of the end of the 2016 Fiscal Year of the Smithsonian, the Smithsonian Research Online includes 80,000 citations from 20 museum and research units and 20,000 titles in our institutional repository.

**SRO Workflow**

Using off the shelf products, citations are gathered from indexing services such as Web of Science. Alerts have been set up with publishers such as John Wiley and Springer. Experiments with alerts with queries specific to Smithsonian in Google Scholar are being tested. Many of these citations need review to establish of the searches are accurate and returning publications associated with Smithsonian staff.

Zotero, a free citation management system that can be integrated in web search applications, is used by staff who use the alerts to quickly capture citation information in the standard RIS format. RefWorks is an off the shelf RIS [2] based product that is use intensely in the SRO
process for global personal name correction, journal, and ISSN corrections, etc. Citations aim to be complete with name of the authors, internal affiliation within the Institution, title of the article, host publication title, year of publication, start / end pages, and standard identifiers including DOIs and ISSN/ISBN. The RIS data is mapped into the sql database that stores a unique identifier.

SRO also provides a very simple web form for staff to enter their own citation information. Full text of articles (with PDF the preferred format) are submitted to the SRO project manager to be included in the repository.

Staff compare metadata with the submitted full text and provides the links to resolve the citations in SRO to the full text loaded in DSpace. Duplication is a consistent problem. Due to the reporting requirements from administrators, each author needs to be assigned an internal qualifier to indicate the research unit – a manual process. We have also begun to institute our own local identifier, and when possible, the ORCID[3] if a researcher has registered.

<table>
<thead>
<tr>
<th>Name</th>
<th>ORCID</th>
<th>Profiles ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marra, Peter P</td>
<td>0000-0002-0508-7577</td>
<td>marrap0242008</td>
</tr>
</tbody>
</table>

White House Policy on Increasing Access to the Results of Federally Funded Scientific Research

The Smithsonian Institution, a Trust Instrument of the U.S. people, is voluntarily complying with the 2013 White House Office of Science and Technology Policy (OSTP) memorandum for public access to federally funded research. SRO was a natural fit to work with the Smithsonian Institution Scholarly Press (SISP) to fulfill this Federal mandate. SISP is responsible for negotiating the publication agreement terms for Smithsonian authors and deposit in openly accessible repositories (local or from the publisher). This fulfills the 1st phase of the mandate: public access to publications resulting from federally funded research. A workflow has been established to track the publications license agreements, partnerships with CHORUS [4], and any embargo period stipulated by the publisher. This process has amended the workflow for the SRO adding in a complex layer of tracking of manuscripts, versions of record, and public access versions. Embargo periods before full text can be available (or brightened) now must be coordinated with publishers that are CHORUS members. Although still a work in progress, the SRO team have improved the workflow since the Public Access program was mandated and are now able to integrate the triggers for various steps via a ticket tracking system.

Currently, the SRO web interface does not provide a robust search interface. Users can search the data via the web for basic keyword on any field in the citation. Dspace can be browsed by collection based on the Smithsonian departmental unit.[5] The majority of use of the SRO publications are via harvesting the data either from a SQL query or OAI PHP call to populate
either departmental websites, scholars’ personal web pages, are ingested into larger indexing services such as ProQuest’s Summon knowledge base. SIL staff were pleased to see an increase of successful discovery of SRO materials via search engines after we implemented a test of schema.org. Each citation is also wrapped in RDFa with schema.org microcitation format. This has led to search engines identifying the materials and returning the appropriate citations in search results.

**Altmetrics for SRO Data**


Along with the traditional citation metrics, Smithsonian Libraries experimenting with Altmetrics. [6] DOIs and ISBNs are submitted to Altmetrics for publications to track and score. We provide these numbers in our interface. Smithsonian staff can sign up for an account and track more in depth information and comparisons between Smithsonian and other research institutes.

**Next Steps and Goals**

The goals for the further development of the SRO will focus on the inclusion of Smithsonian humanities research and related metrics. While the SRO currently does hold some monographs, a few articles, and some peer-reviewed online posts, there is currently debate on what are the most effective criteria for inclusion in the SRO from nontraditional publishing and scholarly platforms.

**Smithsonian Profiles**

This past year Smithsonian Libraries staff were focused on a beta experiment with an expert location services. Smithsonian Research Online is one of the driving data sources in the currently beta testing of the VIVO-based Smithsonian Profiles. The open source software VIVO is an RDF system based on a triple store and data graphs.[7] SRO is providing the initial group of people to profile by ingesting current authors and citations. This immediately provides information on the key topics of research and some of the more visible researchers. Still very much a test project, the profiles of the authors along with other data sources from the Institution (principle investigators of grants, advisors to fellows and post-doctorates), has
proven that the tool could be a very useful way to identify the experts in the disperse Institution.

Conclusion

Smithsonian Libraries stepped easily into the role of the bibliographer of the scholarly output of the Smithsonian Institution. Librarians are trained in the basics of metadata needs for citations and curation of scholarly materials. The program of the Smithsonian Research Online has grown and continues to build to incorporate new ways to support what was initially a simple bibliography to a more robust way to share, highlight, and expose research to an international audience. The depth and availability of research at the Institution is the integral part of the original mandate from 1846: “the increase and diffusion of knowledge.”

[1] List museums units and research programs. Give geographic location and link to Global.Si.Edu
[3] ORCID is a nonprofit organization that assigns unique identifiers to researchers: http://orcid.org/
[4] CHORUS – Chorusaccess.org monitors publishers delivering public access scholarly content
[5] https://Respository.si.edu
[6] Altmetrics.com – tracks a range of online activities around attention to publications including news outlets, twitter, blogs, news, and policy documents.

November 8th, 2016