Evidence Based Library and Information Practice

Evidence Summary

Academic Libraries can Expand Institutional Repository Holdings with Gold Open Access Publications Collected Through Web Scraping

A Review of:
Clark, B. (2023). Proactive institutional repository collection development techniques: Archiving gold open access articles and metadata retrieved with web scraping. *Journal of Library Administration, 63(6),* 743–765. https://doi.org/10.1080/01930826.2023.2240190

Reviewed by:
Kristy Hancock
Evidence Synthesis Coordinator
Maritime SPOR SUPPORT Unit
Halifax, Nova Scotia, Canada
Email: Kristy.Hancock@nshealth.ca

Received: 21 Sept. 2024 **Accepted:** 29 Oct. 2024

DOI: 10.18438/eblip30614

Abstract

Objective – To describe a method for collecting gold open access publications from the web and packaging them for batch deposit in an institutional repository. The goal of this project is to expand institutional repository holdings and increase the comprehensiveness of the collection with gold open access content.

Design – Web scraping and analysis of institutional repository usage metrics.

Setting – A library at a public doctoral university with very high research activity in Alabama, United States.

Subjects – Articles and metadata from the Multidisciplinary Digital Publishing Institute (MDPI) website and the Sponsoring Consortium for Open Access Publishing in Participle Physics (SCOAP3) repository. MDPI is an open access publisher of over 400 journals spanning all disciplines. All articles published in MDPI journals are made freely and immediately accessible on the MDPI website.

SCOAP3 is a global partnership of libraries, funding agencies, and research centers that support open access publishing in the field of high-energy physics. The SCOAP3 repository contains research funded by the organization and published in open access journals.

Methods – The MDPI website and SCOAP3 repository were selected because they contained a substantial amount of scholarship by University of Alabama affiliates. On the MDPI website, an author affiliation search across all journals retrieved University of Alabama publications. The Python library Beautiful Soup was used with the parser package lxml to collect articles and metadata. The first script iterated through the pages of search results, downloaded article PDFs, and wrote abstract page URLs to a text file. The second script collected metadata by iterating through the text file of abstract page URLs, parsing the HTML of each URL, and writing Dublin Core metadata to a CSV file. Articles already archived in the institutional repository were removed from the CSV file, and the remaining metadata were reviewed for errors. To pair each PDF with the correct metadata, the file names of all PDFs were added to the CSV file. Article PDFs and the metadata file were packaged using the DSpace CSV Archive and batch deposited in the University of Alabama's institutional repository.

In SCOAP3, an author affiliation search retrieved University of Alabama publications. The browser automation software Selenium was used to collect articles and metadata. The first script iterated through the pages of search results and wrote article record page URLs to a text file. The second script downloaded article PDFs and extracted DOIs to use for PDF file names. The third script collected metadata by using the article record page URLs to query the SCOAP3 metadata harvesting API and writing MARCXML metadata to a CSV file. To pair each PDF with the correct metadata, the DOI column in the CSV file was duplicated, and the ".pdf" extension added to each DOI. The metadata in the CSV file was reviewed for errors, and citations and keywords were added manually. Articles and the metadata file were packaged and deposited using the MDPI method.

The impact of SCOAP3 content on institutional repository downloads from the physics and astronomy collection was measured in the 100 days preceding and following the deposits.

Main Results – 1,005 articles with corresponding metadata were collected from the MDPI website and SCOAP3 repository. After removing duplicate articles that were already archived in the University of Alabama institutional repository, 937 articles (272 from MDPI, 665 from SCOAP3) were deposited. The amount of faculty research available in the institutional repository increased from 1,639 articles before the project to 2,513 articles, or 37.3%.

678 articles were added to the physics and astronomy collection, which reflects the fact that most of the deposited articles were from a subject repository. The rest of the deposited articles were from MDPI and spanned various disciplines. The next best represented collections were civil, construction, and environmental engineering (26 articles); biological sciences (26 articles); electrical and computer engineering (24 articles); and geography (22 articles). The SCOAP3 articles also contributed to a significant increase in downloads from the physics and astronomy collection. Total downloads increased from 5,765 in the 100 days preceding the deposits to 7,243 in the 100 days following the deposits, with SCOAP3 articles representing 3,421 downloads, or 47.2%.

Conclusion – This project was successful in proactively increasing the amount of scholarship in the institutional repository without faculty or researcher participation. This semi-automated workflow requires considerable technical skills but is manageable for one person. Since the articles and metadata were freely accessible and issued under permissive Creative Commons licenses, there was no need to consult publisher self-archiving policies or solicit permission to copy the articles to the institutional repository. This project did not make any research openly accessible that was otherwise unavailable or behind a paywall, but the added publications contribute to making the institution's scholarly record more complete.

This approach may be particularly helpful for academic library staff looking to build the holdings of a brand-new institutional repository, or for those dealing with an underpopulated institutional repository due to low self-archiving rates. Additional repositories containing a substantial amount of University of Alabama scholarship will be identified and considered for web scraping, to continue expanding the institutional repository holdings. The MDPI website and SCOAP3 repository will also be re-scraped in the future for research added since this project.

Commentary

This study contributes to the literature on content recruitment strategies for institutional repositories. As emphasized by the author, self-archiving is critical to institutional repository longevity, and low self-archiving rates can lead to an underpopulated and underused repository. To address the issue of stagnant content growth, institutional repository staff have developed workflows for harvesting content from other sources and using it to populate their own repositories. For example, Lappalainen and Narayanan (2023) describe a semi-automated process for harvesting publication metadata and full text files, where available and appropriately licensed, from Scopus, Web of Science, Dimensions, and Unpaywall. Harvesting content is one of several strategies that staff employ to ensure that new institutional repository content is added regularly. The author's contribution is a process for populating an institutional repository with gold open access publications from MDPI and SCOAP3.

The study was assessed using a critical appraisal tool developed by Perryman and Rathbun-Grubb (2014). The reason for the study is clear, and the literature review is extensive. In the literature review, the author covers topics such as the emergence of open access publishing and institutional repositories, the evolution of open access mandates in the United States, common reasons for low faculty and researcher self-archiving participation, and the use of mediated deposit models to ensure content growth. On the topic of web scraping, they outline methods, general etiquette, and legal considerations. Several existing content harvesting workflows are also highlighted.

In terms of data collection, the author clearly describes the unit of analysis and the reason for choosing this type of data. They provide an in-depth description of the web scraping methods for the MDPI website and the SCOAP3 repository. The critical appraisal tool also includes a question about whether the right kind of information was examined to address the problem. The author measured the impact of SCOAP3 articles on downloads but didn't measure the same outcome for the MDPI articles. It may simply have been too challenging to collect download metrics for the MDPI articles, given that they were added to a range of collections, but analyzing downloads across disciplines would have strengthened the study.

According to the author, the deposited SCOAP3 articles had a significant impact on downloads from the physics and astronomy collection, but readers may wonder if there had been other factors impacting downloads. For example, some institutional repositories offer a personalized email alert feature, which allows users to keep up with new content. It would have been interesting to know more about the 100 days following the deposits, and whether users were alerted to the new content or discovering the new publications while browsing. This aspect of the study could have been reported more clearly.

Overall, this study is novel and interesting. The literature review is excellent, and the author acknowledges study weaknesses, identifies plans for continued content recruitment at their institution, and discusses the wider implications of their findings. The project did not address the issue of low faculty and researcher self-archiving participation, but the author achieved their goal of increasing the amount of scholarship in the University of Alabama institutional repository without having to plan and sustain outreach efforts or mediated deposit services. More research is needed to determine the return on investment for this content recruitment strategy, and to help institutional repository staff decide whether this strategy will align with their institution's goals. Researchers can build on this

study by further exploring the relationship between content volume and usage metrics across multiple disciplines in institutional repositories.

References

Clark, B. (2023). Proactive institutional repository collection development techniques: Archiving gold open access articles and metadata retrieved with web scraping. *Journal of Library Administration, 63(6)*, 743–765. https://doi.org/10.1080/01930826.2023.2240190

Lappalainen, Y., & Narayanan, N. (2023). Harvesting publication data to the institutional repository from Scopus, Web of Science, Dimensions and Unpaywall using a custom R script. *The Journal of Academic Librarianship, 49(1),* Article 102653. https://doi.org/10.1016/j.acalib.2022.102653

Perryman, C., & Rathbun-Grubb, S. (2014). The CAT: A generic critical appraisal tool. http://www.jotform.us/cp1757/TheCat