How the Digital Commons Harvesting Tool with Scopus integration is transforming workflows at Western Libraries

The automated data collection from Scopus not only saves the team time, it also promises to simplify compliance with the university's proposed open access policy.
“With the Digital Commons Harvesting Tool I save so much time. Much of the process — including deduplication — is automated, data wrangling is limited, and errors are reduced. It’s fantastic!”

— Joanne Paterson, Head of the Research and Scholarly Communication Team, Western Libraries, University of Western Ontario, Canada

As Head of Research & Scholarly Communication at Western Libraries, Joanne Paterson manages a dedicated but busy team who support researchers at every stage of the research cycle. For activities such as research dissemination and impact, the institutional repository (IR) at Canada’s University of Western Ontario is fast becoming a go-to resource. Powered by Digital Commons and launched in 2009, Scholarship@Western is now home to more than 34,000 papers that have been downloaded over 11 million times.* According to Joanne, figures like these have sparked growing interest from the university’s research institutes, who see the IR’s potential to help them understand their outputs and impact.

At the same time, the university is poised to adopt a new campus-wide open access policy that will encourage authors to deposit the author-accepted version of their manuscript on an open digital platform. Joanne and her colleagues are confident that Scholarship@Western will become the university’s open repository of choice, with the library helping researchers to comply with the new policy.

However, she admits: “These two developments are likely to greatly impact the workload of our already stretched scholarly communication team. Although there are primarily three of us, the IR is only one of our tasks. In fact, the time we can devote to it probably equates to only one FTE.”

For Joanne, the Digital Commons Harvesting Tool with Scopus integration has the potential to ease the burden for her and her colleagues. Western Libraries was an early adopter of the tool.

In this case study, she walks us through life in the library prior to using the Harvesting Tool and the benefits the team is now reaping, from smoother and faster workflows to error- and duplicate-free content.

* Figures accurate as of March 2022.
The challenge: Tackling a “cumbersome and time-consuming” process

Until recently, the library largely relied on individuals to populate the IR. But, as Joanne notes ruefully: “That didn’t generate a ton of traffic for us. Despite our outreach and education efforts to explain the benefits of open access and the repository to showcase your work, the uptake wasn’t overwhelming.” She adds: “I do get it — faculty members are busy folks. This was just another thing on their list, and it’s a system they are unfamiliar with.”

But, increasingly, the university’s research institutes are realizing the IR’s potential to help them solve an ongoing problem — the difficulty in identifying institute-generated research articles, showcasing those achievements, and highlighting stories of impact.

One of the first to seek the help of Joanne and her colleagues was the Bone and Joint Institute (BJI). Joanne says: “It is a transdisciplinary group of 145 researchers that hold positions in other faculties such as Engineering, Health and Business. In total, six different faculties/schools are involved.

“The problem the BJI faces is that not all the research the faculty produce is for the institute — some is for their own departments. In addition, articles are often attributed to our overarching organization, University of Western Ontario.”

In the fall of 2019, Joanne began working with the BJI to identify five-years’ worth of their articles and load them to Scholarship@Western. She recalls: “In addition to the problems the BJI had encountered, we struggled with the fact that we didn’t have in-depth knowledge of the subject matter and couldn’t always tell affiliation just by looking at a paper. And because the work is interdisciplinary, there was a danger it had already been uploaded to the IR by a co-author or the researcher’s department, which meant we had duplicates to deal with.”

Together, the BJI and the library developed a workflow that, for Joanne and her team, proved “cumbersome and really time-consuming.”

Step 1: Data clean-up

The BJI shared files with the library to help with the article identification process. Joanne says: “We received PDF or Word versions of the researchers’ CVs — those aren’t really re-usable formats for us. We were also given some CSV files of metadata and citations exported from various databases. I had to use OpenRefine to clean up the data; for example, separate out author names into first name/last name columns and merge, or strip out unnecessary data. We also had to individually remove all the duplicates — not an easy task — and add missing details, such as DOIs, full journal titles and complete publication dates.”

On top of these challenges, the team also had to identify and single out the articles generated by the institute.

Step 2: Upload to the IR

Once Joanne had completed her changes in OpenRefine, she exported that data into Excel, then cut and pasted the remaining information into a bepress (Digital Commons) batch upload spreadsheet. She admits: “Even using batch loading, I still had a lot of data entry spreadsheets on my hands.”

Step 3: Add article information

Joanne and her colleagues used Zotero to check whether an open access copy of each article was available. If it was, they pulled it into Dropbox, generated a Dropbox URL, then from that created an OpenURL, and loaded that to the IR. If the article wasn’t open access, they approached researchers to request a manuscript version.

“Prior to the Digital Commons Harvesting Tool, it took around 15 minutes to load an article to the IR. Even if you work all day without a break, the maximum you can manage in an eight-hour period is 32 articles. And that’s probably an overestimate!”

— Joanne Paterson

The solution: Saving time and streamlining workflows with the Harvesting Tool’s Scopus integration

Since Western Libraries joined the Digital Commons Harvesting Tool’s Scopus integration pilot project in June 2020, workflows for populating the IR have changed dramatically.

Joanne says: “You can see from the table below the difference it has made. The 2021 figure is lower than in 2020 because we had cleared the five-year backlog by then.”

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of BJI papers uploaded to the IR by Western Libraries</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>17</td>
</tr>
<tr>
<td>2020</td>
<td>1,086</td>
</tr>
<tr>
<td>2021</td>
<td>250</td>
</tr>
</tbody>
</table>
She adds: “In October 2020 alone, we uploaded 107 articles for the BJI, and that was without working on it full time. At our original pace of 15 minutes per article, that number of articles would have taken us 201 hours — if you work 40 hours per week, there’s only 160 working hours in a month!”

Joanne recently began helping the university’s Brain and Mind Institute (BMI) capture their publications in the IR. And the workflow they’ve developed with the help of the Digital Commons Harvesting Tool’s Scopus integration looks very different.

Step 1: Data collection
Joanne uses the Digital Commons Harvesting Tool to collect publication data from Scopus.

Step 2: Data clean-up
The tool automatically checks the entire IR — not just that collection — for any duplicates and flags them in the spreadsheet it produces. Joanne removes these then checks her file against the researcher information the BMI has uploaded to a dedicated Zotero shared library. Joanne says: “If I see something in the harvested data that’s not in the Zotero file, then I know it’s probably not a product of the institute and I can strip that out. At some point I can go back and load those papers, but at this point I’m just focusing on this institute.” Joanne also has the option to add email addresses, so that researchers are notified when the upload to the IR takes place and receive monthly readership reports. Finally, she removes any excess columns and makes any necessary tweaks to the metadata.

Step 3: Add open article information
For Joanne, along with the automated deduplication function, another big plus point of the Digital Commons Harvesting Tool is the fact that it flags open access content types in the spreadsheet, along with information on version control and sharing permissions. “That tells me whether I need to spend time looking for an open copy or not — a huge time saver,” says Joanne. If there is an open copy available, she uses the OA Works Open Access Button as a Chrome extension to get the relevant URL. She says: “I can usually then put that URL right in my spreadsheet.” If there isn’t an open URL available online, she creates one using a re-jigged Dropbox link. For non-open access articles, she adds a link to the journal.

Step 4: Upload to the IR
According to Joanne, this process is easy and intuitive. “When I do the upload, if I’ve done something wrong in the spreadsheet, I get an error message and it tells me where that error is, which is really helpful.”

Step 5: Supplement the data
Because the library is keen to increase access to their research and share it openly, Joanne has begun approaching institute administrators to request preprints and post prints. These are also uploaded to the repository.

Working with the Digital Commons Harvesting Tool — Joanne’s 4 key take-aways

1. Work with department administrators to help you figure out which outputs are important to them. They have a closer connection to the researcher and can also assist you in getting any earlier versions of a paper (such as an accepted manuscript) for upload.

2. A Zotero shared library is a really efficient way to collaborate and gather bibliographic metadata.

3. Know the limitations of the database you are harvesting from — none of them are absolutely complete, they index different things, and metadata may be incomplete, etc.

4. The Digital Commons Harvesting Tool is very handy! You not only get a lot of good information and save time, there are also fewer errors because you don’t have to rekey content.

Introducing the Digital Commons Harvesting Tool

The Digital Commons Harvesting Tool can collect data from Scopus, Pure, ORCID, PubMed and Sherpa-Romeo.

Regardless of whether your IR hosts metadata-only records or requires a full-text file/copy with every record, it automates the following steps to create a more efficient workflow:

• Find all your institution’s works
• Identify OA content
• Map and prepopulate high-quality metadata for upload, plus import of full texts from Pure
• Check for duplicate records already in your IR
• Check permissions for multiple journals at a time

Administrators of Digital Commons with appropriate permissions can access the Harvesting Tool via their My Account page. Access to Scopus metadata through the Harvesting Tool is available free to Scopus subscribers. For Digital Commons customers without a Scopus subscription an additional fee is required.

Learn more about Scopus integration on Elsevier.com.
About University of Western Ontario (UWO)

UWO, often shortened to Western University or Western, was established in 1878 in London, Canada, and is now ranked among the top 1% of higher education institutions worldwide. It is a founding member of the UIS Group of Canadian Research Universities — members of which undertake 80 percent of all competitive dollar university research in Canada.

The UWO campus comprises 12 faculties, including Business, Dentistry, Engineering, Law, Medicine and Music, and three affiliated university colleges. UWO also has around 70 centers and research institutes connected to local hospitals; for example, the Bone & Joint Institute and the Brain and Mind Institute.

In total, there are about 40,000 students — a figure which includes more than 5,800 international students from 128 countries — and circa 1,500 full-time faculty. UWO offers more than 400 combinations of undergraduate majors, minors and specializations.

The university has a strong track record in Rhodes Scholars and more than 95% of graduates going on to find employment. It was also at UWO that Sir Frederick Banting rose from a restless sleep in 1920 and wrote out 25 words that led to his discovery of insulin.
About Scopus:
Scopus is a source-neutral abstract and citation database curated by independent subject matter experts who are recognized leaders in their fields. Scopus puts powerful discovery and analytics tools in the hands of researchers, librarians, research managers and funders to promote ideas, people and institutions.

1.8+ billion cited references dating back to 1970
84+ million records
17.6+ million author profiles
94.8+ thousand affiliation profiles
7+ thousand publishers

Interested in learning more about Scopus?
Visit elsevier.com/solutions/scopus

About Digital Commons:
Digital Commons is a state-of-the-art institutional repository solution with integrated professional-grade institution-led publishing, research data management, and faculty profiles tools so you can openly publish, manage and showcase the full spectrum of your institution’s research, scholarship and expertise.

>4.3 million full-text OA articles
>920 million downloads
>2,100 journals
>500 institutions worldwide

Interested in learning more about Digital Commons?
Visit elsevier.com/solutions/digital-commons