NOTES ON OPERATIONS

Growing an Institutional Repository

Leveraging a Citation Database as a Tool for Sourcing Deposits and Conducting Outreach

Savannah Lake and Stephannie Regenauer

Many institutional repositories continue to struggle with low engagement. A combination of factors is often at play, including overburdened faculty, confusion about copyright, and lack of awareness. Adding to these barriers on the researcher side are resource constraints on the administrative side, with many libraries citing limitations in budget and staffing for institutional repositories.[1] Atkins Library at the University of North Carolina at Charlotte sought to address these issues by strategically leveraging citation and copyright information that already existed in Web of Science to grow their institutional repository, Niner Commons. Keeping user needs and staff limitations top of mind, Atkins Library launched a project to reframe the approach to increasing participation with the repository: instead of continuing to expect users to deposit works on their own, the library developed a service in which staff could quickly and sustainably deposit works on behalf of users.

Atkins Library launched University of North Carolina at Charlotte's institutional repository, Niner Commons, in 2019. In its early years, Niner Commons was promoted through various outreach initiatives, including faculty champions who deposited in the repository and promoted it to their peers, a promotional video describing the repository and its benefits, and presentations at various college and department meetings.[2] Despite these efforts, faculty and researcher engagement with Niner Commons had been fairly low, totaling 126 works at the end of 2021, despite a campus of more than 3,000 faculty who, according to Web of Science, had published 4,817 works during that period—1,742 of which are categorized by Web of Science as having some level of open access. This lack of engagement was due to many of the same reasons cited by other repositories. We heard from faculty and liaison librarians that researchers were not submitting to the repository because they were confused about copyright and what they could deposit, they already had busy schedules and were reluctant to take on additional responsibilities, or that they simply did not know about it. Given this feedback, we wanted to create an outreach solution that would introduce faculty to Niner Commons to raise awareness and demonstrate its value while also not adding to already burdened workloads.

Accordingly, in early 2022 we began exploring the possibility of creating a mediated workflow in which we would identify works eligible for the repository and deposit them on behalf of the researcher. In addition to having success at other universities in the literature, we felt this engagement strategy took a user-oriented approach that reduced barriers for faculty and researchers while also serving as an easy way for faculty to learn more about the repository and enjoy some of its benefits without having to do

Savannah Lake (savannah.lake@charlotte.edu; https://orcid.org/0000-0001-9998-5185) is the Digital Scholarship Librarian at the University of North Carolina at Charlotte. **Stephannie Regenauer** (regenauer@ifls.lib.wi.us; https://orcid.org/0009-0001-8497-8422) is a Bibliographic Services Cataloger at the IFLS Library System.

Library Resources & Technical Services | January/April 2024
https://doi.org/10.5860/lrts.68n1.8217

any additional work. Although this emphasis on reducing barriers and burdens for faculty was integral to our thinking, we also recognized that our library has only one librarian in charge of supporting the repository, among other duties. Given these staffing constraints, it was important to develop a mediated workflow process that was sustainable and manageable. As such, in developing the workflow we focused on using easy-to-learn, accessible software platforms as well as batch strategies that required as little manual mediation as possible. We ultimately devised a workflow that used citation data from Web of Science, OpenRefine, Oxygen XML Editor, and a hybrid cloud ingest strategy between Amazon Web Services and Islandora to batch ingest works into our repository. The workflow was developed over a ten-week period and is implemented by a staff of one. By detailing our process and its outcomes, this case study will explore how to develop a mediated deposit workflow while facing staffing and technical constraints.

Literature Review

Open access repositories have been an established part of scholarly communication practices for more than twenty years, with more than 900 repositories in the United States and more than 6,000 repositories worldwide registered with the Directory of Open Access Repositories.[3] Despite this established history of practice and the many benefits that institutional repositories bring—including benefits to faculty in increasing the impact and discovery of works, benefits to institutions in showcasing and preserving their scholarly output, and benefits to readers and other researchers in facilitating information access—low engagement continues to be a challenge.[4] In their guide on institutional repositories, the Scholarly Publishing & Academic Resources Coalition (SPARC) describes the biggest impediment to faculty engagement with repositories as being "the inertia of the traditional publishing paradigm."[5] Indeed, there is much in the literature documenting factors that speak to this inertia and prevent faculty from submitting works to institutional repositories, which range from not knowing about institutional repositories, to not trusting works in repositories, to the perception of inconvenience tied to using a new service or platform.[6]

Libraries have employed various approaches to encourage engagement with institutional repositories, including measures such as targeting different audiences with personalized outreach, focusing on "low-hanging fruit" materials without copyright interference (such as grey literature), and integrating manuscript deposits with already existing workflows.[7] One common approach to increasing faculty engagement involves mediated deposits, in which library staff source scholarship generated at their institution and deposit works on behalf of researchers. In fact, in a survey collecting input from repository administrators at eighty-five institutions, 54 percent of administrators indicated that all repository content was mediated by staff, while the remaining 46 percent of repositories used a combination of mediated deposits and faculty self-archiving—with no respondents indicating repository materials are self-archived by faculty only.[8] For many institutions, actively soliciting and submitting deposits on behalf of faculty is the best way to ensure materials are posted to the repository. A study from Oregon State University, for example, evaluated deposit metrics in connection with various

promotional strategies, finding that direct solicitation of manuscripts and subsequent mediated deposit was the most effective.[9]

This approach, of course, requires more work for library staff than relying on faculty to deposit directly. Soliciting and mediating deposits can take many forms, depending on the resources and technology available at the institution as well as staff bandwidth and expertise. The University of Massachusetts Amherst, for example, was able to harvest citation data from Web of Science by connecting the backend of their repository through Web Services.[10] This built on their previous workflows harvesting through various subject repositories like PubMed, arXiv, and RePEc, requiring "upfront technical work," which they cite as a limitation of setting up Web Services.[11] Robust resources—in the form of technical support and staffing—proves a common element of many of the more advanced workflows for mediating deposits. A case study from Zayed University, for example, describes a workflow leveraging Scopus, Web of Science, Dimensions, and Unpaywall as well as strong technical support staff to develop a custom R script.[12] Kansas State University describes a process in which a cross-departmental team of catalogers and repository staff was able to build out a Wiki complete with publisher policies to help identify and ingest works.[13] The College of Wooster, who created scripts to crosswalk RefWorks data to Dublin Core and verify works' rights situation against Sherpa Romeo, benefitted from strong technical support to develop the script and student assistants to implement the workflow.[14]

Alternatively, a case study from Valparaiso University describes an automated process for generating metadata records using largely free and open-source tools, which include email alerts to collect publications, Zotero and then Excel to format those citations, and finally a script (from the aforementioned College of Wooster case study) to search publications against Sherpa Romeo for embargo policies.[15] As a more resource-constrained library, both in terms of staffing and technical support, we found this case study helpful in devising our own workflow, but it had marked differences that did not allow for complete adoption, including different metadata schemas (we use MODS instead of Dublin Core), additional staff, and an emphasis on generating the metadata records (as opposed to also obtaining the full-text files). More applicable was a case study from Florida State University, which used OpenRefine and Web of Science in their workflow. As we have access to Web of Science and previous experience with OpenRefine, this case study was especially helpful to review.[16] Their workflow did, however, involve student worker support, which we could not plan on.

Although all these varied approaches to identifying, soliciting, and ingesting works are illuminating, many of these workflows do require a generous set of resources, such as multiple database subscriptions, several staff members supporting the process, and strong technical support. For institutions lacking some or all of these elements, the prospect of soliciting and mediating deposits can feel overwhelming. Furthermore, our institutional repository is full-text only, as we do not post publication metadata records without also including a full-text file of the work. Accordingly, we were interested in developing a workflow that would include the back-end, technical work as well as the more forward-facing outreach work with faculty. This article seeks to add to the growing literature on

populating institutional repositories via mediated deposits by exploring how to leverage batch processes to harvest, crosswalk, and ingest records—all under staffing and resource constraints.

Designing the Workflow

Niner Commons runs on Islandora and uses MODS for its item metadata. Since the repository's launch in 2019, researcher self-archiving has been minimal. In its first two years, the repository had only received 126 works from faculty and staff; recognizing this low engagement was the impetus for designing this workflow in early 2022. A complicating factor in designing this workflow was staffing constraints. The repository has always been principally supported by a sole librarian, the digital scholarship librarian, receiving additional support for special projects like ingests of electronic theses and dissertations from the metadata librarian and a software developer. Aside from managing the repository, the digital scholarship librarian is also responsible for other open access services at the library, such as journal and book publishing programs. Accordingly, in designing any workflow for Niner Commons, it was important to factor in these staffing constraints while also grounding the workflow in the needs of its users—the faculty and researchers we hope to engage further.

With these priorities and considerations in mind, we began developing this workflow in earnest during the summer of 2022. In reviewing Web of Science, to which our institution has access, we found that many faculty and researchers at UNC Charlotte already published via some sort of open access. At the time of our research, we identified more than 5,000 open access works by UNC Charlotte authors in Web of Science alone, yet the repository only had 126 works self-archived by faculty and staff. We planned to address this considerable discrepancy by creating a workflow that essentially would serve as a pipeline for identifying open access works from UNC Charlotte authors and creating a pathway for ingestion into Niner Commons.

Although there is only one librarian supporting the repository, our library hosts a fellowship program in which a current or recently graduated master's in library and information science (MLIS) student spends the summer onsite working on a specific project. We were fortunate to receive a fellow for this project, who over a ten-week period was able to test, develop, and document a workflow for creating a mediated deposit pathway for publications from Web of Science to our repository, with the thought that the digital scholarship librarian would implement the workflow once the fellowship ended. Accordingly, the project team for developing this workflow consisted of the digital scholarship librarian and the MLIS fellow. The overall workflow developed ultimately followed four steps:

1. Identify open access works from UNC Charlotte researchers using Web of Science.
2. Reach out to the authors of the identified works.
3. Crosswalk Web of Science metadata to local standards.
4. Batch ingest works into the repository.

Identifying Open Access Works

The first step was to formalize a way of identifying scholarship produced by our faculty and researchers that would be eligible for inclusion in Niner Commons. We used Web of Science for this step given our institutional access to this service. Practically speaking, Web of Science offers many options for searching and filtering results that make it useful for identifying potential works, including filters for date, institutional affiliation, and whether the article is open access.

Because we saw this mediated workflow as an outreach strategy to introduce faculty to Niner Commons, we initially focused on works that would not require soliciting postprints from researchers. Accordingly, this workflow focused on articles that were published gold open access with Creative Commons licenses. By focusing first on the "low-hanging fruit" for outreach, we hoped to more quickly grow the number of items in the repository while also increasing awareness among researchers, which would in turn lead to more engagement should we use the same workflow to support green open access and solicit postprints from researchers.

Web of Science provides an abundance of data but requires significant cleanup to make the data useful for filtering, sorting, and grouping in meaningful ways for outreach. A common issue was variations in author names, such as whether a middle initial was used. Another issue was that the data from Web of Science, which was exported as a .CSV file, often combined multiple data values into a single cell. This occurred with author names, resulting in all the authors of each article appearing in one cell and thus making it impossible to sort and filter works by individual authors—a necessary measure for collocating all works by a single author. We used OpenRefine to address such challenges, largely because both the digital scholarship librarian and other metadata teams within the library had previously used OpenRefine for metadata remediation.[17] OpenRefine is a powerful, open-source data cleaning tool that allows users to clean large datasets by "clustering" similar data and making batch edits and transformations.

Data normalization was undertaken with the next step—outreach—in mind. Ultimately, within the data cleanup we prioritized standardizing the open access designations for works, academic department information, and author names. All steps taken to normalize the data, including General Refine Expression Language (GREL) statements and functions, were thoroughly documented so that we could repeat the same process in the future.

Scholar Outreach

Once we had a list of all the open access works identified in Web of Science as authored by UNC Charlotte scholars, we turned to the issue of outreach. Given the high number of works and our relatively low staffing, it was necessary to determine which user groups to prioritize contacting first. Ultimately, we identified two potential approaches to outreach. The first was to reach out to researchers who would be more inclined to participate, which included researchers from disciplines familiar with repositories, such as physics and biological sciences.[18] We also considered starting with the faculty

champions of Niner Commons. Champions had been recruited as early supporters of the repository and had profiles complete with a research biography, but few had submitted works; of these thirty-seven champions, twenty-one had no works in the repository. The second approach to outreach was to focus on academic departments that had little if any representation in the repository to expand the repository's reach and introduce faculty to the service who had not used it yet. After much thought, we decided to take the first approach to outreach initially, to better populate the repository, and then later take the second approach to outreach to broaden the repository's coverage.

Accordingly, the MLIS fellow sorted and filtered the normalized Web of Science data to identify these groups for initial outreach. Rather than rely on the scholar to submit their work to Niner Commons themselves, the digital scholarship librarian would contact these groups regarding recent publications with a request for permission to submit the work on their behalf. We consulted the literature on best practices for repository outreach, which included personalizing communications as much as possible, such as including the faculty member's citations in the email, as well as avoiding jargon, using language like "online repository of scholarly works" instead of "institutional repository."[19] Accordingly, a template email was sent to authors listing identified works, requesting their permission to submit it to the repository on their behalf, and explaining the repository and its benefits to faculty.

Metadata Crosswalk

Once faculty confirmed we could deposit the works on their behalf, we generated the metadata records so that we could batch ingest the works. Normally, researchers would complete a form to submit their works. This form would supply most of the metadata, which would then be edited by the digital scholarship librarian to meet local standards before being ingested. This workflow instead made use of the rich citation data available from Web of Science to create the necessary metadata records. Developing the crosswalk took a fair amount of work, as it involved testing various OpenRefine functions and GREL statements to modify the Web of Science citation data so that it met local standards for the MODS metadata records used in the repository. Here, especially, the staffing constraints of the repository were top of mind, so in developing the crosswalk we did our best to incorporate batch edits and transformations into the workflow instead of manual interventions.[20]

Although this worked to an extent because OpenRefine allows for many global changes, it did not eliminate the need for manual intervention entirely—reconciliation services for automatically linking Open Researcher and Contributor IDs (ORCIDs), Library of Congress Name Authority File records, and Faceted Application of Subject Terminology (FAST) subject terms, for example, still required manual review.[21] Batch transformations were most successful on text-based changes, such as transforming journal titles to sentence case and formatting Digital Object Identifier (DOI) URLs. However, some text transformations still needed manual intervention; article titles, for example, could be batch edited to sentence case but then required manual review to catch any proper nouns or acronyms.

Batch Ingest

The final step in the workflow involved batch ingesting the works, inclusive of metadata records and full-text files, into the repository. Beginning with the metadata records, using OpenRefine's templating function we exported the crosswalked metadata as a single XML file that included all of the MODS records for the various works.[22] We then cleaned the XML file in Oxygen XML Editor by conducting some simple search-and-replace queries and running XSLTs prepared by our metadata librarian to clean spacing issues and remove null fields. Finally, we ran a third XSLT to split the single XML file into individual metadata records for each work.

For the PDF files, since our initial runs of the workflow have been for gold open access works, obtaining the full-text files was straightforward—we simply downloaded them manually from the journal website. Once downloaded, we renamed the PDF files to our local file-naming conventions, which is a combination of author name, title, and year. This file name identifier was also included in the metadata record, which the final XSLT used to rename the metadata XML record. This enables us to pair the PDF files with their corresponding metadata record.

Once both the metadata records and full-text files were ready, we posted them to our repository through a batch-ingest process involving a hybrid cloud strategy developed by Atkins Library software developers that pushes materials from Amazon S3 storage to Islandora.[23] Sizes of batches depended on how many files we had ready at the time, ranging anywhere from ten to 126 files. Fortunately, this piece of the workflow—inclusive of generating an XML file of MODS records through OpenRefine, cleaning the file in Oxygen XML Editor, and batch ingesting metadata records along with full-text files into the repository—was already created by Special Collections and IT staff for their work in posting digitized collections materials to a different repository. We were able to replicate this process for Niner Commons in consultation with them.

Outcomes

Establishing this workflow has been a valuable mechanism for growing the repository. In its first week of implementation alone, the project increased faculty work in the repository by 10 percent. Overall, in the first year of its implementation, from September 2022 through August 2023, this workflow added 158 faculty works to the repository from fourteen faculty members. Previously, the repository only had 210 works. Furthermore, elements of this workflow—including the crosswalking in OpenRefine, generating MODS records in OxygenXML, and batch ingesting works through the hybrid cloud process—were also used to mediate deposits of grey literature, which added an additional fifty-five works. This means that within its first year of implementation, this workflow can be credited with 77 percent of the repository's growth that year. Like many other repositories, we found that depositing on behalf of faculty and researchers yielded more participation and engagement with the repository than the "if we build it, they will come" self-archiving approach.

Table 1. Repository growth over the years, with the new Web of Science batch ingest workflow implemented in the 2022–2023 academic year

Academic Year (September 1–August 31)	No. of Works Added to the Repository	No. of Works Ingested through the Web of Science Workflow
2018–2019	20	0
2019–2020	41	0
2020–2021	26	0
2021–2022	123	0
2022–2023	276	158

However, even though we strove to make this process as effortless and accessible as possible for faculty, we still found some faculty and researchers nonresponsive to our queries. This was even the case with our first outreach efforts, which focused on the faculty who had been early adopters and champions of the institutional repository. Of the five faculty we initially reached out to who were champions, for example, two never responded with a confirmation to proceed. Ultimately, of the sixteen faculty members we contacted in the first year of implementation, six did not respond. This lack of response could be for a variety of reasons, including people feeling burdened by too many emails or the email address looking unfamiliar owing to staff turnover for the repository since they had participated as faculty champions. Accordingly, we have adjusted our process to be more "opt out" by letting authors know that we are offering this service to deposit on their behalf and that if they would not like to participate, they can let us know. This approach has been more effective, as previously there had never been any researchers not wanting to participate, just unresponsive. We initially decided to solicit author approval because we did not want faculty to feel as if we were overreaching. However, ultimately as the works we have focused on thus far are gold open access works and have clear Creative Commons licensing, authors have already essentially granted permission to distribute through this licensing, so we are able to easily proceed with ingesting works.

As we look ahead to the future of leveraging this workflow to support ingesting postprints via green open access, we will need to determine new strategies to get responses to emails. Because we will be requesting faculty to share their postprint, the opt-out model we have been using with gold open access will not work. Although not yet implemented, some initial conversations and further areas for exploration include collaborating with liaison librarians on communications with faculty so that faculty see a familiar contact and may be more incentivized to engage. Additionally, our team is working on developing an Open Access Author Toolkit with user-oriented language and resources that we hope will resonate with faculty. This includes information tailored by discipline, as researcher engagement with open access and journal postprint policies vary widely by discipline, as well as very clear definitions and examples of postprints.[24] With these potential resources in hand, we will aim to expand our Web of Science mediated deposit workflow to include green open access works in the coming years, likely not until the 2024–2025 academic year, as we currently have more immediate, time-sensitive projects.

A key goal of creating this workflow was sustainable growth, as the repository is supported by a staff of one. In keeping with this goal, we sought to develop a streamlined process that relied on batch strategies to empower the digital scholarship librarian to manage this additional task alongside other responsibilities. The workflow certainly has been able to accomplish this to an extent and has resulted in additional deposits to the repository that would not have happened otherwise. Even with all the care we took to make the workflow as expedient as possible, however, it was still difficult to integrate this workflow into the digital scholarship librarian's already full workload. Although the Web of Science citation data certainly reduces much of the work of metadata creation that comes with ingesting works one-by-one via the submission form, not all steps of this workflow are automated, so it still requires a fair amount of manual intervention and time to execute. A batch of sixty works, for example, took approximately two working days to process and ingest. It became clear that a monthly calendar block dedicated to supporting this workflow was the only way to make headway on this project. Fortunately, the digital scholarship librarian will be overseeing a student assistant in the coming academic year, with whose support we hope to make more progress with this workflow. That said, this workflow was developed with our staffing and resource constraints top of mind and demonstrates that even with this care and prioritization of automation, limited staffing can only support so much progress.

It is also important to note that even though we had limited staffing and technical support to build out this workflow, we did have access to Web of Science and Oxygen XML Editor, as well as the help of a full-time MLIS fellow for ten weeks. These resources were integral to creating this workflow and could present a limitation to other institutions looking to do the same. Testing and fine-tuning the data transformations was especially time-consuming, particularly for devising solutions to tricky data issues we encountered, such as sourcing departmental affiliations for researchers because we did not have ready access to aggregated campus data sources. Of the ten-week fellowship, the first and last weeks were spent reading the literature and presenting the work to the library, respectively. It is relatively safe to plan for eight weeks of full-time work to develop and document such a workflow. Institutions with repositories supported by Islandora, hosting MODS records, and having access to Web of Science can use our project documentation to help jumpstart conducting such work.[25] And fortunately, OpenRefine as a software has a fairly low barrier to entry. With ample online tutorials available and a strong user community, staff new to the platform can readily pick it up; both the MLIS fellow and the digital scholarship librarian first used the software within libraries, with the MLIS fellow completing online training during her first week of the fellowship.[26]

Another goal of this workflow was to enhance relationships with faculty by raising familiarity with Niner Commons as a service that could support their research dissemination. The perception around institutional repositories can often be that they are "extra work" and confusing.[27] In reaching out to faculty, we described the repository succinctly in terms of benefits to them. Mediating deposits also showed an awareness of faculty workloads and an effort to impose as little as possible. With these mediated deposits, in tandem with a streamlined self-archiving submission form we introduced that reduced the form from twenty-three fields to six fields, we aim to build trust in establishing the repository as a user-oriented service that is respectful of faculty needs. Additionally, there has been a

fair amount of staff turnover in library positions supporting the repository, so even for researchers who have used Niner Commons previously, this workflow provided an opportunity for them to work with the new digital scholarship librarian and be reacquainted with the service. Many of the faculty champions for Niner Commons, for example, had not been in contact with the new digital scholarship librarian nor deposited articles on their own. Reaching out to them through this workflow facilitated this reintroduction and reminder of the institutional repository as a service that they could leverage to collocate, preserve, and amplify their research. In fact, after reaching out to one faculty member with two works that we sourced from Web of Science, he responded with a list of seven additional open access works for us to deposit.

Looking ahead, we are excited to use this workflow to highlight work by scholars underrepresented in the repository. Having the citation data from Web of Science on hand means that we can easily filter by departments to source works by disciplines underrepresented in the repository; filter by subject keywords to identify works related to diversity, equity, and inclusion topics; and use join tables with the citation data and lists of new and emerging scholars to surface eligible works for the repository. Such measures would expand the reach and coverage of the repository while aligning with elements of the library's strategic plan, which includes goals to showcase the work of diverse scholars.[28] We also are interested in exploring outreach to emeritus faculty, who have decades' worth of scholarship and may be interested in cementing their legacy in an increasingly digital world.[29]

Institutional repositories are important services within the scholarly communications landscape, for research institutions, authors, and readers alike. Hosting a copy of scholarly works locally is important for institutions as it ensures access to their research for the long term, regardless of journal subscription fees. Furthermore, without the green open access that institutional repositories facilitate, inevitably reliance would increase on gold open access, which overlooks the high article processing charges often associated with that mode of publishing—sometimes thousands of dollars. For research that is not grant funded, this can be a significant enough barrier to prevent researchers from publishing open access. Institutional repositories and green open access offer a no-cost pathway to open access publishing, as well as a way to widely disseminate scholarship that is not formally published, such as grey literature. Despite the well-documented problems with faculty engagement and self-depositing in institutional repositories, we will continue to strive for new ways to invest in our repository and increase participation in this worthwhile enterprise.

Conclusion

Encouraging engagement with institutional repositories is something of a perennial problem. Confusion about copyright and a lack of time to deposit works present significant barriers to participation for faculty. These complications are compounded by limited staff support for repositories, as institutional repositories are often supported by minimal library staff who also have additional responsibilities beyond managing the repository. It is imperative to acknowledge these various constraints when devising potential solutions and repository workflows. At Atkins Library, establishing a mediated

deposit workflow that leverages existing citation data, along with batch processes for crosswalking metadata and ingesting works, has helped populate the repository and engage faculty members and researchers.

In thinking through achieving large-scale participation with institutional repositories, ideal yet perhaps exceedingly optimistic solutions would involve widespread administrative buy-in and promotion, technical solutions that seamlessly integrate deposit into existing workflows, additional staffing, and perhaps even changes to promotion and tenure that take into consideration the value of open access publishing. For many universities, however, such solutions simply may not come to fruition, at least not in the near future. Atkins Library has seen success in using existing metadata and batch processes to expedite repository deposits, both demonstrating the value of an institutional repository to faculty and researchers and contributing to a more active culture and practice of open access.

Appendix 1. Metadata crosswalk to transform Web of Science citation data to a MODS metadata record following local standards

Web of Science Field	Niner Commons Field	OpenRefine Data Transformations	Manual Interventions Still Needed
Article Title	Title	• Split the Article Title Column using ": " (colon-space) as the separator • Transform using value[0].toUppercase()+value.toLowercase(). substring(1,value.length()) • Rejoin the column using " : " (space-colon-space) as the separator	Review for proper nouns and acronyms
[n/a]	UNC Charlotte Constituent Type	n/a	Enter manually
Author Full Name	Author: Name	• Split Author Full Names column using split multi-valued cells with "; " as the separator. • Cluster and edit the column. Use key collision/fingerprint then nearest neighbor/ppm. • Re-join the column cells using ";" as the separator. • Split the column into several columns using ";" as the separator. Split into 6 columns at most. Uncheck "Remove this column." • Delete the 6th column.	Will need to manually review items with many authors as these transformations account for only 6 authors.
[n/a]	Author: E-mail Address	n/a	Look up in directory
Author Full Name	Author: Assigned Linked ID By Controlled Vocabulary	• Reconcile Author Full Name columns using LC VIAF API: http://refine.codefork.com/ • Add new column based on Author Full Name column called "LCNAF" using cell.recon.match.id • Transform new column using "http://id.loc.gov/authorities/names/" + cells['LCNAF'].value	In testing, only about 10% of these matched using the auto-match, so there were many to review. It may be easier to look them up manually depending on how many rows are in the file. Consider a numeric facet on the column to focus on the high probability matches.
Reprint Addresses	Author: Department	• Add column based on Reprint Addresses column using the following Python/Jython: • import re • pattern = re.compile(r"((Dept|Sch|School|Department).+?),", re.I) • list = [] • for i in pattern.findall(value): • list.append(i[0]) • return ";".join(list) • Find "Dept" and replace with "Department of" and "Sch" with "School of" • Cluster and edit on column (use nearest neighbor and ppm) and create a text facet to clean normalize department names as able. Reference this document for controlled terms.	Reference directory for any blanks

Author Full Name	Author: ORCID ID	• Reconcile Author Full Name column using ORCID API: http://refine.codefork.com/ • Add new column based on Author Full Name column called "ORCID" using cell.recon.match.id • Transform new column using "https://orcid.org/" + cells['ORCID'].value	This can also be cumbersome to review. Consider a numeric facet on the column to focus on the high-probability matches.
[n/a]	Author: Role	n/a	Set to "author"
Abstract	Abstract	n/a	Review for typos/formatting issues
Publication Year	Single Date of Publication (YYYY)	• Create a text facet on Publication Year column; select blanks • Transform column using cells['Early Access Date'].value.substring(4)	
Author Keywords	Subjects	• Split the Author Keywords column into several columns using "; " (semicolon-space) as the separator. Split into 4 columns at most. Uncheck "Remove this column." • Reconcile first three columns using FAST reconciliation: https://github.com/remerjohnson/conda-reconcile. Delete 4th column. • Add new column based on each reconciled column using cell.recon.match.id	
Document Type	Genre Terms	• Create a text facet and edit document types as needed. They should largely be articles or conference proceedings. Make a copy of the column. • Reconcile new column using Getty API: https://www.getty.edu/research/tools/vocabularies/obtain/openrefine.html • Add new column based on reconciled column called "Getty URI" using cell.recon.match.id • Transform new column using "http://vocab.getty.edu/" + cells['Getty URI'].value	
Language	Language	n/a	Filter list to identify any not in English and check those against MARC Code list for Languages for authorized term
Source Title	Journal/Book/Host Title	• Transform using: value[0].toUppercase()+value.toLowercase().substring(1,value.length())	Review for proper nouns and acronyms
DOI	DOI	• Create duplicate column (DOI value will be used for DOI and URI) • Facet column by blank; select false • Transform using: "doi:" + cells['DOI'].value	
DOI	URI	• Facet column by blank; select false • Transform using: "https://doi.org/" + cells['DOI'].value	

Editor's note: This project was presented as a poster at the ACRL 2023 Conference.

References and Notes

1. Soohyung Joo, Darra Hofman, and Youngseek Kim, "Investigation of Challenges in Academic Institutional Repositories: A Survey of Academic Librarians," *Library Hi Tech* 37, no. 3 (2018): 536–37, https://doi.org/10.1108/LHT-12-2017-0266.
2. "About Niner Commons," accessed July 4, 2023, https://ninercommons.charlotte.edu/niner-commons/about-us.
3. "OpenDOAR Statistics—Sherpa Services," OpenDOAR, accessed June 23, 2023, https://v2.sherpa.ac.uk/view/repository_visualisations/1.html.
4. Shahla Asadi, Rusli Abdullah, Yusmadi Yah, and Shah Nazir, "Understanding Institutional Repository in Higher Learning Institutions: A Systematic Literature Review and Directions for Future Research," *IEEE Access* 7 (2019): 2, https://doi.org/10.1109/ACCESS.2019.2897729; Peter Suber, *Open Access* (Cambridge: MIT Press, 2012): 30, https://doi.org/10.7551/mitpress/9286.001.0001.
5. "SPARC Institutional Repository Checklist & Resource Guide," *The Scholarly Publishing & Academic Resources Coalition* (2002): 11, https://sparcopen.org/wp-content/uploads/2016/01/IR_Guide__Checklist_v1_0.pdf.
6. Rose Fortier and Emily Laws, "Marketing an Established Institutional Repository: Marquette Libraries' Research Stewardship Survey," *Library Hi Tech News* 31, no. 6 (2014): 14, https://doi.org/10.1108/LHTN-05-2014-0038; Zheng Y. (Lan) Yang and Yu Li, "University Faculty Awareness and Attitudes towards Open Access Publishing and the Institutional Repository: A Case Study," *Journal of Librarianship and Scholarly Communication* 3, no. 1 (2015): 2, https://doi.org/10.7710/2162-3309.1210; Ruth Kitchin Tillman, "Where Are We Now? Survey on Rates of Faculty Self-Deposit in Institutional Repositories," *Journal of Librarianship and Scholarly Communication* 5, no. 1 (2017): 13, https://doi.org/10.7710/2162-3309.2203.
7. Marisa L. Ramírez and Michael D. Miller, "Approaches to Marketing an Institutional Repository to Campus," in *The Institutional Repository: Benefits and Challenges*, ed. Pamela Bluh and Cindy Hepfer (Chicago: ALA Editions, 2013), 28; David Scherer, "Incentivizing Them to Come: Strategies, Tools, and Opportunities for Marketing an Institutional Repository," in *Making Institutional Repositories Work*, ed. Burton B. Callicott, David Scherer, and Andrew Wesolek (West Lafayette: Purdue University Press, 2015): 165, https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1040&context=purduepress_ebooks; Sean Aery, "Revitalizing DSpace at Duke," *Bitstreams: The Digital Collections Blog (blog)*, May 25, 2018, https://blogs.library.duke.edu/bitstreams/2018/05/25/revitalizing-dspace-at-duke/.
8. Ellen Dubinsky, "A Current Snapshot of Institutional Repositories: Growth Rate, Disciplinary Content and Faculty Contributions," *Journal of Librarianship and Scholarly Communication* 2, no. 3 (2014): 13, https://doi.org/10.7710/2162-3309.1167.

9. Hui Zhang, Michael Boock, and Andrea A. Wirth, "It Takes More than a Mandate: Factors That Contribute to Increased Rates of Article Deposit to an Institutional Repository," *Journal of Librarianship and Scholarly Communication* 3, no. 1 (2015): 12, https://doi.org/10.7710/2162-3309.1208.

10. Yuan Li, "Harvesting and Repurposing Metadata from Web of Science to an Institutional Repository Using Web Services," *D-Lib Magazine* 22, no. 3/4 (2016), https://doi.org/10.1045/march2016-li.

11. Yuan Li and Marilyn Billings, "Strategies for Developing an Institutional Repository: A Case Study of ScholarWorks@ UMass Amherst," *Journal of Library and Information Science* 37, no. 1 (2011): 89, https://surface.syr.edu/sul/69; Li, "Harvesting and Repurposing Metadata."

12. Yrjo Lappalainen and Nikesh Narayanan, "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, no. 1 (2023): 102653. https://doi.org/10.1016/j.acalib.2022.102653.

13. Debora L. Madsen and Jenny K. Oleen, "Staffing and Workflow of a Maturing Institutional Repository," *Journal of Librarianship and Scholarly Communication* 1, no. 3 (2013), https://doi.org/10.7710/2162-3309.1063.

14. Stephen X. Flynn, Catalina Oyler, and Marsha Miles, "Using XSLT and Google Scripts to Streamline Populating an Institutional Repository," *The Code4Lib Journal* 19 (2013), https://journal.code4lib.org/articles/7825.

15. Jonathan Bull and Teresa Auch Schultz, "Harvesting the Academic Landscape: Streamlining the Ingestion of Professional Scholarship Metadata into the Institutional Repository," *Journal of Librarianship and Scholarly Communication* 6 no. 1 (2018): 8–11, https://doi.org/10.7710/2162-3309.2201.

16. Rachel Smart, "What Is an Institutional Repository to Do? Implementing Open Access Harvesting Workflows," *Publications* 7, no. 2 (2019): 37, https://doi.org/10.3390/publications7020037.

17. "OpenRefine," accessed August 16, 2023, https://openrefine.org.

18. Paul Royster, "How to Fill Your Institutional Repository: Or, Practical Lessons I Learned by Doing," *Library Conference Presentations and Speeches* (2008): 40, https://digitalcommons.unl.edu/library_talks/40.

19. Ramírez and Miller, "Approaches to Marketing," 31; Megan Gaffney, "Involving the Library and Campus Community in Institutional Repository Projects," *The Serials Librarian* 55, no. 4 (2008): 573, https://doi.org/10.1080/03615260802380411.

20. See Appendix 1.

21. "Reconciliation Services for OpenRefine," accessed July 4, 2023, https://refine.codefork.com/; Ryan Johnson, "Remerjohnson/Fast-Reconcile," accessed July 4, 2023, https://github.com/remerjohnson/fast-reconcile.

22. Savannah Lake, “WoS-IR-pathway,” accessed August 22, 2023, https://github.com/savannahlake/WoS-IR-pathway.

23. Brad Spry, “Industrial-Strength Ingest,” presented at Islandoracon 2019, Vancouver, British Columbia, 2019, http://hdl.handle.net/20.500.13093/work:75.

24. Anthony J. Olejniczak and Molly J. Wilson, “Who’s Writing Open Access (OA) Articles? Characteristics of OA Authors at Ph.D.-Granting Institutions in the United States,” *Quantitative Science Studies* 1, no. 4 (2020): 1436, https://doi.org/10.1162/qss_a_00091.

25. Savannah Lake, “WoS-IR-pathway.”

26. “Library Carpentry: OpenRefine,” accessed July 4, 2023, https://librarycarpentry.org/lc-open-refine/.

27. See, for example, Fortier and Laws, “Marketing an Established Institutional Repository,” 14; Yang and Li, “University Faculty Awareness and Attitudes,” 2; Tillman, “Where Are We Now?” 13.

28. “J. Murrey Atkins Library Strategic Plan: 2021-2031,” accessed July 4, 2023, 8, https://drive.google.com/file/d/1-5fb1ifsDXPbqg3cbHUwuz_CF_DwH1Sq/view.

29. Royster, “How to Fill Your Institutional Repository,” 31.