The Journal of Academic Librarianship 49 (2023) 102653

Contents lists available at ScienceDirect

The Journal of Academic Librarianship

journal homepage: www.elsevier.com/locate/jacalib

# Harvesting publication data to the institutional repository from Scopus, Web of Science, Dimensions and Unpaywall using a custom R Script

Yrjo Lappalainen [*], Nikesh Narayanan

*Zayed University, Library and Learning Commons, Dubai, United Arab Emirates*

ARTICLE INFO

*Keywords:*
Institutional repositories
Open access
Content harvesting
Metadata
Workflow automation
Unpaywall
R programming language

ABSTRACT

Institutional repositories are established tools for archiving and increasing the visibility and availability of academic outputs. Although the potential benefits of institutional repositories are well researched and many funders and institutions already mandate open access publishing via gold or green open access routes, institutional repositories often struggle with lack of growth and sustained workflows for content recruitment. Institutions have come up with various (and often creative) workflows for populating their repositories, including institutional open access mandates, library-mediated self-archiving, fully or partially automated content harvesting and integrations between repositories and Current Research Information Systems (CRIS).

Zayed University launched the ZU Scholars[1] institutional repository in fall 2021. Since the beginning, a semi-automated workflow was introduced to populate the repository with publication data from Scopus, Web of Science, Dimensions and Unpaywall using a custom R script. Full text files are added automatically for all Creative Commons licensed articles. This article describes the data harvesting and conversion process, its current limitations and plans for future development. The article also reviews similar content harvesting projects in the context of institutional repositories.

## Introduction

An institutional repository is considered a powerful medium to archive an institution's intellectual output by its faculty, researchers, and students and to make it accessible within and outside the organization. Since open repositories have greater significance in increasing the visibility and accessibility of research, many research institutions have implemented an institutional repository, evident from the increasing number of registrations (4700+) in the Registry of Open Access repositories.[2] Identifying its importance, Zayed University also implemented and launched the ZU Scholars institutional repository in fall 2021 to showcase the university's research in one place and to make works openly available and accessible to a larger audience. ZU's repository software is Digital Commons,[3] originally developed by Bepress and acquired by Elsevier in 2017.

According to de Castro (2014), institutional repositories and Current Research Information Systems (CRIS) are two different kinds of platforms for collecting and disseminating information about research outputs and activities. Repositories generally emphasize the dissemination of research outputs, whereas CRIS systems are more geared toward reporting and comprehensive research analysis. Repositories usually host the full text files, while CRIS systems host a wider set of research information (e.g. researcher profiles, various types of research outputs, activities, research infrastructure, funding information and other master data records such as journals, publishers and events). The key difference between CRISs and IRs lies in the metadata standards: CRIS systems use complex metadata models such as the Common European Research Information Format (CERIF) to describe a wide set of research activities, whereas repositories generally use simpler models that are often perceived “too flat”, offering less flexibility for describing complex semantic areas (de Castro, 2014). However, CRISs and IRs keep evolving toward an increasing level of integration, more complex data models and deeper interoperability (de Castro, 2014).

Although the number of faculty publications at Zayed University has

[*] Corresponding author at: Zayed University, Library and Learning Commons, P.O. Box 19282, Dubai, United Arab Emirates.
*E-mail addresses:* yrjo.lappalainen@zu.ac.ae (Y. Lappalainen), nikesh.narayanan@zu.ac.ae (N. Narayanan).

[1] https://zuscholars.zu.ac.ae

[2] http://roar.eprints.org

[3] https://bepress.com/products/digital-commons

https://doi.org/10.1016/j.acalib.2022.102653
Received 29 September 2022; Received in revised form 9 December 2022; Accepted 11 December 2022
Available online 21 December 2022

risen significantly over the last decade (from 94 annual outputs in 2012 to over 700 in 2021[4]) and the university has an increasing focus on research activities, ZU doesn't have a dedicated CRIS system yet. Therefore, ZU Library decided to implement CRIS-like features in the institutional repository and also use it for tracking and analyzing institutional research. This approach includes harvesting and adding metadata-only records, creating researcher profiles using the SelectedWorks[5] addon for Digital Commons and collecting statistics from the repository using system dashboards and custom scripts. Similar approaches have been successfully implemented earlier by other institutions without a CRIS (see e.g. Bjork, Cummings-Sauls, & Otto, 2019; Bull & Schultz, 2018).

### Statement of the problem

Although the potential benefits of institutional repositories are well researched (Asadi, Abdullah, Yah, & Nazir, 2019) and many funders and institutions already mandate open access publishing via gold or green open access (OA) routes (Mering, 2020), institutional repositories often struggle with lack of growth and sustained workflows for content recruitment (Bull & Schultz, 2018). In general, the proportion of open access articles is growing through gold, green, hybrid and bronze OA channels (Piwowar et al., 2018). However, activating faculty to self-archive their manuscripts into institutional repositories can be challenging. Furthermore, finding existing full text OA versions, especially green OA, is often complex since the files are located in various locations such as institutional repositories, preprint servers and faculty websites. They are also often not indexed consistently (Bulock, 2017). To tackle this problem, various browser extensions have emerged to discover and download OA full text versions from repositories or publisher services. Such tools include Unpaywall, Open Access Button, Lean Library, Lazy Scholar and EndNote Click (Ferguson, 2019; Schultz et al., 2019). Unpaywall also has an Application Programming Interface (API) which makes it possible to make queries into the database programmatically (Dhakal, 2019). This makes it a potential source for automated content harvesting.

### Objectives and scope

To avoid the common issue of stagnation in repository growth and content recruitment, ZU Library decided to introduce a semi-automated workflow for content harvesting early on. This article describes the developed workflow and reviews similar harvesting projects in the context of institutional repositories.

## Literature review

Since first introduced in the early 2000s, institutional repositories have been well-researched over the last two decades and a comprehensive review of IRs is beyond the scope of this article. The potential benefits and trouble spots of IRs have been identified in many earlier studies and review articles. Asadi et al. (2019) reviewed 115 IR-related studies published between 2007 and 2018. They identified several potential benefits and key challenges of institutional repositories. Typical benefits include:

- Showcasing institution's intellectual quality
- Enhancing reputation, visibility and prestige of an organization
- Preserving and disseminating scholarship to a larger audience
- Providing a single consolidated system for academic outputs
- Helping institutions organize their outputs and preserve it long-term
- Enabling to track and analyze research performance
- Supporting learning and teaching

Typical challenges of implementing and maintaining institutional repositories include limited resources, absence of institutional policies, lack of awareness, lack of technical expertise, copyright, ethical and quality concerns (e.g. plagiarism, quality of student works) and a lack of motivation for data sharing (Asadi et al., 2019; Joo, Hofman, & Kim, 2019). Creaser et al. (2010) conducted a survey of authors' awareness and attitudes toward open access repositories. In a survey of over 3000 respondents and four focus groups, they identified three main barriers for depositing articles: concerns over copyright infringement, uncertainty over embargo periods and the unwillingness to place their outputs where other content had not been peer reviewed.

The challenge of activating faculty to self-archive their articles into IRs has been a long standing and widely reported issue (see e.g. Salo, 2008). Solutions to this challenge include implementing institutional open access policies that mandate self-archiving, library-mediated self-archiving and fully or partially automated content harvesting. Zhang, Boock, and Wirth (2015) measured the increase of article deposit rates after two different strategies: soliciting manuscripts directly from authors based on new Web of Science records and implementing an institutional open access policy. They concluded that library outreach and mediated deposit services played an important role and that the direct solicitation was more successful in increasing the rate of article deposits than the institutional open access policy. They also discussed that open access mandates should be linked to research funding and institutional promotion processes in order to make them more effective.

Another way to approach IR content recruitment is to automate processes fully or partially. Automatic content harvesting and workflow automation are relatively common topics in the context of institutional repositories. Research organizations have come up with various (and often creative) ways to populate their repositories and to automate their processes fully or partially. Solutions include the use of external tools and services such as reference management software, Google sheets, Sherpa/Romeo and Openrefine, custom scripting and even fully self-developed systems to facilitate IR workflows.

Li (2016) described a semi-automatic workflow where faculty works were harvested from Web of Science into Digital Commons using WoS web services. According to Li, the workflow greatly improved the efficiency of metadata ingestion compared to a previous manual workflow. However, manual steps were still required for checking copyright policies and acquiring full text files.

Bull and Schultz (2018) reported another semi-automatic workflow where metadata was harvested from Web of Science, Google Scholar and major journal publishers into Digital Commons by setting up email alerts for new works. The metadata was then processed in Zotero and the publisher's copyright policies were fetched from Sherpa/Romeo using a Google Script originally developed by Flynn, Oyler, and Miles (2013). According to Bull and Schultz, the workflow increased the number of faculty works in their IR but still proved to be more time-consuming and problematic than expected, partially due to certain limitations and design choices of Digital Commons. For example, batch uploads can only be done one collection at a time and the system occasionally has issues with handling full text URLs automatically. Although the workflow was partially automated, it still included several manual steps such as formatting and splitting the data manually. It also relied heavily on external for-profit tools which was considered a potential risk for future development.

A similar workflow was described by Smart (2019) who harvested metadata from Web of Science into DSpace, utilizing Zotero and Google Sheets to manage metadata and fetch publisher policies from Sherpa/Romeo. The data was further processed in OpenRefine and converted into Metadata Object Description Schema (MODS) format using a script. Open access articles were also added into the repository and author manuscripts were requested directly from authors by using email templates and a mediated submission process. Smart reported an upward

[4] Based on SciVal data.

[5] https://works.bepress.com

trend of repository content growth since the workflow implementation. However, author manuscript submission rates remained low despite the library's outreach efforts. Only 21.7 % of WoS metadata records were paired with an article and submitted into the repository.

Sergiadis (2019) reported another workflow where Zotero, Sherpa/Romeo and Unpaywall were used to search and add faculty works into Digital Commons. The workflow was mostly manual and required much involvement from the repository manager and student assistants. The process began by requesting Curricula Vitae (CVs) from faculty. Student assistants then searched for publications mentioned in the CV and imported them into Zotero, where copyright policies were added using Zotero's Sherpa/Romeo plugin. Full texts were then located with Unpaywall's browser extension. While Sergiadis focused mainly on the availability and accuracy of metadata in different sources and the differences between various disciplines, the article concluded that these tools can improve the manual IR workflow.

Alsaedi et al. (2021) described a self-developed system that harvests publication data into DSpace. Independent of the institutional repository, it regularly queries various publisher and indexer APIs for new publications, checks for relevant journal policies, identifies institutional authors and makes it easier to request accepted author manuscripts. The solution, named Institutional Research Tracking Service (IRTS),[6] was also made available as open source software. According to the authors, the system helped them to make their repository more comprehensive and the key source of publication data in their institution for other integrations as well, such as ORCID and PlumX (Alsaedi et al., 2021; Baessa et al., 2016).

Zhang (2020) reported another self-developed system designed to automate IR tasks and support self-archiving. The system, named Easy Deposit 2, automatically harvests articles from Web of Science API, parses the metadata and saves it into a local database, requests manuscripts from authors by email and also facilitates the IR submission process by generating a custom deposit link. The system also features an admin dashboard, providing information such as the total number of harvested and self-archived articles. Zhang reported that the system significantly increased the number of article manuscripts deposited by faculty members and that the article deposit rate increased from 7.40 % to 25.60 %. The results showed that a highly automated solution can successfully perform routine IR tasks and significantly increase article deposit rates.

## Methodology

This case study illustrates the development and implementation of a semi-automated workflow for harvesting content into the ZU Scholars institutional repository. The developed process includes metadata harvesting from Scopus, Web of Science, Dimensions and Unpaywall using a custom-made R script. This study provides a detailed description of the harvesting process, discusses its current limitations and presents ideas for further development. The study also describes the repository acquisition process, the rationale behind custom development and the selection of data sources for content harvesting.

### *ZU Scholars institutional repository*

In fall 2020, Zayed University University Library set up a task force to identify the best-suited repository for ZU. The task force benchmarked other universities in the United Arab Emirates to learn about their IR functionalities and best practices. The task force charted a policy framework for the ZU repository with the details of services that are core to the IR, contribution parameters, content creation and harvesting, preservation, management, and content visibility. All ZU Library systems are hosted, so preference was given to hosted IR solutions to avoid the trouble of setting up and hosting a dedicated server environment. The task force analyzed various solutions and decided to proceed with Digital Commons which satisfied most of the requirements specified in the policy framework.

Digital Commons forms a network of more than 600 repositories around the world, which can further increase the visibility of research outputs in individual repositories.[7] Digital Commons also has Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) metadata harvesting compatibility. Additionally, the system is optimized for indexing by all major search engines. Another factor considered was Digital Commons' Scopus metadata harvesting tool which enables the Library to harvest ZU publications into the repository. The expandable scope of Digital Commons, including the SelectedWorks profile service, Journal publication platform and Conference hosting platform was another favorable factor influencing the selection of Digital Commons.

### *Custom development*

Although Digital Commons satisfies most of ZU's requirements, effective content harvesting from multiple sources requires custom development outside the system. The R programming language and RStudio[8] development environment were chosen for the task primarily due to the developer's personal preferences and previous experience, relatively easy setup, widely available documentation, active developer community and many existing program libraries for open science. RStudio requires no server setup and it enables easy data manipulation and views to objects stored in the environment. The environment is free and open source and also well suitable for future development such as statistical analysis and machine learning operations. A minor downside is that RStudio is not a database environment. Setting up a dedicated database would facilitate automation in the long term but it would also require additional effort and resources to set up and maintain. The bright side is that the script can be run in any desktop environment that supports R.

### *Data sources*

Initially, metadata was harvested from Scopus using the built-in harvesting tool in Digital Commons. However, the built-in tool doesn't harvest full text files or some metadata fields such as the article's discipline or the author's unit. This quickly led into the idea of harvesting additional data from other sources and automating the process. Ultimately, the following data sources were chosen:

**Elsevier Scopus**, launched in 2004, is a subscription-based abstract and citation database. Digital Commons, also owned by Elsevier, has a built-in feature to harvest metadata from Scopus, which made Scopus a convenient starting point for the harvesting project.

**Web of Science**, launched in 1997 and previously known as Web of Knowledge, is a subscription-based abstract and citation database, owned by Clarivate. Web of Science was chosen as a source for additional coverage.

**Dimensions,** launched in 2018, is an abstract and citation database owned by Digital Science. In a comparison between various citation databases, Martín-Martín, Thelwall, Orduna-Malea, and Delgado López-Cózar (2020) described Dimensions as a worthy alternative to Scopus and Web of Science in terms of coverage in many subject areas, although it displayed some coverage gaps. Dimensions was chosen as a source due to its easy export functionality, API capabilities and additional coverage.

**Unpaywall**, launched in 2017, is a free database and a tool that harvests open access content from various open indexes (e.g. Crossref and Directory of Open Access Journals), repositories and publisher's services. It is commonly used as a browser extension, but an API is also

[6] https://github.com/kaust-library/irts

[7] https://network.bepress.com

[8] https://www.rstudio.com

available (Dhakal, 2019), making it a potential source for content harvesting. Unpaywall was chosen as a source because of its wide coverage and an existing R library (roadoi[9]) for API access. Unpaywall makes it possible to easily identify and harvest full text files for Creative Commons licensed articles.

*Harvesting process*

The process consists of three parts: 1) collecting the source data, 2) merging and processing the data and 3) batch uploading the data into Digital Commons.

*Part 1: collecting the source data*

The process begins by exporting the latest ZU publication data manually from various sources in Excel format. This is usually done once a week. Also, the current records in the repository are harvested from the repository's API. A list of faculty members is also used as a source for adding additional metadata, e.g. author departments, disciplines and ORCID identifiers.

The following Excel files are collected:

1. Latest ZU articles from Scopus
2. Latest ZU articles from Web of Science
3. Latest ZU articles from Dimensions
4. List of current faculty members

*Part 2: merging and processing the data*

*Preparing the data.* After fetching the Excel files, the data is processed and merged with a custom R script in RStudio. The process begins by loading the Excel files into separate data frames. The script then harmonizes column names and metadata content between different sources and uses regular expressions to split the data into various fields. Initially, suffixes with the name of the source database are added into several column names (e.g. title_scopus, title_wos). This approach enables us to keep the data from all sources and supplement missing data from another source if required.

The script uses Digital Object Identifiers (DOIs) to identify and merge articles. DOIs are first converted into lower case to facilitate matching and deduplication. Duplicate records are identified and removed by comparing newly harvested DOIs to existing DOIs in the repository. Finally, all new DOIs are also compared to each other to make sure there are no duplicates within the latest batch of articles. Additionally, specific DOIs are excluded from the process because they have been identified as false matches in earlier harvests. These DOIs are maintained in a separate list.

After removing all duplicates, the script processes the author data. Digital Commons has separate fields for the first name, last name and institution for the first 30 authors of each article. The built-in Scopus import feature in Digital Commons automatically splits Scopus data into separate fields but importing data from Web of Science and Dimensions requires additional steps. In Dimensions, the author data is in one field, with the institution in parentheses and each author separated by a semicolon as shown in Table 1. In Web of Science, the author data is in two separate fields with the former containing the authors separated by semicolon and the latter containing the names with the institutions.

To convert these fields into the Digital Commons format, regular expressions are used to split the author data. Dimensions has a straightforward structure since the names are separated by a comma and the institution's name is in parentheses. However, Web of Science requires additional lookup since the institution's name is in another field and several authors from the same organization can also be presented within the same bracket. Also, the correct order of authors can only be determined from the first author data field. To connect the author to the right institution, the author's name is looked up from the second author data field with another regular expression (Table 2).

**Table 1**
Author data in Dimensions and Web of Science.

| Source | Author data field 1 | Author data field 2 |
|---|---|---|
| Dimensions | Shin, Donghee (Zayed University); Zaid, Bouziane (University of Sharjah) | |
| Web of Science | Shin, Donghee; Rasul, Azmat; Fotiadis, Anestis | [Shin, Donghee; Rasul, Azmat] Zayed Univ, Abu Dhabi, U Arab Emirates; [Fotiadis, Anestis] Zayed Univ, Coll Business, Abu Dhabi, U Arab Emirates |

*Merging and adding data from Unpaywall.* Once the data has been prepared and the author data has been processed, all contents from Scopus, Web of Science and Dimensions are merged into a new data frame based on DOI. After merging, further data is fetched and merged from Unpaywall with the roadoi R package using DOI as the identifier. This step is essential to the harvesting process since Unpaywall makes it possible to identify Creative Commons licensed full text files and import them into the repository. All fields from all sources are still kept in this phase for the purpose of comparing and supplementing data between the sources.

*Supplementing missing data from other sources.* After merging all data, the script attempts to fill empty fields with data from other sources. The availability of metadata varies between sources (Table 3). For example, Scopus lacks Discipline information and Dimensions doesn't have the International Standard Serial Number (ISSN) or keywords. Unpaywall, on the other hand, is the only source that contains the Creative Commons license information and a direct link to the CC-licensed PDF file. Combining data from multiple sources gives us the option to supplement missing data and to prefer specific sources if they are easier to process or appear to contain higher quality metadata. For example, due to its clear structure, Unpaywall is our preferred choice for OA type and Publisher although this information would be available in other sources as well.

*Converting data and generating identifiers.* Once the data has been supplemented, the script processes several conversions and generates additional data. A unique identifier is generated for each record by joining the title, source of publication (name of journal, book or series) and year of publication. This string is then converted into lower case with white spaces and special characters removed to avoid possible duplicates from alternative special characters or type cases. Finally, the string is converted into a 10-digit hash digest. This identifier is highly useful since it can be used to identify possible duplicate records and to perform quick searches within the repository.

Another check string is created for each author by joining the author's last and first names. To avoid duplicates, this string is also converted into lower case with spaces and special characters removed. The identifier is then used to look up the author's department and discipline from the list of faculty members. An obvious downside of using names as check strings is that namesakes would result in false matches. Since we

**Table 2**
Splitting the author data with regular expressions.

| Source | Regular expression 1 | Regular expression 2 |
|---|---|---|
| Dimensions | Last name, First name (Institution) | |
| Web of Science | Last name, First name | [Last name, First name …] Institution, Institution's location |

[9] https://github.com/ropensci/roadoi

**Table 3**
Availability of selected metadata fields in different sources.

| Source | DOI | ISSN | Keywords | Discipline | OA type | Publisher | CC License | PDF link |
|---|---|---|---|---|---|---|---|---|
| Scopus | X | X | X | | X | X | | |
| WoS | X | X | X | X | X | X | | |
| Dimensions | X | | | X | X | X | | |
| Unpaywall | X | X | | | X | X | X | X |

are a medium-sized university with (currently) no namesakes, this hasn't been an issue so far. Using other identifiers (e.g. ORCID, Scopus ID) would be more reliable, but they are not always available in source databases.

Discipline is an important piece of information as it is used as a classification in Digital Commons. Scopus and Unpaywall don't have any discipline information so the data originates from Web of Science and Dimensions. The built-in classification in Digital Commons is different from the classifications used in Web of Science and Dimensions, so the script harmonizes the data by converting commonly occurring disciplines into the correct Digital Commons classification. If no discipline information is found in the source databases, the discipline is looked up from the list of current faculty members. If the author is not present in the list, the script automatically looks up the author's previous publications from the current IR records and selects the discipline with the most occurrences.

Additional links are also generated for DOI, ISSN and the Scopus identifier. These appear as hyperlinks in the repository, allowing users to access additional information sources conveniently. Finally, the Web of Science publication date requires additional processing since the month is in written format and the date is split into two separate fields (date and year). The month is converted into a number and the two fields are merged to form the complete date. Examples of data conversions are displayed in Table 4.

### *Part 3: batch uploading the data*

After all conversions are complete, the script produces an Excel file which can be batch uploaded into the repository. Before uploading, the data is checked by the repository manager. If some information is still missing (e.g. discipline or department), it will be added manually. The batch upload is also done manually since the Digital Commons API is read-only and it doesn't allow contents to be added into the repository programmatically. To facilitate the upload process, the collection consists of one master collection which includes all faculty works. These works are then displayed in several sub-collections according to certain attributes (e.g. department name, open access status and Scopus indexing status).

### *Summary*

To summarize, the process consists of the following steps:
Part 1: Collecting the source data (manual)
1. Data is downloaded from various sources in Excel format.
Part 2: Merging and processing the data (automatic script)
2. Data is loaded into several data frames in RStudio.
3. Data is prepared and harmonized.
4. Duplicates are removed based on DOI.
5. Data is harvested from Unpaywall.
6. Data is merged into one data frame based on DOI.
7. Missing data is supplemented.
8. Additional data is generated.
9. Data is converted into Digital Commons format.
Part 3: Batch uploading the data (manual)
10. Data is batch uploaded into Digital Commons.

## Assessment of the method

Adopting a semi-automatic harvesting process early on has proven very successful and effective. The harvesting process has enabled the Library to populate the repository with a significant number of works in a short period of time without any author involvement. Currently there are over 5300 records in the repository. Additionally, over 1000 CC-licensed full text files have been harvested using Unpaywall. This has already resulted in over 38,000 downloads.

Although the workflow requires manual steps for exporting and importing Excel files, the whole process takes less than one hour to complete. This includes the time that Digital Commons takes to process new records and re-index the collection. The harvesting process is repeated once a week. On average, it yields around 50–100 new records each month, depending on the availability of publications in source databases. The process can be run by the repository administrator alone.

The script automatically processes and merges data from different sources and removes all duplicates. No manual manipulation of source data is required, except occasional additions and corrections before uploading the final batch upload Excel. Digital Commons also fetches full text files automatically if a direct link to the PDF file is provided. However, sometimes this feature fails to work and individual files need to be uploaded manually afterwards which slightly increases the total time required for updating the collection.

By design, it is not possible to batch upload contents into multiple collections at once in Digital Commons. Therefore, we decided to use a single master collection to facilitate batch uploading. If the collection was structured differently (e.g. own master collection for each college), we would have to repeat the same process separately for each collection and the benefits of automation would be limited. This issue was also noted and discussed by Bull and Schultz (2018). The downside of a single collection is that the batch revision file is rather large and it takes a relatively long time to re-index the whole collection. However, our solution still requires much less time and effort than maintaining multiple collections separately.

## Discussion and future development

Currently, the process still includes manual steps for exporting and importing data. To further automate the workflow, we are planning to use Scopus, Web of Science and Dimensions APIs instead of exported Excel files. This would allow us to drop the first manual step from our current process. The script could also be scheduled to run automatically at specific intervals. However, the Digital Commons API is read-only, so manual steps will still be required for batch uploading and re-indexing the collection in Digital Commons.

The next major initiative will be to start collecting accepted author manuscripts from authors. Identifying the common challenge of activating faculty, we are planning to use a semi-automatic process to facilitate self-archiving. Publisher's open access policies will first be collected from Sherpa/Romeo. This can be done either by using a built-in tool in Digital Commons or by harvesting data directly from the Sherpa/Romeo API. The downside of the built-in tool is that it can fetch policies only for Scopus-indexed articles. Collecting these policies will allow us to maintain a journal-specific list of policies and request specific manuscripts from authors whenever self-archiving is permitted by the publisher. Building on previous IR automation projects (especially

**Table 4**
Examples of data conversion.

| Field | Initial data | Final converted data | Description |
|---|---|---|---|
| Document ID | asystematicanalysisofcommunitydetectionincomplexnetworksprocediacomputerscience2022 | 485e5ca062 | 10-digit hash of the title, name of publication and year |
| Author check string | Feras Al-Obeidat | alobeidatferas | Last and first name with special characters and spaces removed |
| Distribution license | cc-by | CC BY 4.0 | Link to license text |
| DOI link | https://doi.org/10.1016/j.procs.2022.03.046 | https://doi.org/10.1016/j.procs.2022.03.046 | Link to publisher's service |
| ISSN link | 2071-1050 | 2071–1050 | Link to Sherpa/Romeo |
| Scopus ID | 85128732942 (Scopus) | 85128732942 | Link to Scopus |
| Discipline | 01 Mathematical Sciences | Mathematics | Harmonized discipline classification |
| Web of Science date | MAR 22 (date) 2021 (year) | 2021-03-22 | Fixed date string |

Zhang, 2020; Alsaedi et al., 2021), we are aiming to automate the process at least partially by generating custom lists and personalized email messages automatically. Having metadata-only records in the repository will also facilitate the self-archiving process since authors will only need to submit the full text file instead of the whole record.

In addition to the current data sources, other sources such as Lens[10] and ORCID[11] will be considered for additional coverage. However, initial testing indicates that some sources have inadequate or non-existent affiliation data which makes it difficult to identify our organization's publications. We are also exploring the possibility of exporting records from the repository into ORCID. Finally, we are planning to collect statistics (e.g. number of works by department, number of open access works etc.) automatically from the repository using another custom-made script.

## Conclusion

This article described a semi-automatic process for harvesting and converting publication data from Scopus, Web of Science, Dimensions and Unpaywall into Digital Commons. The results show that adopting automated workflows can significantly increase the number of records and Open Access content in institutional repositories and thus reduce the common problem of stagnation in IR growth and content recruitment. By adopting a semi-automatic process early on, ZU Library has been able to populate the repository with over 5300 records and 1000 full text files in a short period of time without any author involvement. Including metadata-only records into the repository will also potentially facilitate self-archiving of full text files later on.

Although proven effective, there are certain challenges associated with the described method. Firstly, developing an automated script requires advanced programming expertise which is not always available in academic libraries. The script also needs to be maintained to accommodate any changes in metadata schemas or APIs. At ZU Library, the script is currently developed and maintained by one key person only which is an obvious risk for continuity. Furthermore, ZU Scholars has customer-specific metadata fields and the script would require further customization if used elsewhere. Another challenge is that Scopus and Web of Science are both subscription-based databases and therefore not available in every organization. Dimensions can be accessed for free but a subscription is required for API access. Although R and Unpaywall are both free tools, the benefits of automation would be limited without harvesting and merging metadata from commercial sources. Finally, there's also a risk that the method can actually induce passiveness in some faculty members since they assume that the Library takes care of everything and no further actions are required. This is why the Library needs to continue promoting the repository and finding ways to get the faculty more involved. We're hoping to increase involvement by establishing a library-mediated process for requesting accepted author manuscripts.

This article demonstrates that the R programming language is well suitable for harvesting, processing, harmonizing and merging publication data from various sources. Using R also opens up many possibilities for further development, such as research analysis, statistics and machine learning operations. The combination of reference databases, Unpaywall API and a custom R script is proven to be an effective solution for populating the repository. Although this article focused on R and Digital Commons, the process could be replicated with other programming languages and any repository software that supports bulk import of records or writing data directly through an API. We are also planning to publish the script so that others can use it and build upon it.

## CRediT authorship contribution statement

**Yrjo Lappalainen:** Conceptualization, Methodology, Software, Writing - Original draft preparation, Data curation. **Nikesh Narayanan**: Writing - Review & Editing, Resources.

## Declaration of competing interest

None.

## References

Alsaedi, Y., Grenz, D. M., & Baessa, M. A. (2021, June 7–10). Leveraging open services to enhance institutional research tracking workflows [Conference presentation]. In *16th International Open Repositories Conference (OR2021), Denver, CO, United States*. https://hdl.handle.net/10754/669460.

Asadi, S., Abdullah, R., Yah, Y., & Nazir, S. (2019). Understanding Institutional Repository in Higher Learning Institutions: A systematic literature review and directions for future research. *IEEE Access, 7*, 35242–35263. https://doi.org/10.1109/ACCESS.2019.2897729

Baessa, M. A., Grenz, D. M., & Wang, H. (2016, June 13–16). Towards a comprehensive and up-to-date institutional repository: Development of a publications tracking process [Conference presentation]. In *11th International Open Repositories Conference (OR2016), Dublin, Ireland*. https://hdl.handle.net/10754/615855.

Bjork, K., Cummings-Sauls, R., & Otto, R. (2019). Opening up open access institutional repositories to demonstrate value: Two universities' pilots on including metadata-only records. *Journal of Librarianship and Scholarly Communication, 7*(1), Article eP2220. https://doi.org/10.7710/2162-3309.2220

Bull, J., & Schultz, T. A. (2018). Harvesting the academic landscape: Streamlining the ingestion of professional scholarship metadata into the institutional repository. *Journal of Librarianship and ScholarlyCommunication, 6*(1), Article eP2201. https://doi.org/10.7710/2162-3309.2201

Bulock, C. (2017). Delivering open. *Serials Review, 43*(3–4), 268–270. https://doi.org/10.1080/00987913.2017.1385128

Creaser, C., Fry, J., Greenwood, H., Oppenheim, C., Probets, S., Spezi, V., & White, S. (2010). Authors' awareness and attitudes toward open access repositories. *The New Review of Academic Librarianship, 16*(1), 145–161. https://doi.org/10.1080/13614533.2010.518851

De Castro. (2014, September). *7 things you should know about institutional repositories, CRIS Systems, and their interoperability*. Retrieved September 19, 2022, from. COAR

[10] https://www.lens.org
[11] https://orcid.org

https://www.coar-repositories.org/news-updates/7-things-you-should-know-aboutirs.

Dhakal, K. (2019). Unpaywall. *Journal of the Medical Library Association, 107*(2), 286–288. https://doi.org/10.5195/jmla.2019.650

Ferguson, C. L. (2019). Leaning into browser extensions. *Serials Review, 45*(1–2), 48–53. https://doi.org/10.1080/00987913.2019.1624909

Flynn, S. X., Oyler, C., & Miles, M. (2013). Using XSLT and Google Scripts to streamline populating an institutional repository. *Code4Lib, 19*. https://journal.code4lib.org/articles/7825.

Joo, S., Hofman, D., & Kim, Y. (2019). Investigation of challenges in academic institutional repositories: A survey of academic librarians. *Library Hi Tech, 37*(3), 525–548. https://doi.org/10.1108/LHT-12-2017-0266

Li, Y. (2016). Harvesting and repurposing metadata from web of science to an institutional repository using web services. *D-Lib Magazine, 22*(3). https://doi.org/10.1045/march2016-li

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2020). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics, 126*(1), 871–906. https://doi.org/10.1007/s11192-020-03690-4

Mering, M. (2020). Open access mandates and policies: The basics. *Serials Review, 46*(2), 157–159. https://doi.org/10.1080/00987913.2020.1760707

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of open access articles. *PeerJ, 2018*(2), Article 6:e437. https://doi.org/10.7717/peerj.4375

Salo, D. (2008). Innkeeper at the roach motel. *Library Trends, 57*(2), 98–123. https://doi.org/10.1353/lib.0.0031

Schultz, T. A., Azadbakht, E., Bull, J., Bucy, R., & Floyd, J. (2019). Assessing the effectiveness of open access finding tools. *Information Technology and Libraries, 38*(3), 82–90. https://doi.org/10.6017/ital.v38i3.11009

Sergiadis, A. D. R. (2019). Evaluating zotero, SHERPA/RoMEO, and unpaywall in an institutional repository workflow. *Journal of Electronic Resources Librarianship, 31*(3), 152–176. https://doi.org/10.1080/1941126X.2019.1635396

Smart, R. (2019). What is an institutional repository to do? Implementing open access harvesting workflows. *Publications, 7*(2), Article 37. https://doi.org/10.3390/publications7020037

Zhang, H., Boock, M., & Wirth, A. A. (2015). 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*(1), Article eP1208. https://doi.org/10.7710/2162-3309.1208

Zhang, H. (2020). Toward easy deposit: Lowering the barriers of green open access with data integration and automation. *Publications, 8*(2), Article 28. https://doi.org/10.3390/publications8020028