Workflow Process Document: Identifying affiliated author publications by publisher

Original copy by Marco Chau with edits from Erin Calhoun and Chloe Thierstein. October 6, 2025.

Table of Contents

Purpose:

This workflow exports institutionally affiliated author publications in given date range for specific publishers. The data corpus includes corresponding and non-corresponding author information for open access and closed articles. The resulting corpus will provide insights into the number and types of articles and adjacent documents that affiliated authors publish, which may be used to inform future open access publishing policies. The University of Toronto Libraries Respond to Updated Tri-Agency OA Mandate Working Group used this data corpus to identify publishers for evaluation of author rights and compliance in light of the Tri-Agency Policy on Open Access Publications.

Downloading Data

Web of Science (WOS)

1. Navigate to Web of Science
    a. UofT catalogue link: https://librarysearch.library.utoronto.ca/permalink/01UTORONTO_INST/14bjeso/alma991106138429606196
2. Navigate to the **advanced search** and build the following search:
    a. Affiliation = [enter affiliation], ex. “University of Toronto”
    b. Publication date = [specific date or date range], ex. 2019-01-01-2024-12-31
    c. Publisher = [name of publisher], ex. Lippincott Williams & Wilkins
3. Conduct search. In the results page, navigate to the filter menu bar on the left. Under “Document Types” select the following:
    - Article
    - Review Article
    - Editorial Material
    - Proceeding Paper
    - Book Chapters
    - Early Access
    - Book Review
    - Correction
    - Data Paper
    - Retraction
    - Retracted Publication
    - Book

    **We are looking for document types that might be subject to APC payments.*
4. Select “Export” and choose “Excel” as the format type. Export records in increments of 1000 with Recorded Content set to “Full Record”
    - Example workflow:
        - Extract 1: Records from 1 to 1000
        - Extract 2: Records from 1001 to 2000
        - Extract 3: Records from 2001 to 3000
        - Etc.
    - Repeat until all records are downloaded.
        - Tips:
            - Keep track of exported files by naming files with their export range, ex. Frontiers_wos_0001-1000, frontiers_wos_1001-2000, etc.

- Save exported files in a designated folder for the source ("wos" or “scopus”) and publisher (ex. Frontiers).

Scopus

1. Navigate to Scopus
    a. UofT catalogue link: https://librarysearch.library.utoronto.ca/permalink/01UTORONTO_INST/14bjeso/alma991106482060206196
2. In the search query builder, enter the following
    a. Affiliation = [affiliation], ex. “University of Toronto”
    b. Add date range = [specific date or range date], ex. 2019-2024
    c. Added to Scopus = Anytime
    Note: Unlike Web of Science, Scopus does not allow for searches by publisher. Data will need to be exported for the entire institution and then filtered out later for specific publishers.
3. Conduct **Search**. On the menu bar to the left under “Document type” select the following:
    - Article
    - Review
    - Conference paper,
    - Book chapter,
    - Erratum
    - Book
    - Data paper
4. Export records
    - **Note:** Scopus **does not** extract files in order, so you **cannot** follow the same steps as WoS for extracting. You **must** extract complete records in one pull, or publications will be missing. Scopus has an export limit of 20,000. If your institution publishes/your result list is more than 20,000, you will need to pull data in controlled portions. To do this, you can do the following;
        1. Split the data by publication year
        2. After splitting the data by year, split the data by document type.
        3. Extract the data once the total number of documents is less than or equal to 20,000 so all the publications are captured.
    - Extract to CSV file type
    - Export all information:
        - Citation information
        - Bibliographic information
        - Abstract and keywords
        - Funding details
        - Other information

- Tips:
  - Save exported files in a designated folder for the source ("wos" or "scopus") and publisher (ex. Frontiers)

OpenAlex

The workflow below describes the process undertaken by the provided code to retrieve and compile the OpenAlex data.

1. Initialize the OpenAlex API using a coding software (R, Python)
   - Documentation for the OpenAlex API can be found here: https://docs.openalex.org/how-to-use-the-api/api-overview
2. Query for works where any author is affiliated with UofT.
   - Entity = "works"
   - Locations.source.publisher_lineage = "https://openalex.org/P4310320990"
     - Filters the publisher for Elsevier BV and its branches
   - From_publication_date = "2019-01-1"
   - To_publication_date = "2024-12-31"
   - Authorship.institutions.id = https://openalex.org/I185261750
     - Filters for UofT as an affiliated institution with any author in the work
3. Query for works where corresponding institution is UofT
   - Entity = "works"
   - Locations.source.publisher_lineage = "https://openalex.org/P4310320990"
     - Filters the publisher for Elsevier BV and its branches
   - From_publication_date = "2019-01-1"
   - To_publication_date = "2024-12-31"
   - Corresponding_institution_ids = https://openalex.org/I185261750
     - Filters for UofT as a corresponding institution for the work
4. Add a "Corresponding Author" column to each of the files. For the file that is only UofT affiliated papers, leave it blank. For the file with UofT as a corresponding institution, fill the column with "University of Toronto Author (Name Unknown)"
5. Combine the OpenAlex files

Cleaning the data

The workflow below describes the process undertaken by the provided code to compile and clean the data from each data source.

Web of Science (WoS)

1. Remove duplicate DOIs
   - Convert DOIs to uppercase
   - Remove duplicate

2. Filter to keep only rows where the “Affiliations” column contains text “University of Toronto”
3. Rename reprint address column to “Corresponding Author” column
4. Select to keep only relevant columns
    - Author Full Names
    - Article Title
    - Source Title
    - Document Type
    - Affiliations
    - Publisher
    - DOI
    - Corresponding Author
5. Add new columns and values
    - “Source” column -> populate with “Web of Science”
6. Rename columns
    - “Author Full Names” -> “Authors”
    - “Article Title” -> “Title”
7. Reorder columns
    a. Authors
    b. Corresponding Author
    c. Document Type
    d. Title
    e. Source Title
    f. DOI
    g. Affiliations
    h. Publisher
    i. Source

Scopus

1. Remove duplicate DOIs
    - Convert all DOI to upper case
    - Remove duplicates
2. Filter to keep relevant publishers from “Publisher” column by checking for matching string
    - Elsevier
    - Cell Press
    - Lancet Publishing
3. Filter to keep only rows where “Affiliations” column contains text “University of Toronto”
4. Rename Corresponding Address column to Corresponding Author

5. Keep only relevant columns
    - Authors
    - Title
    - Source title
    - DOI
    - Affiliations
    - Publisher
    - Document Type
    - Source
    - Corresponding Author
6. Rename columns
    - Rename “Source title” to “Source Title”
7. Reorder columns
    a. Authors
    b. Corresponding Author
    c. Document Type
    d. Title
    e. Source Title
    f. DOI
    g. Affiliations
    h. Publisher
    i. Source

OpenAlex

1. Remove duplicate DOIs
    - First, convert all DOIs to uppercase
    - Remove duplicate DOIs and prioritize keeping instance where the corresponding author is a University of Toronto author
2. Filter for relevant document type
    - Article
    - Editorial
    - Review
    - Book-chapter
    - Retraction
    - Erratum
3. Remove weblink text in front of DOI number
    - Remove “https?://doi.org/”
4. Keep only relevant columns
    - Title

- Host_organization
- Doi
- Type
- Corresponding Author

5. Add new columns and values
    - Source column -> populate with “OpenAlex”
    - Authors column -> populate with NA
    - Source Title column -> populate with NA
    - Affiliations column -> populate with NA
6. Rename stock columns
    - “title” -> “Title”
    - “host_organization” -> “Publisher”
    - “doi” -> “DOI”
    - “type” > “Document Type”
7. Reorder the OpenAlex columns
    a. Authors
    b. Corresponding Author
    c. Document Type
    d. Title
    e. Source Title
    f. DOI
    g. Affiliations
    h. Publisher
    i. Source

Combine and clean the dataframe

The workflow below describes the process undertaken by the provided code to compile and clean the data from each data source together.

Merge the dataframes together

- Using the column names as matchers, merge the Web of Science, Scopus, and OpenAlex dataframes together.

2. Remove duplicate publications
    - Establish the priority for keeping 1 duplicate instance (WoS > Scopus > OpenAlex)
    - With the DOIs all in upper case, remove duplicates and keep one instance based on the priority
    - Convert all Article Titles to lowercase
    - Remove duplicate publications based on article title, keep one instance based on priority

3. Keep only relevant document types
    - In the Document Type column, remove publications that contain the string “book” or “editorial” or “early” ignoring case
4. Shorten author lists
    - To prevent formatting issues, if publication has more than 10 authors, remove the 11th author and beyond and replace with “etc.”
5. Standardize document type names
    - In the “Document Type” column, rename:
        i. Any string with “data” to Data Paper
        ii. Any string with “proceeding” to “Proceedings Paper”
        iii. Any string with “retract” to “Retraction”
6. **Dataframe with all UofT affiliated publications is done!**

Create dataframe with only publications with UofT as corresponding institution

1. In the “Corresponding Author” column, filter for any string with “Univ Toronto” or “utoronto” or “University of Toronto”