

# CONCORDIA UNIVERSITY LIBRARY

## OPEN ACCESS HARVESTER

June 2019

The Open Access Harvester starts with an xml list of numbered, bibliographic citations (say, from a CV), which has been parsed. The harvester uses the xml file as input, and identifies any citations published in journals where [Sherpa/Romeo](#) indicates that a "Publisher PDF" may be deposited in an institutional repository.

It is important to note that the program relies on the presence of citation numbers. If these are not already present, citation numbers must be added/created before parsing takes place:

- 1.Appel, R., & Trofimovich, P. (2017). Transitional probability predicts native and non-native use of formulaic sequences. *International Journal of Applied Linguistics*, 27, 24–43.  
doi:https://doi.org/10.1111/ijal.12100 Trofimovich, P.
- 2.Appel, R., Trofimovich, P., Saito, K., Isaacs, T., & Webb, S. (in press). Lexical aspects of comprehensibility and nativeness from the perspective of native-speaking English raters. *ITL - International Journal of Applied Linguistics*. Trofimovich, P.
3. Ayotte-Beaudet, J.-P., Potvin, P., Lapierre, H. G., & Glackin, M. (2017). Teaching and learning science outdoors in schools' immediate surroundings at K-12 levels: A meta-synthesis. *Eurasia Journal of Mathematics Science and Technology Education*, 13(9), 5343-5363.  
doi:10.12973/eurasia.2017.00833a Potvin, P.

Xml file after parsing:

```
<?xml version="1.0" encoding="UTF-8"?>
- <references>
-   <reference>
      <citation_number>1.</citation_number>
      <author>Appel, R., & Trofimovich, P.</author>
      <date>(2017).</date>
      <title>Transitional probability predicts native and non-native use of formulaic sequences.</title>
      <journal>International Journal of Applied Linguistics,</journal>
      <volume>27,</volume>
      <pages>24–43.</pages>
      <doi>doi:https://doi.org/10.1111/ijal.12100</doi>
      <note>Trofimovich, P.</note>
    </reference>
-   <reference>
      <citation_number>2.</citation_number>
      <author>Appel, R., Trofimovich, P., Saito, K., Isaacs, T., & Webb, S. (in</author>
      <title>press). Lexical aspects of comprehensibility and nativeness from the perspective of native-speaking English
      raters.</title>
      <journal>ITL - International Journal of Applied Linguistics.</journal>
      <note>Trofimovich, P.</note>
    </reference>
-   <reference>
      <citation_number>3.</citation_number>
      <author>Ayotte-Beaudet, J.-P., Potvin, P., Lapierre, H. G., & Glackin, M.</author>
      <date>(2017).</date>
      <title>Teaching and learning science outdoors in schools' immediate surroundings at K-12 levels: A meta-
      synthesis.</title>
      <journal>Eurasia Journal of Mathematics Science and Technology Education,</journal>
      <volume>13(9),</volume>
      <pages>5343-5363.</pages>
      <doi>doi:10.12973/eurasia.2017.00833a</doi>
      <note>Potvin, P.</note>
    </reference>
- </reference>
```

We used <http://anystyle.io> to produce the list of citations in xml format, but you can use any parser you like to prepare the list of citations.

Using your parsed citations, the tool finds the DOIs of the citations, checking each journal against Sherpa/Romeo database to determine whether the journal allows a "green" Publisher PDF deposit.

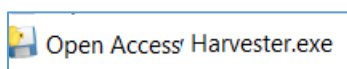
If the journal publisher allows such a deposit, the harvester retrieves the publisher PDF and puts it into a "green" folder. Where Sherpa/Romeo classifies as yellow [OA restricted] but depositing the publisher's PDF is permitted, these pdfs are also harvested and placed in respective "yellow" folder.

When the program is finished, two reports are generated: report.json, and a formatted version of the results report.html, along with the folders containing the pdfs retrieved.

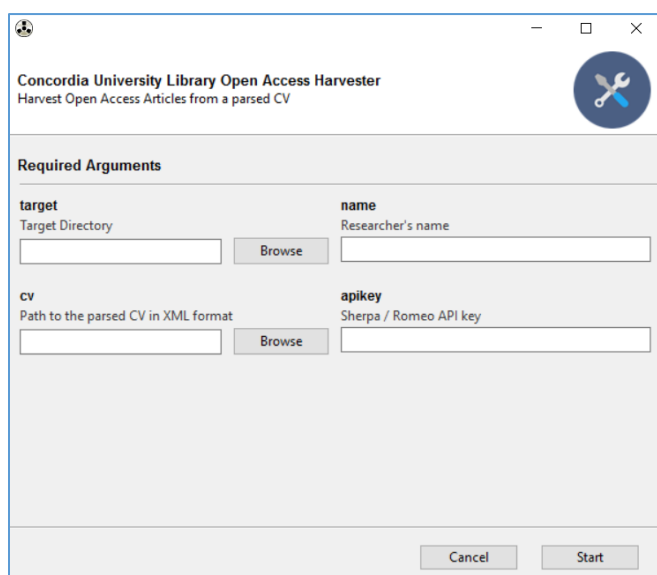
The html report includes a summary of all the citations checked, errors encountered, PDFs successfully fetched, and indicates which PDFs could not be fetched, and articles which may need a manual download instead.

## DOWNLOAD AND RUN

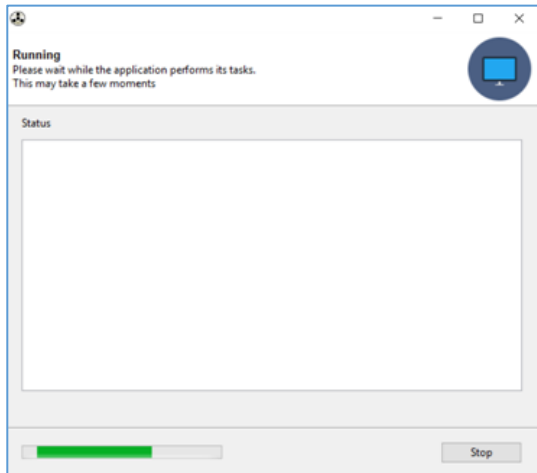
1. Download the **Open Access Harvester** from github <https://github.com/fberrizbeitia/open-access-harvester>
2. Click on the .exe executable file to launch:



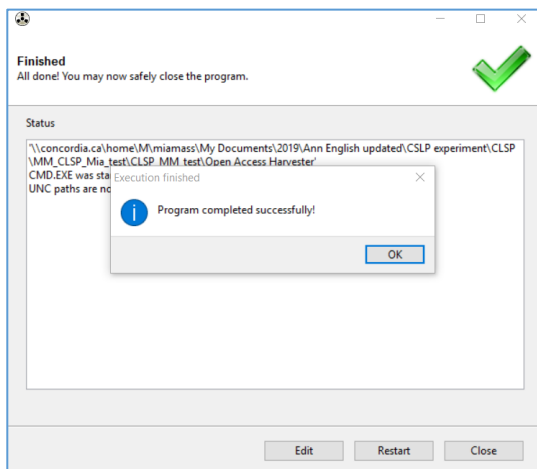
OA Harvester interface opens:



3. Enter your **Target Directory**, using the **Browse** button to navigate to the directory where you want the program's output to be located, e.g. a shared network drive, local PC, etc.
4. Enter **Researcher's name** use consistent naming convention, e.g. *LastName\_FirstName\_MiddleInitial*
5. In the **cv** section, provide the directory path to the parsed CV in XML format.
6. Apply for Sherpa/Romeo api key by visiting <http://sherpa.ac.uk/romeo/apiregistry.php>. After you receive your email from Sherpa, enter **apikey**.
7. Click **Start** to run the program. While the program is running, a green progress bar displays. Depending on the size of the file, it may take a while to run. To cancel, press Stop.

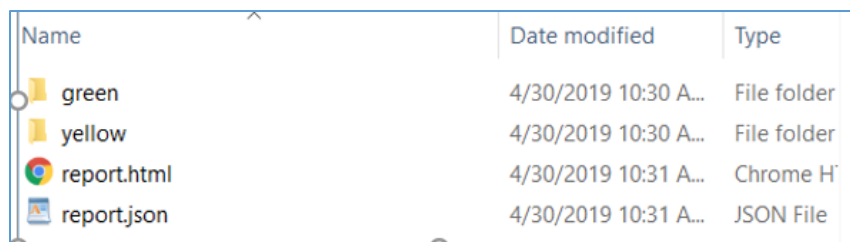


When the program is finished, a dialog box appears. Click Ok and Close to continue.



# VIEWING OUTPUT RESULTS

Navigate to your target folder. Two files are generated: **report.html** and **report.json**. If any PDFs were retrieved, a corresponding 'OA color' folder is generated (e.g., green, yellow, blue).



Name	Date modified	Type
green	4/30/2019 10:30 A...	File folder
yellow	4/30/2019 10:30 A...	File folder
report.html	4/30/2019 10:31 A...	Chrome HTML File
report.json	4/30/2019 10:31 A...	JSON File


Open **report.html**. The report lists every citation processed by OA Harvester, and includes as much detail as can be found for the following elements:

- **Notes** – *possible values*:
  - Trying to fetch the article. PDF archiving is not allowed for this article
  - Unable to retrieve article information from Crossref. PDF archiving is not allowed for this article.
  - Article is possibly open access. Attempting to download full text from the publisher's site.
  - Error retrieving the pdf. Try manual download.
  - Attempting to download full text from the publisher's site. PDF downloaded successfully.
- **Citation number**
- **Author(s)** of article
- **Date article published.**
- **Title of the article**
- **Volume/Issue**
- **Pages**
- **Journal in which article was published**
  - **Short name**
  - **Long journal title**
  - **ISSN**
  - **Sherpa link** if found
  - **Colour** (green, yellow, blue, white) as assigned by Sherpa
  - **PDF Archiving allowed** – possible values:
    - Can
    - Cannot
    - Restricted
    - Unknown
  - Conditions of archiving listed at Sherpa
- **DOI**
- **URL**

All of the data contained in **report.html** is contained in the **report.json** file. To view the results data as a spreadsheet, take the **report.json** and use a conversion tool like [json to csv](#) to change it into a csv or xlsx.

## Folders

Inside the green and/or yellow folders, retrieved PDFs are listed according to the *citation number* in the original parsed file. In the example, the PDF for citation #29 was successfully retrieved and placed inside the green folder.

Name	Date modified	Type
 29..pdf	4/30/2019 10:30 A...	Adobe Ac...

In **report.html**, details of citation #29:

Notes	Trying to fetch the article. Article is possibly open access. Attempting to download full text from the publisher's site. PDF downloaded successfully		
citation_number	29.		
author	Jorgensen, M., Budd, J., Fichten, C. S., Nguyen, M. N., & Havel, A.		
date	(2018).		
title	Graduation prospects of college students with specific learning disorder and students with mental health related disabilities.		
volume	7(1),		
pages	19-31.		
journal	shortName	International Journal of Higher Education,	
	name	International Journal of Higher Education	
	issn	1927-6044	
	sherpa link	<a href="http://www.sherpa.ac.uk/romeo/issn/1927-6044/">http://www.sherpa.ac.uk/romeo/issn/1927-6044/</a>	
	colour	green	
	pdf archiving	can	
	conditions	(1) On open access repositories(2) Author's pre-prints on pre-print servers(3) author's pre-print must be updated with citation, DOI and link to published version(4) Creative Commons Attribution License 3.0(5) Authors retain copyright(6) Publisher's version/PDF may be used(7) Published source must be acknowledged(8) Publisher last contacted on 18/07/2014(9) All titles are open access journals	
DOI	10.5430/ijhe.v7n1p19		
URL	<a href="http://dx.doi.org/10.5430/ijhe.v7n1p19">http://dx.doi.org/10.5430/ijhe.v7n1p19</a>		

If a folder is empty, check the **report.html** for occurrences of "manual download" (use CTRL-F). The URL link in the report can be used to download those pdf(s) manually.

The **Open Access Harvester** has been an experimental collaborative project (2019) of Francisco Berrizbeita, Developer, and Mia Massicotte, Systems Librarian.