Need help searching for RORs? Try RORRetriever!

Ted Habermann

Blog post published June 30, 2022 in *Blog - Metadata Game Changers*
https://doi.org/10.59350/4gxfz-4kb47

Abstract

Cite this blog as Habermann, T. (2022). Need help searching for RORs? Try RORRetriever! Front Matter.

Keywords

Information Systems (Subfield)

Recommended citation

Habermann, T. (2022, June 30). Need help searching for RORs? Try RORRetriever!. *Blog - Metadata Game Changers*. https://doi.org/10.59350/4gxfz-4kb47

Copyright

Copyright © Ted Habermann 2022. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Cite this blog as Habermann, T. (2022). Need help searching for RORs? Try RORRetriever! Front Matter. https://doi.org/10.59350/4gxfz-4kb47

::::: {#block-af62d65303870c061734 .sqs-block .html-block .sqs-block-html block-type="2" border-radii="{"topLeft":{"unit":"px","value":0.0},"topRight":{"unit":"px","value":0.0},"bottomLeft":{"unit":"px","value":0.0},"bottomRight":{"unit":"px","value":0.0}}"}

The Research Organization Registry provides access to a growing collection of unique and persistent identifiers (RORs) for research organizations and a web interface and API for searching the collection with an organization name or a affiliation string. The web interface and API work well if you have a small number of affiliation strings that need identifiers, but what if you have several hundred organization names or thousands of affiliation strings?

In a recent blog I showed that many University repositories face this problem. They have large numbers of affiliations that need identifiers to facilitate connections into the global research infrastructure. The process of finding those identifiers can be difficult and time-consuming. RORRetriever is a tool developed to help these repositories and others that are ready to add identifiers to large collections with many affiliations. It can simplify and streamline the process of finding RORs for many affiliations or organization names by providing an automated interface to the ROR affiliation search.

RORRetriever is open-source and available on github.

Usage

Use python RORRetriever.py -h to see this usage description.

:::::

```
usage: RORRetriever [-h] [-a [AFFILIATIONLIST [AFFILIATIONLIST ...]]] [-af AFFILIATIONFILE
                    [--logto FILE]

search organization names and affiliations for RORs using the ROR Affiliation Strategy

optional arguments:
  -h, --help            show this help message and exit
  -a [AFFILIATIONLIST [AFFILIATIONLIST ...]], --affiliationList [AFFILIATIONLIST [AFFILIAT
                        a list of "affiliations in quotes"
  -af AFFILIATIONFILENAME, --affiliationFilename AFFILIATIONFILENAME
                        a file with one affiliation per line in current working directory
  -ad AFFILIATIONDATA, --affiliationData AFFILIATIONDATA
                        datafile (tsv, csv) with affiliations in cwd)
  --noacronyms          Exclude Acronym matches (default=False)
  --max                 Accept max score if no result chosen by ROR algorithm (more result
  --details             Show detailed response data (automatic for one ROR)
  -o OUTPUTINTERVAL, --outputInterval OUTPUTINTERVAL
                        For batch processing output results update interval default:20
  --loglevel
                        Logging level
  --logto FILE          Log file (will overwrite if exists)
```

::::: {#block-yui_3_17_2_1_1656605115552_166810 .sqs-block .html-block .sqs-block-html block-type="2" border-radii="{"topLeft":{"unit":"px","value":0.0},"topRight":{"unit":"px","value":0.0},"bottomLeft":{"unit":"px","value":0.0},"bottomRight":{"unit":"px","value":0.0}}"}

Environment

ROR Reteriever imports the following python modules: requests, json, pandas, urllib.parse, sys, datetime, argparse, datetime, logging, os, re. The file RORRetriever.yml can be used to create the environment using the command **conda env create -f RORRetriever.yml**

Outputs - ROR Search Results

The results of the search are provided in a tab-delimited file called **AffiliationAPI_RORData__TIMESTAMP.tsv**. This file has at least one row for every input affiliation but can have more than one if more than one ROR is identified for a single affiliation. In many cases when multiple RORs are found, some are incorrect and curation is required to select the correct one.

The columns in this file are:

:::::

Name	Definition
affiliation	The complete affiliation being searched
searchString_Affiliation	Complete affiliations are split into substrings during the search. This is the substring that found the match.
ROR_Affiliation	The ROR identifier found using the substring.
organizationLookupName_Affiliation	The name of the organization with the found ROR.
country_Affiliation	The country the found organization is in.
match_Affiliation	The kind of match that was found (provided by algorithm).
chosen_Affiliation	TRUE if the algorithm chose this organization, FALSE if not.
score	The algorithm score for this organization (0 - 1). The chosen ROR typically has the maximun score.
numberOfResults_Affiliation	The number of results returned from the search.
valid	A column for recording validity during curation (always TRUE prior to curation).

Table 1. ROR output fields.

::::: {#block-yui_3_17_2_1_1656605115552_185566 .sqs-block .html-block .sqs-block-html block-type="2" border-radii="{"topLeft":{"unit":"px","value":0.0},"topRight":{"unit":"px","value":0.0},"bottomLeft":{"unit":"px","value":0.0},"bottomRight":{"unit":"px","value":0.0}}"}

Outputs - Search Details

Some details of the ROR algorithm response can be displayed using the **--details** flag. This is automatic if only one affiliation is being searched. if **--details** is set, a table of the search results is shown with the following fields:

:::::

Name	Definition
substring	Search string (can be substring of complete affiliation)
score	Match score between 0 and 1
matchingType	Method that found the match (provided by algorithm)
chosen	True for chosen ROR False for others
organization	Name of organization for ROR (should match substring)
country	Country of organization

Table 2. Search detail output fields.

::::: {#block-yui_3_17_2_1_1656605115552_190890 .sqs-block .html-block .sqs-block-html block-type="2" border-radii="{"topLeft":{"unit":"px","value":0.0},"topRight":{"unit":"px","value":0.0},"bottomLeft":{"unit":"px","value":0.0},"bottomRight":{"unit":"px","value":0.0}}"}

Examples

As an example of some of the capabilities of this tool consider this list of affiliations:

:::::

```
The University of Alabama Libraries
Arizona State University Library
UC Berkeley Library
MIT Libraries
```

::::: {#block-yui_3_17_2_1_1656612885657_53622 .sqs-block .html-block .sqs-block-html block-type="2" border-radii="{"topLeft":{"unit":"px","value":0.0},"topRight":{"unit":"px","value":0.0},"bottomLeft":{"unit":"px","value":0.0},"bottomRight":{"unit":"px","value":0.0}}"}

These strings are affiliation strings (affiliations for short) because they contain the names of research organizations along with other text. These are simple affiliations as they include university names with the words library or libraries. Despite this simplicity, they demonstrate some interesting characteristics of the ROR Affiliation search. These affiliations are in a test file in this repository (OrganizationNamesTest.txt).

The command **RORRetriever -af OrganizationNamesTest.txt** uses the ROR affiliation API to search for RORs for these affiliations and gives this output to the terminal:

:::::

```
2022-06-27 11:25:00:INFO:RORRetriever: Searching OrganizationNamesTest.txt for affiliation
2022-06-27 11:25:00:INFO:RORRetriever: 4 Input Affiliations
2022-06-27 11:25:03:INFO:RORRetriever: 3 new RORs written to AffiliationAPI_RORData__20220
2022-06-27 11:25:03:INFO:RORRetriever: AffiliationAPI_RORData__20220627_11.tsv 3 RORs Foun
```

::::: {#block-yui_3_17_2_1_1656605115552_196034 .sqs-block .html-block .sqs-block-html block-type="2" border-radii="{"topLeft":{"unit":"px","value":0.0},"topRight":{"unit":"px","value":0.0},"bottomLeft":{"unit":"px","value":0.0},"bottomRight":{"unit":"px","value":0.0}}"}

More details are written to a tab-separated file (converted to a table here):

:::::

affiliation

searchString_Affiliation

ROR_Affiliation

organizationLookupName_Affiliation

country_Affiliation

match_Affiliation

chosen_Affiliation

score

numberOfResults_Affiliation

valid

The University of Alabama Libraries

University of Alabama

https://ror.org/03xrrjk67

University of Alabama

United States

HEURISTICS

True

1.0

9

True

Arizona State University Library

No Match

4

False

UC Berkeley Library

UC Berkeley Library

https://ror.org/01bndk311

Berkeley Public Library

United States

COMMON TERMS

True

0.9

10

True

MIT Libraries

MIT

https://ror.org/047m6hg73

Management Intelligenter Technologien (Germany)

Germany

ACRONYM

True

0.9

18

True

Table 3. Search results from Affiliations__TIMESTAMP.csv.

::::: {#block-yui_3_17_2_1_1656605115552_200974 .sqs-block .html-block .sqs-block-html block-type="2" border-radii="{"topLeft":{"unit":"px","value":0.0},"topRight":{"unit":"px","value":0.0},"bottomLeft":{"unit":"px","value":0.0},"bottomRight":{"unit":"px","value":0.0}}"}

The output data shows that between 4 and 18 results were found for each of these affiliations and that RORs were identified for three. The organizationLookupName_Affiliation column gives the name associated with the ROR. In the first row, this name clearly matches the organization name in the input affiliation (The **University of Alabama** Libraries). In the other cases, the names suggest that either no match was chosen or the incorrect match was chosen.

There are two ways to find out what actually happened. First, the **--details** option can be used to show details for all searches at once, i.e. **RORRetriever -af OrganizationNamesTest.txt --details** or we can run each affiliation one at a time which automatically shows details.

Affiliation: The University of Alabama Libraries

The command **RORRetriever -a "The University of Alabama Libraries"** shows what happened for the affiliation "The University of Alabama Libraries". There were 9 results returned and University of Alabama was selected as the best result (score is 1 and chosen is True). Even though everything went well in this case, it demonstrates the kind of challenges that occur when over 100,000 organization names are being searched. First, many Universities have multiple campuses with very similar names. There are also many cases where states have two Universities with names like "University of Alabama" and "Alabama State University". In this case the algorithm overcame these challenges to find the correct organization. Great job!

:::::

```
2022-06-27 12:40:48:INFO:RORRetriever: 1 Input Affiliations
                          substring  score matching_type  chosen                       ror
              University of Alabama   1.00    HEURISTICS    True https://ror.org/03xrrjk67
              University of Alabama   0.89    HEURISTICS   False https://ror.org/010acrp16
              University of Alabama   0.88    HEURISTICS   False https://ror.org/01s7b5y08
              University of Alabama   0.88    HEURISTICS   False https://ror.org/0584fj407
                 Alabama University   0.86    HEURISTICS   False https://ror.org/01eedy375
              University of Alabama   0.86    HEURISTICS   False https://ror.org/051fvmk98
                 Alabama University   0.82    HEURISTICS   False https://ror.org/05hz8m414
The University of Alabama Libraries   0.76  COMMON TERMS   False https://ror.org/008s83205
              University of Alabama   0.75    HEURISTICS   False https://ror.org/02zsxwr40
2022-06-27 12:40:49:INFO:RORRetriever: 1 new RORs written to AffiliationAPI_RORData__20220
2022-06-27 12:40:49:INFO:RORRetriever: AffiliationAPI_RORData__20220627_12.tsv 1 RORs Foun
```

::::: {#block-yui_3_17_2_1_1656605115552_203053 .sqs-block .html-block .sqs-block-html block-type="2" border-radii="{"topLeft":{"unit":"px","value":0.0},"topRight":{"unit":"px","value":0.0},"bottomLeft":{"unit":"px","value":0.0},"bottomRight":{"unit":"px","value":0.0}}"}

Affiliation: "Arizona State University Library"

The command **RORRetriever -a "Arizona State University Library"** gives the following results:

:::::

```
2022-06-30 11:06:36:INFO:RORRetriever: 1 Input Affiliations
                        substring  score matching_type  chosen                       ror
Arizona State University Library   0.86  COMMON TERMS   False https://ror.org/03efmqc40
Arizona State University Library   0.65  COMMON TERMS   False https://ror.org/03y1zyv86
Arizona State University Library   0.54  COMMON TERMS   False https://ror.org/00g3q3z36
Arizona State University Library   0.54  COMMON TERMS   False https://ror.org/01wwqrh75 Ar
2022-06-30 11:06:37:INFO:numexpr.utils: NumExpr defaulting to 8 threads.
2022-06-30 11:06:37:INFO:RORRetriever: 0 new RORs written to AffiliationAPI_RORData__20220
2022-06-30 11:06:37:INFO:RORRetriever: AffiliationAPI_RORData__20220630_11.tsv 0 RORs Foun
```

::::: {#block-yui_3_17_2_1_1656605115552_206842 .sqs-block .html-block .sqs-block-html block-type="2" border-radii="{"topLeft":{"unit":"px","value":0.0},"topRight":{"unit":"px","value":0.0},"bottomLeft":{"unit":"px","value":0.0},"bottomRight":{"unit":"px","value":0.0}}"}

The table shows the 4 results that were found and that the chosen column is False for all four, even though the right answer has the highest score (0.86). The algorithm considers a number of factors in making a choice and, in this case, the correct answer did not pass the test.

The **--max** option can be used to choose the best answer even if the algorithm does not, so the command **RORRetriever -a "Arizona State University Library" --max** gives the results:

:::::

```
2022-06-30 11:09:19:INFO:RORRetriever: 1 Input Affiliations
2022-06-30 11:09:19:INFO:RORRetriever: ************** Best match is being found (may not b
                        substring  score matching_type  chosen                       ror
Arizona State University Library   0.86  COMMON TERMS   False https://ror.org/03efmqc40
Arizona State University Library   0.65  COMMON TERMS   False https://ror.org/03y1zyv86
Arizona State University Library   0.54  COMMON TERMS   False https://ror.org/00g3q3z36
Arizona State University Library   0.54  COMMON TERMS   False https://ror.org/01wwqrh75 Ar
2022-06-30 11:09:20:INFO:numexpr.utils: NumExpr defaulting to 8 threads.
2022-06-30 11:09:20:INFO:RORRetriever: 1 new RORs written to AffiliationAPI_RORData__20220
2022-06-30 11:09:20:INFO:RORRetriever: AffiliationAPI_RORData__20220630_11.tsv 1 RORs Foun
```

::::: {#block-yui_3_17_2_1_1656605115552_210823 .sqs-block .html-block .sqs-block-html block-type="2" border-radii="{"topLeft":{"unit":"px","value":0.0},"topRight":{"unit":"px","value":0.0},"bottomLeft":{"unit":"px","value":0.0},"bottomRight":{"unit":"px","value":0.0}}"}

and the correct ROR is found. Note that the terminal output warns that this flag is set to make sure the user is aware that, while some results may improve, as in this example, there is also more noise introduced into the results. Also note that the results file contains a column named chosen which will be True for RORs selected by the algorithm and False for those not selected by the algorithm as is the case here.

Affiliation: "UC Berkeley Library"

The details for the affiliation "UC Berkeley Library" can be seen using the command **RORRetriever -a "UC Berkeley Library"** with the results:

::::::

```
2022-06-30 11:11:38:INFO:RORRetriever: 1 Input Affiliations
           substring  score matching_type  chosen                       ror
UC Berkeley Library   0.88         FUZZY   False https://ror.org/01bndk311
UC Berkeley Library   0.77         FUZZY   False https://ror.org/02jbv0t02
UC Berkeley Library   0.57         FUZZY   False https://ror.org/01an7q238
UC Berkeley Library   0.56         FUZZY   False https://ror.org/02xewxa75
UC Berkeley Library   0.51  COMMON TERMS   False https://ror.org/01n7gem85
UC Berkeley Library   0.44  COMMON TERMS   False https://ror.org/03fgher32
UC Berkeley Library   0.43  COMMON TERMS   False https://ror.org/048vdhs48
UC Berkeley Library   0.33  COMMON TERMS   False https://ror.org/05ehe8t08
UC Berkeley Library   0.33  COMMON TERMS   False https://ror.org/05q8kyc69
UC Berkeley Library   0.32  COMMON TERMS   False https://ror.org/04wvygn49
UC Berkeley Library   0.30  COMMON TERMS   False https://ror.org/04gyf1771
UC Berkeley Library   0.27  COMMON TERMS   False https://ror.org/00d9ah105
UC Berkeley Library   0.27  COMMON TERMS   False https://ror.org/05rrcem69
UC Berkeley Library   0.25  COMMON TERMS   False https://ror.org/006dpe828 University of
2022-06-30 11:11:40:INFO:numexpr.utils: NumExpr defaulting to 8 threads.
2022-06-30 11:11:40:INFO:RORRetriever: 0 new RORs written to AffiliationAPI_RORData__20220
2022-06-30 11:11:40:INFO:RORRetriever: AffiliationAPI_RORData__20220630_11.tsv 0 RORs Foun
```

:::::: {#block-yui_3_17_2_1_1656605115552_212624 .sqs-block .html-block .sqs-block-html block-type="2" border-radii="{"topLeft":{"unit":"px","value":0.0},"topRight":{"unit":"px","value":0.0},"bottomLeft":{"unit":"px","value":0.0},"bottomRight":{"unit":"px","value":0.0}}"}

In this case, the algorithm choses the ROR for "Berkeley Public Library" which has the maximum score (0.90) but is incorrect. The correct answer, University of California, Berkeley is in the results with a score of 0.58.

This example is included here to remind us that there is an algorithm involved in these searches and we should help it as much as possible when we are providing affiliations to journals or repositories. In this case the acronym UC is the problem. The algorithm cannot guess that it stands for University of California. Help it out by writing the affiliation as University of California Berkeley Library and the algorithm gets the right answer (**RORRetriever -a "University of California Berkeley Library"**).

Affiliation: "MIT Libraries"

The final example (MIT Libraries) illustrated another problem case, using an acronym in an affiliation rather than spelling it out. In this case, the command **RORRetriever -a "MIT Libraries"** shows the 7 organizations with the acronym MIT:

:::::

```
2022-06-30 11:13:19:INFO:RORRetriever: 1 Input Affiliations
    substring  score matching_type  chosen                       ror
          MIT   0.90       ACRONYM   False https://ror.org/002r67t24                Manuk
          MIT   0.90       ACRONYM   False https://ror.org/02mb6z761                  Mya
          MIT   0.90       ACRONYM   False https://ror.org/02qk9nj07       Ministry of Inf
          MIT   0.90       ACRONYM   False https://ror.org/02yyma482                Inter
          MIT   0.90       ACRONYM   False https://ror.org/042nb2s44           Massachuset
          MIT   0.90       ACRONYM    True https://ror.org/047m6hg73 Management Intelligen
          MIT   0.90       ACRONYM   False https://ror.org/04mtcj695                Unive
MIT Libraries   0.77  COMMON TERMS   False https://ror.org/00n3xk635
MIT Libraries   0.76  COMMON TERMS   False https://ror.org/05mgwy297
MIT Libraries   0.59         FUZZY   False https://ror.org/03qa42z85
MIT Libraries   0.58  COMMON TERMS   False https://ror.org/00w2b1v71
MIT Libraries   0.57  COMMON TERMS   False https://ror.org/022z6jk58
MIT Libraries   0.56  COMMON TERMS   False https://ror.org/0239rpj17
MIT Libraries   0.55  COMMON TERMS   False https://ror.org/00k8bhh21
MIT Libraries   0.50  COMMON TERMS   False https://ror.org/00rj4dg52
MIT Libraries   0.46  COMMON TERMS   False https://ror.org/03fg5ns40
MIT Libraries   0.44  COMMON TERMS   False https://ror.org/00v140q16
MIT Libraries   0.22         FUZZY   False https://ror.org/05nfbnp91             Academy
2022-06-30 11:13:20:INFO:numexpr.utils: NumExpr defaulting to 8 threads.
2022-06-30 11:13:20:INFO:RORRetriever: 1 new RORs written to AffiliationAPI_RORData__20220
2022-06-30 11:13:20:INFO:RORRetriever: AffiliationAPI_RORData__20220630_11.tsv 1 RORs Foun
```

::::: {#block-yui_3_17_2_1_1656605115552_214723 .sqs-block .html-block .sqs-block-html block-type="2" border-radii="{"topLeft":{"unit":"px","value":0.0},"topRight":{"unit":"px","value":0.0},"bottomLeft":{"unit":"px","value":0.0},"bottomRight":{"unit":"px","value":0.0}}"}

The flag **--noacronyms** can prevent RORs based on acronyms from being selected. In this case, that misses the correct ROR but, in cases where RORs are selected on the basis of abbreviations in many affiliations, this can be helpful. Try **RORRetriever -a "Metadata LLC USA"** for an interesting example.

Affiliation Best Practices

Keep in mind that ROR uses an algorithm to try to find RORs for affiliation strings and, as is always true, the algorithm is designed to work well in most cases. RORRetriever can help find RORs in batch mode if you have many affiliations or organization names to search. I find it very helpful, but always treat the results with care. See the [ROR API Documentation](https://ror.readme.io/docs/match-organization-names-to-ror-ids) for more tips and tricks and try RORRetriever if you have a big ROR search and don't know where to start!

Most important - write clear affiliations when submitting data and publications!

:::::