JournalList Ecosystem Web Crawler

Last Updated: July 4, 2022

This document describes how to use a Python based web crawler to download all the trust.txt files that can be reached by traversing all the trust.txt files referenced from a root URL (by default https://www.journallist.net/). This document is intended for Mac computers running macOS and assumes some familiarity with the "terminal" application and the UNIX shell.

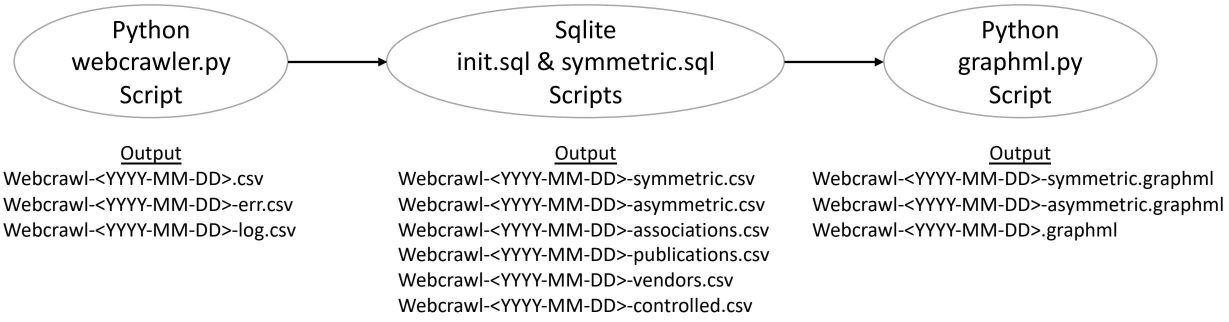
Preparing the environment:

1. Download the Python environment: https://www.python.org/downloads/
2. Install the Python requests module: https://docs.python-requests.org/en/master/user/install/#python-m-pip-install-requests
3. Download the yEd graphml editor/visualizer: https://www.yworks.com/products/yed/download#download

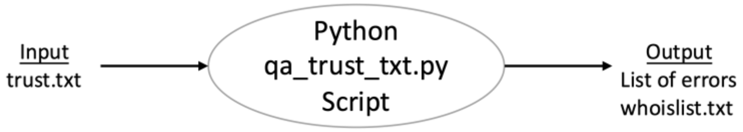
The cron.sh bash shell script captures this flow:

1. A Python webcrawler.py script crawls the trust.txt files generating .csv files capturing the results.
2. A pair of Sqlite scripts process the Webcrawl-<YYYY-MM-DD>.csv and Webcrawl-<YYYY-MM-DD>-err.csv files to derive the .csv files used in the final visualization step.
3. A Pytyon graphml.py script generates three .graphml files visualizing just the symmetric links found, just the asymmetric links found, and the complete social graph.

The processing flow is shown here.


Python
webcrawler.py
Script
Sqlite
init.sql & symmetric.sql
Scripts
Python
graphml.py
Script
Output
Webcrawl-<YYYY-MM-DD>.csv
Webcrawl-<YYYY-MM-DD>-err.csv
Webcrawl-<YYYY-MM-DD>-log.csv
Output
Webcrawl-<YYYY-MM-DD>-symmetric.csv
Webcrawl-<YYYY-MM-DD>-asymmetric.csv
Webcrawl-<YYYY-MM-DD>-associations.csv
Webcrawl-<YYYY-MM-DD>-publications.csv
Webcrawl-<YYYY-MM-DD>-vendors.csv
Webcrawl-<YYYY-MM-DD>-controlled.csv
Output
Webcrawl-<YYYY-MM-DD>-symmetric.graphml
Webcrawl-<YYYY-MM-DD>-asymmetric.graphml
Webcrawl-<YYYY-MM-DD>.graphml

In addition, there is a trust.txt QA python script that takes a trust.txt file as input and lists any errors encountered when parsing it.



Running the webcrawler:

Open a terminal window and change directory to the Journalist directory. Pull the following files from the JournalList GitHub repository: https://github.com/brownwolf1355/JournalList

1. cron.sh - a bash shell script that runs the python webcrawler, processes the results through sqlite, and generates graphml files of the results.
2. webcrawler.py - a python script that recursively crawls trust.txt files to capture the state of the trust.txt ecosystem. It captures a copy of all of the trust.txt files it finds and generates a .csv file of the contents of all of them.
3. init.sql - the initialization sqlite script that creates the intermediate tables used in the following sql script.
4. symmetric - a sql script that generates .csv files containing the symmetric links in the trust.txt ecosystem and list of associations, publishers, and vendors discovered.
5. graphml.py - a python script that generates three graphml files containing the symmetric links, the asymmetric links, and the full ecosystem including both the symmetric and asymmetric links.
6. qa_trust_txt.py - a python script that parses a trust.txt file and lists any errors it contains.

Run the cron.sh shell script:

```
$ bash cron.sh
```

The shell and python scripts will create a directory “Webcrawl-YYYY-MM-DD" that will contain all the results of crawling the trust.txt files. This directory will contain the following:

1. The trust.txt files downloaded through the web crawling process. Each trust.txt file will be given a unique name “www.<domain>-trust.txt”.
2. A log file useful for debugging, should anything go awry.
3. Several .csv files
    a. Webcrawl-YYYY-MM-DD.csv containing three columns: “srcurl,” “attr”, and “refurl”. Representing the Source URL, Attribute, and Referenced URL on each line of all downloaded trust.txt files.
    b. Webcrawl-YYYY-MM-DD-err.csv containing four columns: “srcurl”, “attr”, “refurl”, and “error”. Representing the Source URL, Attribute, Referenced URL, and the Error encountered when the Referenced URL trust.txt file is attempted to be downloaded, a blank “www.<domain>-trust.txt” file will be created in these cases. Errors that are identified as “HTTP Status Code:” or “Content type: text/html” indicate that the server was reached, but a trust.txt file doesn’t exist. Errors that are identified as “HTTP GET connection error exception occurred:” indicate that the server wasn’t able to be reached and may indicate an error in the spelling of the Referenced URL.

c. Webcrawl-YYYY-MM-DD -symmetric.csv containing three columns: “srcurl,” “attr”, and “refurl”. Representing the Source URL, Attribute, and Referenced URL for all symmetric links (e.g., “member” <-> “belongto” or “control” <-> “controlledby” or “vendor” <-> “customer”). The Attribute will be one of “member”, “control”, or “vendor”.
d. Webcrawl-YYYY-MM-DD -asymmetric.csv containing three columns: “srcurl,” “attr”, and “refurl”. Representing the Source URL, Attribute, and Referenced URL for all asymmetric links (e.g., “member” -> “belongto” reference without a corresponding “belongto” -> “member” reference). Note, these are the links listed in the Webcrawl-YYYY-MM-DD-err.csv file.
e. Webcrawl-YYYY-MM-DD-associations.csv containing one column “srcurl” listing all of the Source URLs that have more “member” attributes than “belongto” entries and all of the Referenced URLs that have “control” attributes.
f. Webcrawl-YYYY-MM-DD-publishers.csv containing one column “srcurl” listing all of the Source URLs that have more “belongto” attributes than “member” entries and all of the Referenced URLs that have “control” attributes.
g. Webcrawl-YYYY-MM-DD-vendors.csv containing one column “srcurl” listing all of the Referenced URLs that have “vendor” attributes.
h. Webcrawl-YYYY-MM-DD-symmetric.graphml containing the graphml diagram for the all symmetric links discovered.
i. Webcrawl-YYYY-MM-DD-asymmetric.graphml containing the graphml diagram for the all asymmetric links discovered.
j. Webcrawl-YYYY-MM-DD.graphml containing the graphml for the all links discovered, both symmetric and asymmetric.
k. Webcrawl-YYYY-MM-DD -links.json containing the JSON description of all the links for the ArangoDB graph analysis
l. Webcrawl-YYYY-MM-DD -urls.json containing the JSON description of all the urls for the ArangoDB graph analysis

Visualizing the results:

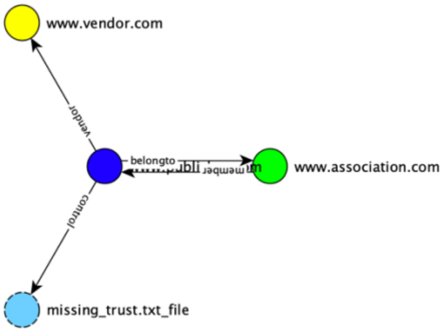
There are a few ways to analyze these results:

1. Through a yEd visual layout of the trust.txt ecosystem social graph
2. Through an Excel Pivot Table
3. Through a social graph analysis using ArangoDB

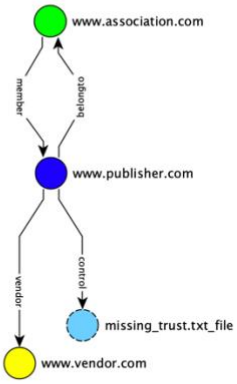
Visualizing the Social Graph with yEd:

To visualize the trust.txt ecosystem social graph, open the graphml files int yEd. Because the generated graphml files do not have any imbedded layout, the graph will be displayed with all the nodes on top of each other:

Use yEd’s layout tool to display the trust.txt social graph. I have found that the “Organic” layout with “Preferred Edge Length” set to 100 to be most useful. You may wish to manually place the graph legend (consisting of 4 nodes and 4 edges):



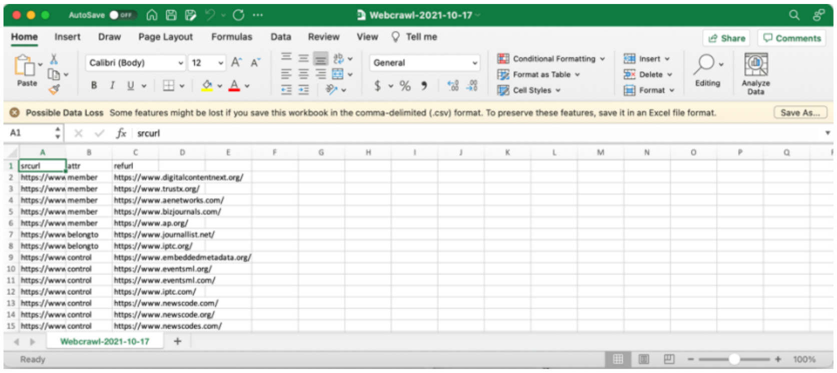
Use the selection tool to select these nodes and edges. You can obtain a better layout using the “Selection (Partial)” layout tool:



The “Component” layout tool has options for organizing the layout of the social graph.

Analysis Using an Excel Pivot Table:

To create the Excel Pivot Table, open the file “Webcrawl-YYYY-MM-DD.csv” with Excel. It will notify you regarding Possible Data Loss.

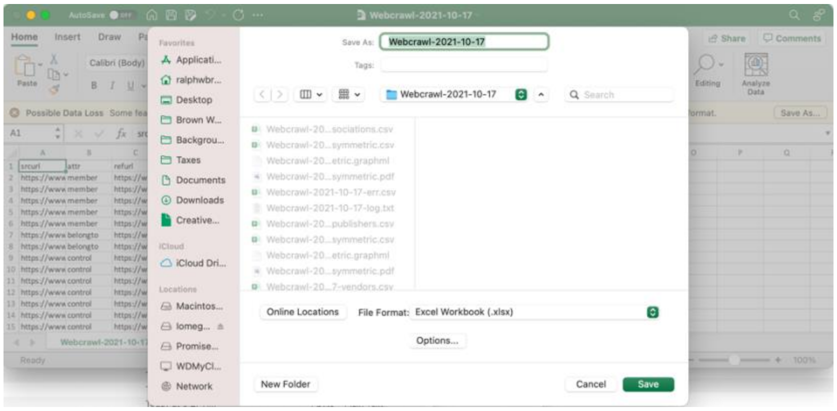


Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

A1 srcurl

	A	B	C
1	srcurl	attr	refurl
2	https://www	member	https://www.digitalcontentnext.org/
3	https://www	member	https://www.trustx.org/
4	https://www	member	https://www.aenetworks.com/
5	https://www	member	https://www.bizjournals.com/
6	https://www	member	https://www.ap.org/
7	https://www	belongto	https://www.journallist.net/
8	https://www	belongto	https://www.iptc.org/
9	https://www	control	https://www.embeddedmetadata.org/
10	https://www	control	https://www.eventsml.org/
11	https://www	control	https://www.eventsml.com/
12	https://www	control	https://www.iptc.com/
13	https://www	control	https://www.newscode.com/
14	https://www	control	https://www.newscode.org/
15	https://www	control	https://www.newscodes.com/

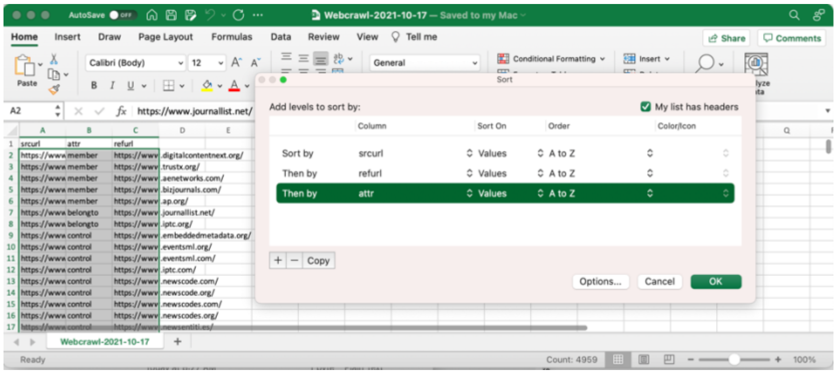
Save the file as an Excel Workbook “Webcrawl-YYYY-MM-DD.xlsx”.



Save As: Webcrawl-2021-10-17

File Format: Excel Workbook (.xlsx)

Select all the contents of this worksheet and custom sort the contents, first by srcurl, then by refurl, and then by attr.

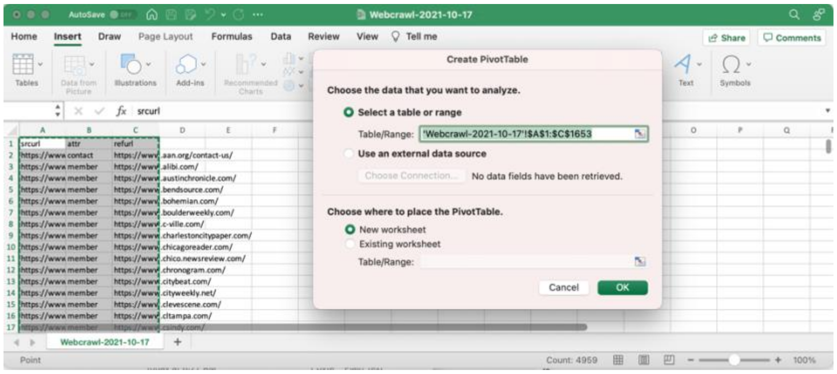


Add levels to sort by: ☑ My list has headers

	Column	Sort On	Order	Color/Icon
Sort by	srcurl	Values	A to Z	
Then by	refurl	Values	A to Z	
Then by	attr	Values	A to Z	

Count: 4959

Select all again and insert a PivotTable using selected range and placing it on a new worksheet.



fx srcurl

	A	B	C
1	srcurl	attr	refurl
2	https://www	contact	https://www.aan.org/contact-us/
3	https://www	member	https://www.alibi.com/
4	https://www	member	https://www.austinchronicle.com/
5	https://www	member	https://www.bendsource.com/
6	https://www	member	https://www.bohemian.com/
7	https://www	member	https://www.boulderweekly.com/
8	https://www	member	https://www.c-ville.com/
9	https://www	member	https://www.charlestoncitypaper.com/
10	https://www	member	https://www.chicagoreader.com/
11	https://www	member	https://www.chico.newsreview.com/
12	https://www	member	https://www.chronogram.com/
13	https://www	member	https://www.citybeat.com/
14	https://www	member	https://www.cityweekly.net/
15	https://www	member	https://www.clevescene.com/
16	https://www	member	https://www.cltampa.com/
17	https://www	member	https://www.csindy.com/

Create PivotTable

Choose the data that you want to analyze.

● Select a table or range

Table/Range: 'Webcrawl-2021-10-17'!$A$1:$C$1653

○ Use an external data source

Choose Connection... No data fields have been retrieved.

Choose where to place the PivotTable.

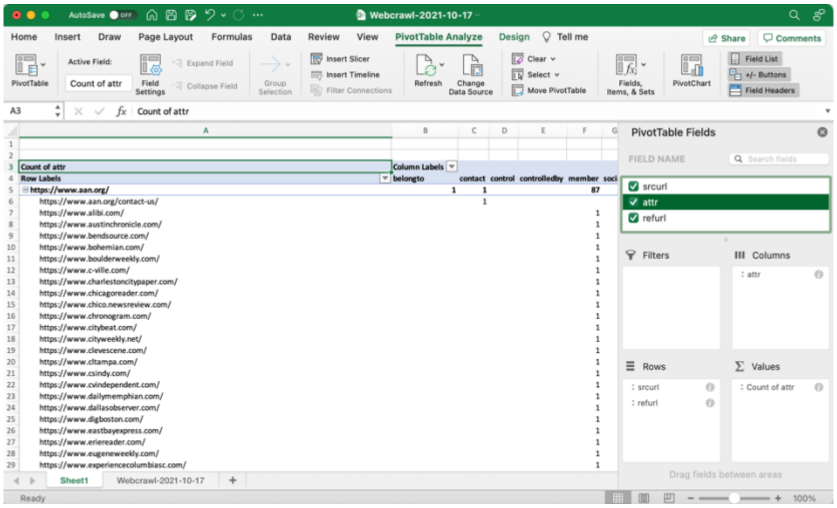
● New worksheet

○ Existing worksheet

Table/Range:

Cancel OK

Add srcurl to the Rows of the PivotTable Fields and refurl below this. Add attr to the Columns of the PivotTable Fields and to the Sum Values.



A3 *fx* Count of attr

	A	B	C	D	E	F	G
1							
2							
3	Count of attr	Column Labels					
4	Row Labels	belongto	contact	control	controlledby	member	soci
5	⊟ https://www.aan.org/	1	1			87	
6	https://www.aan.org/contact-us/		1				
7	https://www.alibi.com/					1	
8	https://www.austinchronicle.com/					1	
9	https://www.bendsource.com/					1	
10	https://www.bohemian.com/					1	
11	https://www.boulderweekly.com/					1	
12	https://www.c-ville.com/					1	
13	https://www.charlestoncitypaper.com/					1	
14	https://www.chicagoreader.com/					1	
15	https://www.chico.newsreview.com/					1	
16	https://www.chronogram.com/					1	
17	https://www.citybeat.com/					1	
18	https://www.cityweekly.net/					1	
19	https://www.clevescene.com/					1	
20	https://www.cltampa.com/					1	
21	https://www.csindy.com/					1	
22	https://www.cvindependent.com/					1	
23	https://www.dailymemphian.com/					1	
24	https://www.dallasobserver.com/					1	
25	https://www.digboston.com/					1	
26	https://www.eastbayexpress.com/					1	
27	https://www.eriereader.com/					1	
28	https://www.eugeneweekly.com/					1	
29	https://www.experiencecolumbiasc.com/					1	

PivotTable Fields

FIELD NAME

☑ srcurl

☑ attr

☑ refurl

Filters

Columns: attr

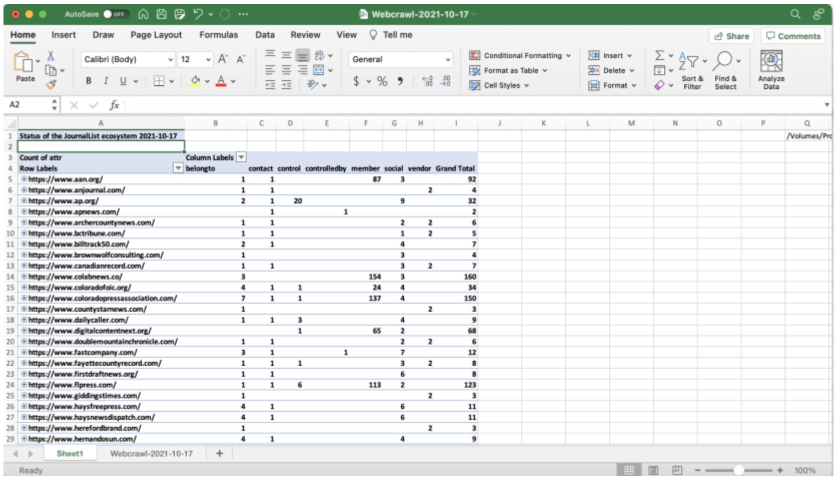
Rows: srcurl, refurl

Values: Count of attr

Collapse all the sub-rows by selecting cell A5 and “Collapse Field” under “PivotTable Analyze”. This will provide a summary of the different attributes used in each of the trust.txt files found.

Add a cell to capture the filename of the Excel workbook and create a title for the using the following formulas:

- Cell Q1
- Cell A1 =CONCAT("Status of the JournalList ecosystem ",MID(Q1,FIND("Webcrawl-",Q1)+9,10))



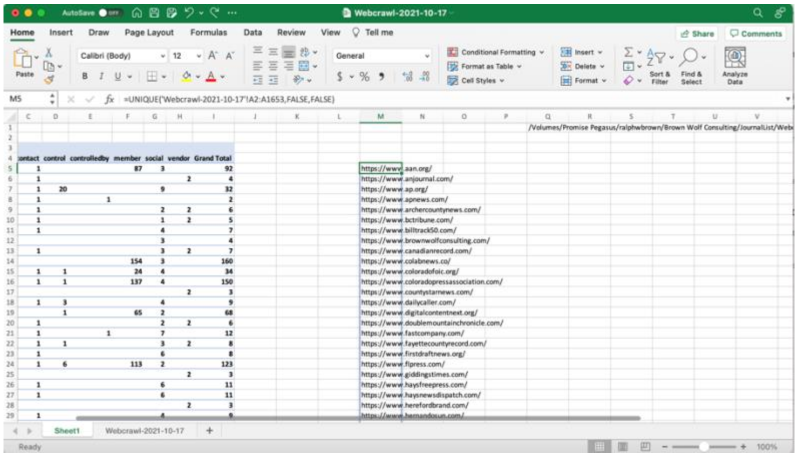
Status of the JournalList ecosystem 2021-10-17

Count of attr	Column Labels							
Row Labels	belongto	contact	control	controlledby	member	social	vendor	Grand Total
https://www.aan.org/	1	1			87	3		92
https://www.anjournal.com/	1	1					2	4
https://www.ap.org/	2	1	20			9		32
https://www.apnews.com/		1		1				2
https://www.archercountynews.com/	1	1				2	2	6
https://www.bctribune.com/	1	1				1	2	5
https://www.billtrack50.com/	2	1				4		7
https://www.brownwolfconsulting.com/	1					3		4
https://www.canadianrecord.com/	1	1				3	2	7
https://www.colabnews.co/	3				154	3		160
https://www.coloradofoic.org/	4	1	1		24	4		34
https://www.coloradopressassociation.com/	7	1	1		137	4		150
https://www.countystarnews.com/	1						2	3
https://www.dailycaller.com/	1	1	3			4		9
https://www.digitalcontentnext.org/			1		65	2		68
https://www.doublemountainchronicle.com/	1	1				2	2	6
https://www.fastcompany.com/	3	1		1		7		12
https://www.fayettecountyrecord.com/	1	1	1			3	2	8
https://www.firstdraftnews.org/	1	1				6		8
https://www.flpress.com/	1	1	6		113	2		123
https://www.giddingstimes.com/	1						2	3
https://www.haysfreepress.com/	4	1				6		11
https://www.haysnewsdispatch.com/	4	1				6		11
https://www.herefordbrand.com/	1						2	3
https://www.hernandosun.com/	4	1				4		9

/Volumes/Pro

Create a list of the organization URLs with that have published a trust.txt file by placing the following in cell M5 of the piviot table:

=UNIQUE('[Webcrawl-YYYY-MM-DD.xlsx]Webcrawl- YYYY-MM-DD!A2:A<last row>,FALSE,FALSE)

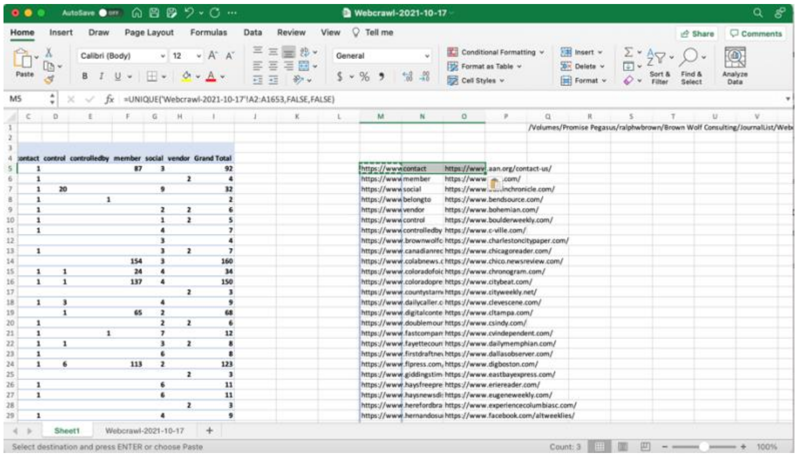


M5 =UNIQUE('Webcrawl-2021-10-17'!A2:A1653,FALSE,FALSE)

/Volumes/Promise Pegasus/ralphwbrown/Brown Wolf Consulting/JournalList/Web

https://www.aan.org/
https://www.anjournal.com/
https://www.ap.org/
https://www.apnews.com/
https://www.archercountynews.com/
https://www.bctribune.com/
https://www.billtrack50.com/
https://www.brownwolfconsulting.com/
https://www.canadianrecord.com/
https://www.colabnews.co/
https://www.coloradofoic.org/
https://www.coloradopressassociation.com/
https://www.countystarnews.com/
https://www.dailycaller.com/
https://www.digitalcontentnext.org/
https://www.doublemountainchronicle.com/
https://www.fastcompany.com/
https://www.fayettecountyrecord.com/
https://www.firstdraftnews.org/
https://www.flpress.com/
https://www.giddingstimes.com/
https://www.haysfreepress.com/
https://www.haysnewsdispatch.com/
https://www.herefordbrand.com/
https://www.hernandosun.com/

Copy this cell to the two adjacent cells



M5 =UNIQUE('Webcrawl-2021-10-17'!A2:A1653,FALSE,FALSE)

/Volumes/Promise Pegasus/ralphwbrown/Brown Wolf Consulting/JournalList/Web

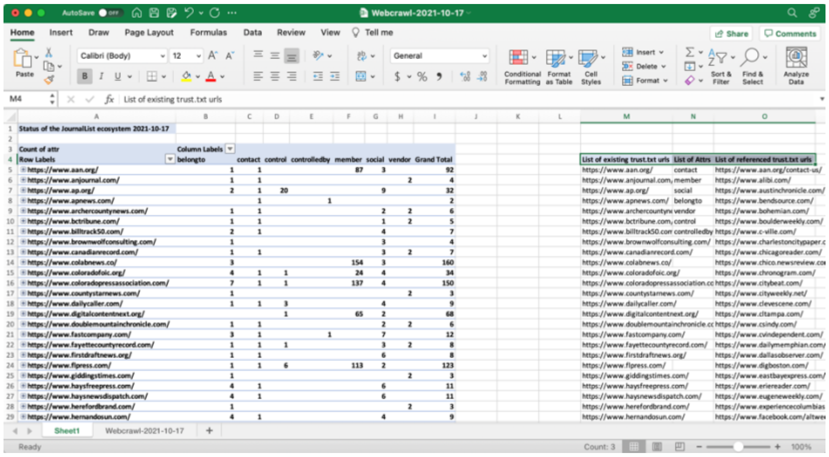
M	N	O
https://www	contact	https://www.aan.org/contact-us/
https://www	member	https://www[illegible].com/
https://www	social	https://www[illegible]inchronicle.com/
https://www	belongto	https://www.bendsource.com/
https://www	vendor	https://www.bohemian.com/
https://www	control	https://www.boulderweekly.com/
https://www	controlledby	https://www.c-ville.com/
https://www	.brownwolfc	https://www.charlestoncitypaper.com/
https://www	.canadianrec	https://www.chicagoreader.com/
https://www	.colabnews.c	https://www.chico.newsreview.com/
https://www	.coloradofoic	https://www.chronogram.com/
https://www	.coloradopre	https://www.citybeat.com/
https://www	.countystarn	https://www.cityweekly.net/
https://www	.dailycaller.c	https://www.clevescene.com/
https://www	.digitalconte	https://www.cltampa.com/
https://www	.doublemour	https://www.csindy.com/
https://www	.fastcompan	https://www.cvindependent.com/
https://www	.fayettecoun	https://www.dailymemphian.com/
https://www	.firstdraftne	https://www.dallasobserver.com/
https://www	.flpress.com,	https://www.digboston.com/
https://www	.giddingstim	https://www.eastbayexpress.com/
https://www	.haysfreepre	https://www.eriereader.com/
https://www	.haysnewsdi	https://www.eugeneweekly.com/
https://www	.herefordbra	https://www.experiencecolumbiasc.com/
https://www	.hernandosu	https://www.facebook.com/altweeklies/

Select destination and press ENTER or choose Paste

Count: 3

Put headers above these in cells M4, N4, and O4 using the following:

List of existing trust.txt urls	List of Attrs	List of referenced trust.txt urls



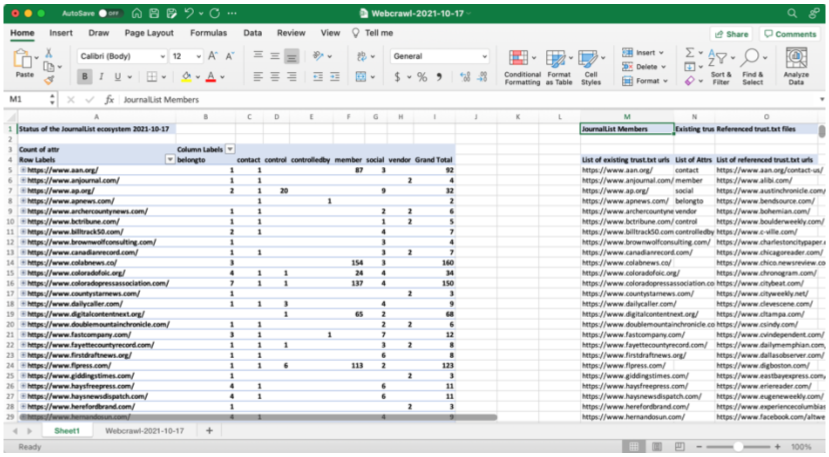
M4 | *fx* List of existing trust.txt urls

	A	B	C	D	E	F	G	H	I
1	Status of the JournalList ecosystem 2021-10-17								
2									
3	Count of attr	Column Labels							
4	Row Labels	belongto	contact	control	controlledby	member	social	vendor	Grand Total
5	https://www.aan.org/	1	1			87	3		92
6	https://www.anjournal.com/	1	1					2	4
7	https://www.ap.org/	2	1	20			9		32
8	https://www.apnews.com/		1		1				2
9	https://www.archercountynews.com/	1	1				2	2	6
10	https://www.bctribune.com/	1	1				1	2	5
11	https://www.billtrack50.com/	2	1				4		7
12	https://www.brownwolfconsulting.com/	1					3		4
13	https://www.canadianrecord.com/	1	1				3	2	7
14	https://www.colabnews.co/	3				154	3		160
15	https://www.coloradofoic.org/	4	1	1		24	4		34
16	https://www.coloradopressassociation.com/	7	1	1		137	4		150
17	https://www.countystarnews.com/	1						2	3
18	https://www.dailycaller.com/	1	1	3			4		9
19	https://www.digitalcontentnext.org/			1		65	2		68
20	https://www.doublemountainchronicle.com/	1	1				2	2	6
21	https://www.fastcompany.com/	3	1		1		7		12
22	https://www.fayettecountyrecord.com/	1	1	1			3	2	8
23	https://www.firstdraftnews.org/	1	1				6		8
24	https://www.flpress.com/	1	1	6		113	2		123
25	https://www.giddingstimes.com/	1						2	3
26	https://www.haysfreepress.com/	4	1				6		11
27	https://www.haysnewsdispatch.com/	4	1				6		11
28	https://www.herefordbrand.com/	1						2	3
29	https://www.hernandosun.com/	4	1				4		9

	M	N	O
4	List of existing trust.txt urls	List of Attrs	List of referenced trust.txt urls
5	https://www.aan.org/	contact	https://www.aan.org/contact-us/
6	https://www.anjournal.com,	member	https://www.alibi.com/
7	https://www.ap.org/	social	https://www.austinchronicle.com/
8	https://www.apnews.com/	belongto	https://www.bendsource.com/
9	https://www.archercountyne	vendor	https://www.bohemian.com/
10	https://www.bctribune.com,	control	https://www.boulderweekly.com/
11	https://www.billtrack50.com	controlledby	https://www.c-ville.com/
12	https://www.brownwolfconsulting.com/		https://www.charlestoncitypaper.c
13	https://www.canadianrecord.com/		https://www.chicagoreader.com/
14	https://www.colabnews.co/		https://www.chico.newsreview.co
15	https://www.coloradofoic.org/		https://www.chronogram.com/
16	https://www.coloradopressassociation.cc		https://www.citybeat.com/
17	https://www.countystarnews.com/		https://www.cityweekly.net/
18	https://www.dailycaller.com/		https://www.clevescene.com/
19	https://www.digitalcontentnext.org/		https://www.cltampa.com/
20	https://www.doublemountainchronicle.cc		https://www.csindy.com/
21	https://www.fastcompany.com/		https://www.cvindependent.com/
22	https://www.fayettecountyrecord.com/		https://www.dailymemphian.com/
23	https://www.firstdraftnews.org/		https://www.dallasobserver.com/
24	https://www.flpress.com/		https://www.digboston.com/
25	https://www.giddingstimes.com/		https://www.eastbayexpress.com/
26	https://www.haysfreepress.com/		https://www.eriereader.com/
27	https://www.haysnewsdispatch.com/		https://www.eugeneweekly.com/
28	https://www.herefordbrand.com/		https://www.experiencecolumbias
29	https://www.hernandosun.com/		https://www.facebook.com/altwe

Above this in cell M1, N1, and O1 create the following headers:

JournalList Members	Existing trust.txt files	Referenced trust.txt files

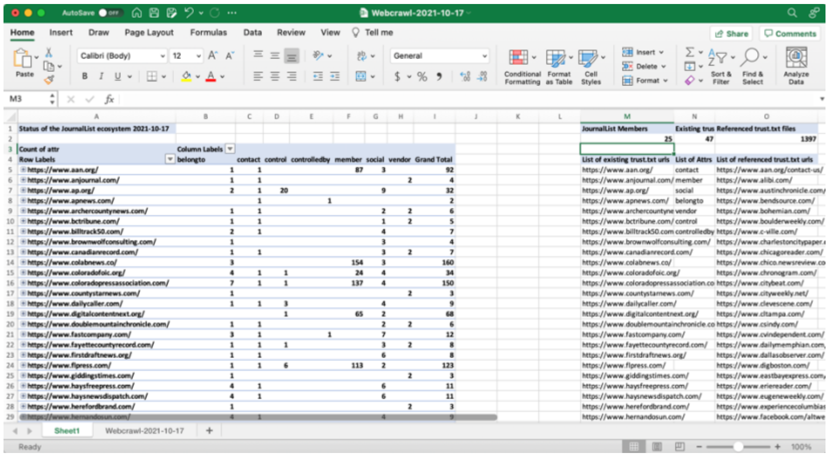


M1 | *fx* JournalList Members

	A	B	C	D	E	F	G	H	I	M	N	O
1	Status of the JournalList ecosystem 2021-10-17									JournalList Members	Existing trus	Referenced trust.txt files
2												
3	Count of attr	Column Labels										
4	Row Labels	belongto	contact	control	controlledby	member	social	vendor	Grand Total	List of existing trust.txt urls	List of Attrs	List of referenced trust.txt urls

	A	B	C	D	E	F	G	H	I	M	N	O
5	https://www.aan.org/	1	1			87	3		92	https://www.aan.org/	contact	https://www.aan.org/contact-us/
6	https://www.anjournal.com/	1	1					2	4	https://www.anjournal.com/	member	https://www.alibi.com/
7	https://www.ap.org/	2	1	20			9		32	https://www.ap.org/	social	https://www.austinchronicle.com/
8	https://www.apnews.com/		1		1				2	https://www.apnews.com/	belongto	https://www.bendsource.com/
9	https://www.archercountynews.com/	1	1				2	2	6	https://www.archercountyne	vendor	https://www.bohemian.com/
10	https://www.bctribune.com/	1	1				1	2	5	https://www.bctribune.com/	control	https://www.boulderweekly.com/
11	https://www.billtrack50.com/	2	1				4		7	https://www.billtrack50.com	controlledby	https://www.c-ville.com/
12	https://www.brownwolfconsulting.com/	1					3		4	https://www.brownwolfconsulting.com/		https://www.charlestoncitypaper.
13	https://www.canadianrecord.com/	1	1				3	2	7	https://www.canadianrecord.com/		https://www.chicagoreader.com/
14	https://www.colabnews.co/	3				154	3		160	https://www.colabnews.co/		https://www.chico.newsreview.co
15	https://www.coloradofoic.org/	4	1	1		24	4		34	https://www.coloradofoic.org/		https://www.chronogram.com/
16	https://www.coloradopressassociation.com/	7	1	1		137	4		150	https://www.coloradopressassociation.co		https://www.citybeat.com/
17	https://www.countystarnews.com/	1						2	3	https://www.countystarnews.com/		https://www.cityweekly.net/
18	https://www.dailycaller.com/	1	1	3			4		9	https://www.dailycaller.com/		https://www.clevescene.com/
19	https://www.digitalcontentnext.org/			1		65	2		68	https://www.digitalcontentnext.org/		https://www.cltampa.com/
20	https://www.doublemountainchronicle.com/	1	1				2	2	6	https://www.doublemountainchronicle.co		https://www.csindy.com/
21	https://www.fastcompany.com/	3	1		1		7		12	https://www.fastcompany.com/		https://www.cvindependent.com/
22	https://www.fayettecountyrecord.com/	1	1	1			3	2	8	https://www.fayettecountyrecord.com/		https://www.dailymemphian.com
23	https://www.firstdraftnews.org/	1	1				6		8	https://www.firstdraftnews.org/		https://www.dallasobserver.com/
24	https://www.flpress.com/	1	1	6		113	2		123	https://www.flpress.com/		https://www.digboston.com/
25	https://www.giddingstimes.com/	1						2	3	https://www.giddingstimes.com/		https://www.eastbayexpress.com/
26	https://www.haysfreepress.com/	4	1				6		11	https://www.haysfreepress.com/		https://www.eriereader.com/
27	https://www.haysnewsdispatch.com/	4	1				6		11	https://www.haysnewsdispatch.com/		https://www.eugeneweekly.com/
28	https://www.herefordbrand.com/	1						2	3	https://www.herefordbrand.com/		https://www.experiencecolumbias
29	https://www.hernandosun.com/	4	1				4		9	https://www.hernandosun.com/		https://www.facebook.com/altwe

The number of JournalList members can be selected from the pivot table, number of existing published trust.txt files is obtained by =COUNTA(M5#), the number of referenced, but not published trust.txt files is obtained by =COUNTA(O5#). The list of attributes used across all the existing trust.txt files is show in Column N.



Webcrawl-2021-10-17

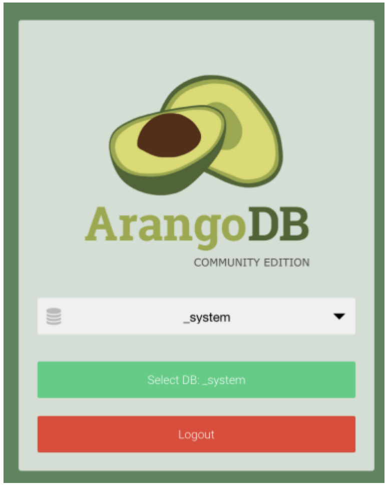
	A	B	C	D	E	F	G	H	I
1	Status of the JournalList ecosystem 2021-10-17								
2									
3	Count of attr	Column Labels							
4	Row Labels	belongto	contact	control	controlledby	member	social	vendor	Grand Total
5	https://www.aan.org/	1	1			87	3		92
6	https://www.anjournal.com/	1	1					2	4
7	https://www.ap.org/	2	1	20			9		32
8	https://www.apnews.com/		1		1				2
9	https://www.archercountynews.com/	1	1				2	2	6
10	https://www.bctribune.com/	1	1				1	2	5
11	https://www.billtrack50.com/	2	1				4		7
12	https://www.brownwolfconsulting.com/	1					3		4
13	https://www.canadianrecord.com/	1	1				3	2	7
14	https://www.colabnews.co/	3				154	3		160
15	https://www.coloradofoic.org/	4	1	1		24	4		34
16	https://www.coloradopressassociation.com/	7	1	1		137	4		150
17	https://www.countystarnews.com/	1						2	3
18	https://www.dailycaller.com/	1	1	3			4		9
19	https://www.digitalcontentnext.org/			1		65	2		68
20	https://www.doublemountainchronicle.com/	1	1				2	2	6
21	https://www.fastcompany.com/	3	1		1		7		12
22	https://www.fayettecountyrecord.com/	1	1	1			3	2	8
23	https://www.firstdraftnews.org/	1	1				6		8
24	https://www.flpress.com/	1	1	6		113	2		123
25	https://www.giddingstimes.com/	1						2	3
26	https://www.haysfreepress.com/	4	1				6		11
27	https://www.haysnewsdispatch.com/	4	1				6		11
28	https://www.herefordbrand.com/	1						2	3
29	https://www.hernandosun.com/	4	1				4		9

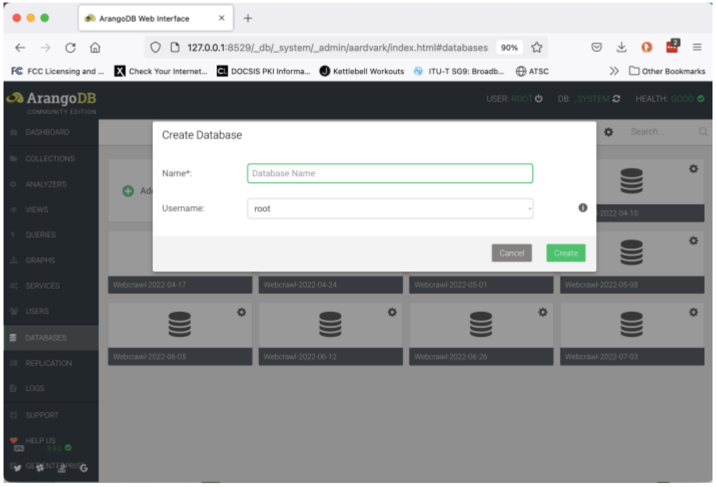
M	N	O
JournalList Members	Existing trus	Referenced trust.txt files
25	47	1397
List of existing trust.txt urls	List of Attrs	List of referenced trust.txt urls
https://www.aan.org/	contact	https://www.aan.org/contact-us/
https://www.anjournal.com/	member	https://www.alibi.com/
https://www.ap.org/	social	https://www.austinchronicle.com/
https://www.apnews.com/	belongto	https://www.bendsource.com/
https://www.archercountyne	vendor	https://www.bohemian.com/
https://www.bctribune.com/	control	https://www.boulderweekly.com/
https://www.billtrack50.com	controlledby	https://www.c-ville.com/
https://www.brownwolfconsulting.com/		https://www.charlestoncitypaper.
https://www.canadianrecord.com/		https://www.chicagoreader.com/
https://www.colabnews.co/		https://www.chico.newsreview.co
https://www.coloradofoic.org/		https://www.chronogram.com/
https://www.coloradopressassociation.co		https://www.citybeat.com/
https://www.countystarnews.com/		https://www.cityweekly.net/
https://www.dailycaller.com/		https://www.clevescene.com/
https://www.digitalcontentnext.org/		https://www.cltampa.com/
https://www.doublemountainchronicle.co		https://www.csindy.com/
https://www.fastcompany.com/		https://www.cvindependent.com/
https://www.fayettecountyrecord.com/		https://www.dailymemphian.com,
https://www.firstdraftnews.org/		https://www.dallasobserver.com/
https://www.flpress.com/		https://www.digboston.com/
https://www.giddingstimes.com/		https://www.eastbayexpress.com,
https://www.haysfreepress.com/		https://www.eriereader.com/
https://www.haysnewsdispatch.com/		https://www.eugeneweekly.com/
https://www.herefordbrand.com/		https://www.experiencecolumbias
https://www.hernandosun.com/		https://www.facebook.com/altwe

Social Graph Analysis Using ArangoDB:

To create a social graph analysis using ArandoDB (https://www.arangodb.com/) first create a graph database for the webcrawl:

Login to system DB and create a new graph database for the webcrawl.





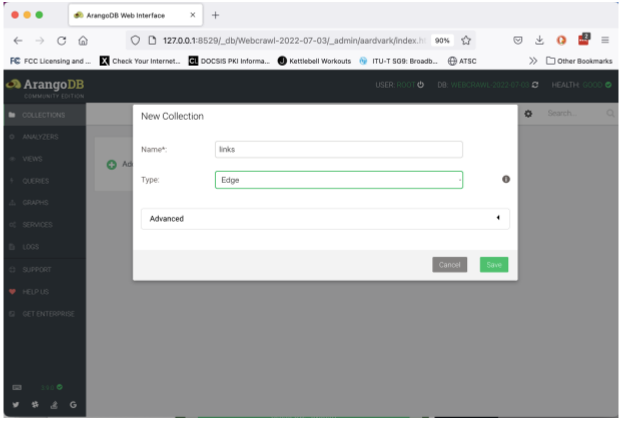
127.0.0.1:8529/_db/_system/_admin/aardvark/index.html#databases

Create Database

Name*: Database Name

Username: root

Login to the new DB and create a "links" edge collection and a "urls" document collection.

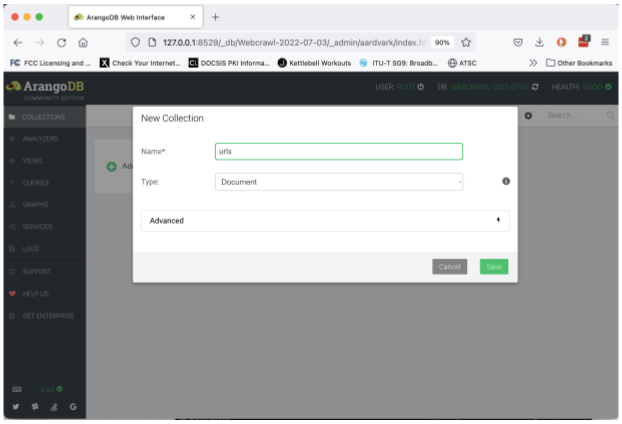


127.0.0.1:8529/_db/Webcrawl-2022-07-03/_admin/aardvark/index.h

New Collection

Name*: links

Type: Edge



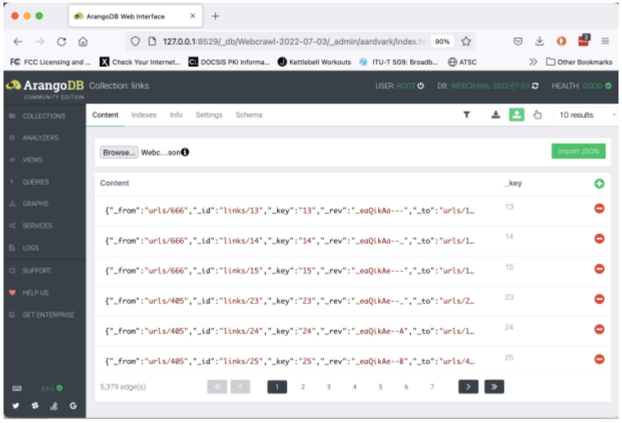
127.0.0.1:8529/_db/Webcrawl-2022-07-03/_admin/aardvark/index.h

New Collection

Name*: urls

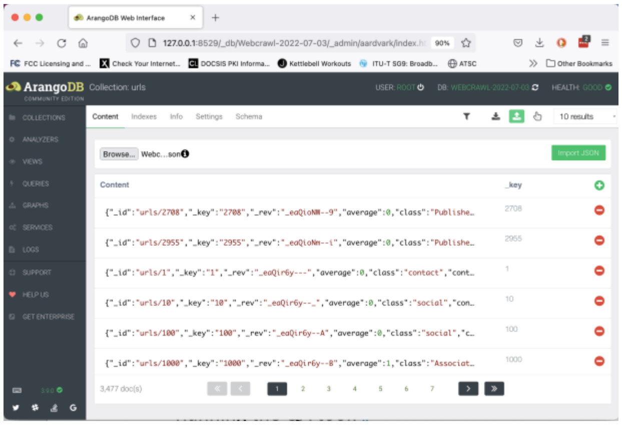
Type: Document

Upload the Webcrawl-YYYY-MM-DD -links.json and Webcrawl-YYYY-MM-DD -urls.json to their respective collections:


ArangoDB Web Interface
127.0.0.1:8529/_db/Webcrawl-2022-07-03/_admin/aardvark/index.h
ArangoDB Collection: links
USER: ROOT DB: WEBCRAWL-2022-07-03 HEALTH: GOOD
Content Indexes Info Settings Schema
10 results
Browse... Webc...son
Import JSON

Content	_key
{"_from":"urls/666","_id":"links/13","_key":"13","_rev":"_eaQikAa---","_to":"urls/1...	13
{"_from":"urls/666","_id":"links/14","_key":"14","_rev":"_eaQikAa--_","_to":"urls/1...	14
{"_from":"urls/666","_id":"links/15","_key":"15","_rev":"_eaQikAe---","_to":"urls/1...	15
{"_from":"urls/405","_id":"links/23","_key":"23","_rev":"_eaQikAe--_","_to":"urls/2...	23
{"_from":"urls/405","_id":"links/24","_key":"24","_rev":"_eaQikAe--A","_to":"urls/1...	24
{"_from":"urls/405","_id":"links/25","_key":"25","_rev":"_eaQikAe--B","_to":"urls/4...	25

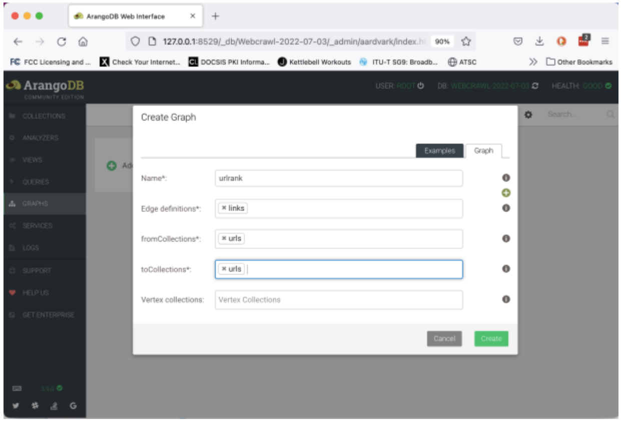
5,379 edge(s)


ArangoDB Web Interface
127.0.0.1:8529/_db/Webcrawl-2022-07-03/_admin/aardvark/index.h
ArangoDB Collection: urls
USER: ROOT DB: WEBCRAWL-2022-07-03 HEALTH: GOOD
Content Indexes Info Settings Schema
10 results
Browse... Webc...son
Import JSON

Content	_key
{"_id":"urls/2708","_key":"2708","_rev":"_eaQioNM--9","average":0,"class":"Publishe...	2708
{"_id":"urls/2955","_key":"2955","_rev":"_eaQioNm--i","average":0,"class":"Publishe...	2955
{"_id":"urls/1","_key":"1","_rev":"_eaQir6y---","average":0,"class":"contact","cont...	1
{"_id":"urls/10","_key":"10","_rev":"_eaQir6y--_","average":0,"class":"social","con...	10
{"_id":"urls/100","_key":"100","_rev":"_eaQir6y--A","average":0,"class":"social","c...	100
{"_id":"urls/1000","_key":"1000","_rev":"_eaQir6y--B","average":1,"class":"Associat...	1000

3,477 doc(s)

Then create a graph “urlrank” with Edge definitions from the “links” collection and the from/toCollections from the “urls” collection. If this is the first time, import the Webcrawl queries.



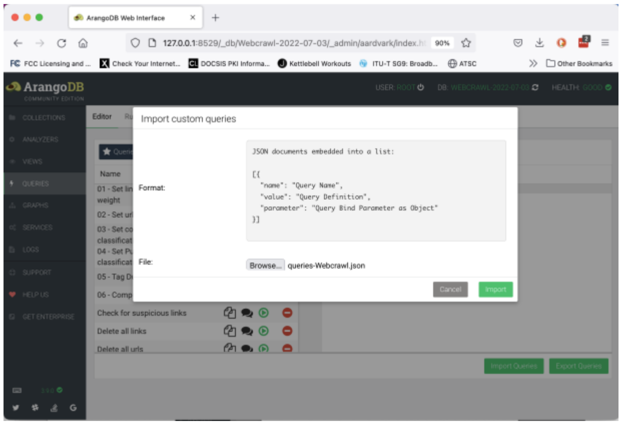
Name*: urlrank

Edge definitions*: × links

fromCollections*: × urls

toCollections*: × urls

Vertex collections: Vertex Collections



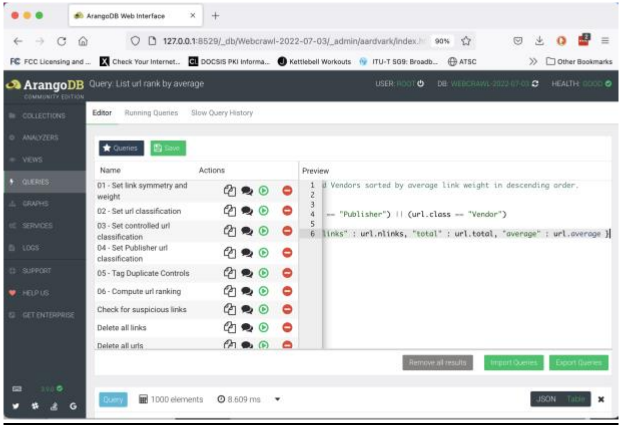
Format:

```
JSON documents embedded into a list:

[{
  "name": "Query Name",
  "value": "Query Definition",
  "parameter": "Query Bind Parameter as Object"
}]
```

File: Browse... queries-Webcrawl.json

Run the queries “01 - Set link symmetry and weight” through “06 - Compute url ranking” in sequence. The url ranking can be generated by running the “List url rank by average” query and saving the CSV file.



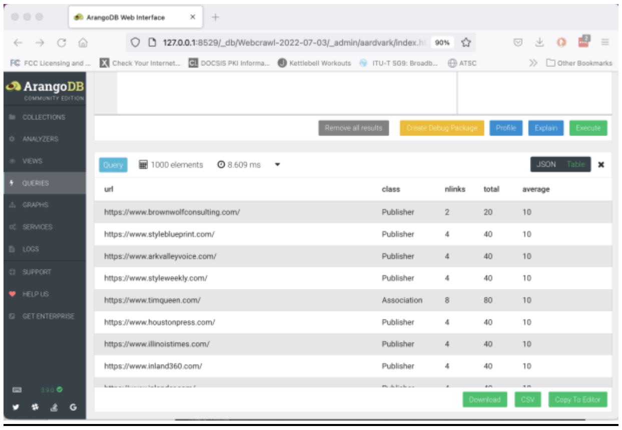
Name
01 - Set link symmetry and weight
02 - Set url classification
03 - Set controlled url classification
04 - Set Publisher url classification
05 - Tag Duplicate Controls
06 - Compute url ranking
Check for suspicious links
Delete all links
Delete all urls

Preview

```
d Vendors sorted by average link weight in descending order.


 == "Publisher") || (url.class == "Vendor")

links" : url.nlinks, "total" : url.total, "average" : url.average }
```



url	class	nlinks	total	average
https://www.brownwolfconsulting.com/	Publisher	2	20	10
https://www.styleblueprint.com/	Publisher	4	40	10
https://www.arkvalleyvoice.com/	Publisher	4	40	10
https://www.styleweekly.com/	Publisher	4	40	10
https://www.timqueen.com/	Association	8	80	10
https://www.houstonpress.com/	Publisher	4	40	10
https://www.illinoistimes.com/	Publisher	4	40	10
https://www.inland360.com/	Publisher	4	40	10

The resulting CSV file can be imported into Excel and filtered as desired.

Generating trust.txt files from a website:

There are two options for running the scrapesite.py python script:

```
$ python3.10 sitescrape.py "Publication Name" Error! Hyperlink reference not valid.
```

This generates an output file <website>-trust.txt containing the trust.txt file for the referenced website.

```
$ python3.10 sitescrape.py input.csv output.csv
```

Where input.csv contains a list of publication names and their websites.

- input.csv - a .csv file containing a list of sites to scan [Name, Website]
- output.csv - the output .csv file containing [Website, Contact, Facebook, Instagram, Twitter, Youtube, LinkedIn, Vendor, Copyright, Control]

This output.csv can be run through an awk script to generate trust.txt files based on this output. The awk script will need to be modified for the specific Association. This example is for the Texas Press Associaiton.

```
$ awk -F "," -f tpa.awk tpa-output.csv
```

Open the output.csv file in Excel and ad the following columns M through R (where 339 is replaced by the number of rows in the sheet. Copy the equations through all the rows in the sheet and filter all of the columns.

Start	End	Subdirectory	Copyright	Top URL	Count
=FIND("//",B2)+2	=IF(RIGHT(B2,1)="/",LEN(B2)-1,LEN(B2))	=IFERROR(FIND("/",MID(B2,M2,N2-M2)),0)	=COUNTIF(K$2:K$339,K2)	=IF(O2<>0,LEFT(B2,M2+O2-1),"")	=IF(O2<>0,COUNTIF(Q$2:Q$339,Q2),0)

Filter on column R (“Count”) to select all values are not zero or one. This will generate a list of publications that published as subdirectories of a common website. You can then filter on column Q (“Top URL”) for each website that has multiple publications. Check to see that there has been a trust.txt file generated for the Top URL. If not, you can use the scrapesite.py Python script to generate it as described above.

All the unique social media and contact sites for all these publications must be added to the trust.txt file of the Top URL. You can then delete the trust.txt files that were generated for the subdirectories. They will be of the form “<website>-<subdirectory>-trust.txt”. You should have one “<website>-trust.txt” file. Do this for each Top URL.

It may be necessary to generate a trust.txt file for publications with common ownership or control that are published on different websites, as opposed to through subdirectories of a common website. In this case you can filter on column L (“Control”) and generate a trust.txt file for each controlling website. You then need to add “control=” entries for each website in column B (“Website”).

Running the QA tool:

Run the qa_trust_txt.py python script:

```
$ python3.10 qa_trust_txt.py filename
```

The *filename* is a local trust.txt file to be checked.

For example, here is a sample trust.txt file:

```
# Brown Wolf Consulting LLC test trust.txt file
belongto=https://journallist.net/
controleby=https://www.brownwolfconsulting.com
member=https://www.brownwolfconsulting.org/
control=https://www.google.com/
social=https://www.linkedin.com/in/ralphwbrown/
social=https://twitter.com/RalphWBrown/
social=https://www.facebook.com/ralph.brown.5220/
```

The output looks like:

```
Invalid attribute at line 3 : controleby
Error at line 4 :  member=https://www.brownwolfconsulting.org/ - HTTP GET connection error
exception occurred: HTTPSConnectionPool(host='brownwolfconsulting.org', port=443): Max
retries exceeded with url: /trust.txt (Caused by
NewConnectionError('<urllib3.connection.HTTPSConnection object at 0x107d849a0>: Failed to
establish a new connection: [Errno 8] nodename nor servname provided, or not known'))
```

```
Error at line 5 :  control=https://www.google.com/ - HTTP Status Code: 404
```

It also generates a list of domain names in the whoislist.txt file that have connection errors. These can be used to check the status of these domains to understand the error further. For example:

```
$ whois brownwolfconsulting.org
```

Will generate the following:

```
% IANA WHOIS server
% for more information on IANA, visit http://www.iana.org
% This query returned 1 object

refer:        whois.pir.org

domain:       ORG

organisation: Public Interest Registry (PIR)
address:      11911 Freedom Drive 10th Floor,
address:      Suite 1000
address:      Reston, VA 20190
address:      United States

contact:      administrative
name:         Director of Operations, Compliance and Customer Support
organisation: Public Interest Registry (PIR)
address:      11911 Freedom Drive 10th Floor,
address:      Suite 1000
address:      Reston, VA 20190
address:      United States
phone:        +1 703 889 5778
```

```
fax-no:      +1 703 889 5779
e-mail:      ops@pir.org

contact:      technical
name:         Senior Director, DNS Infrastructure Group
organisation: Afilias
address:      Building 3, Suite 105
address:      300 Welsh Road
address:      Horsham, Pennsylvania 19044
address:      United States
phone:        +1 215.706.5700
fax-no:      +1 215.706.5701
e-mail:      tld-tech-poc@afilias.info

nserver:      A0.ORG.AFILIAS-NST.INFO 199.19.56.1 2001:500:e:0:0:0:0:1
nserver:      A2.ORG.AFILIAS-NST.INFO 199.249.112.1 2001:500:40:0:0:0:0:1
nserver:      B0.ORG.AFILIAS-NST.ORG 199.19.54.1 2001:500:c:0:0:0:0:1
nserver:      B2.ORG.AFILIAS-NST.ORG 199.249.120.1 2001:500:48:0:0:0:0:1
nserver:      C0.ORG.AFILIAS-NST.INFO 199.19.53.1 2001:500:b:0:0:0:0:1
nserver:      D0.ORG.AFILIAS-NST.ORG 199.19.57.1 2001:500:f:0:0:0:0:1
ds-rdata:     26974 8 2
4fede294c53f438a158c41d39489cd78a86beb0d8a0aeaff14745c0d16e1de32

whois:       whois.pir.org

status:      ACTIVE
remarks:      Registration information: http://www.pir.org

created:      1985-01-01
changed:       2020-10-27
```

```
source:       IANA


# whois.pir.org


NOT FOUND
>>> Last update of WHOIS database: 2021-10-21T15:10:55Z <<<
```