

JournalList Ecosystem Web Crawler

Last Updated: October 21, 2021

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.journalist.net/>). This document is intended for Mac computers running macOS and assumes some familiarity with the “terminal” application and the UNIX shell.

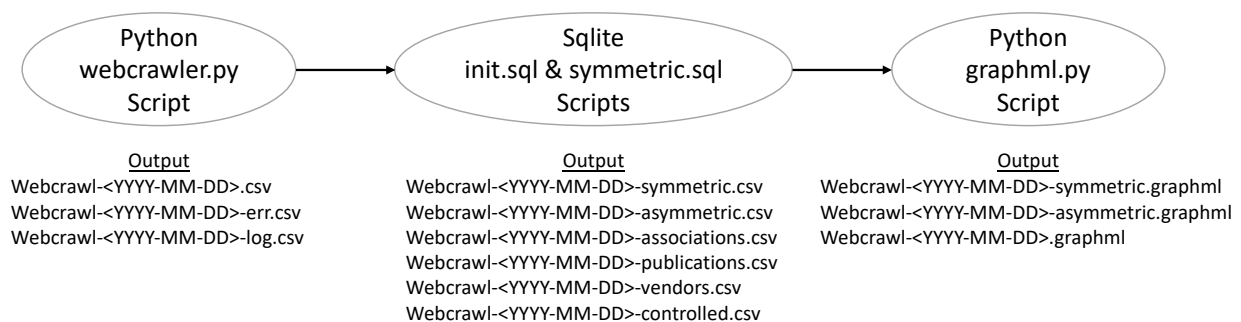
Preparing the environment:

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 Python 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.



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.

Visualizing the results:

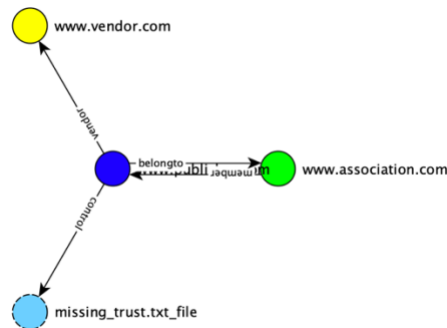
There are a couple ways to analyze these results:

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

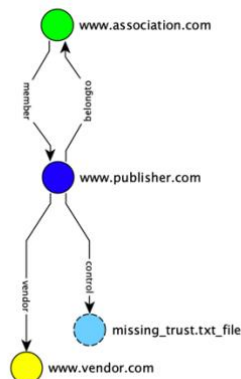
To visualize the trust.txt ecosystem social graph, open the graphml files into yEd. Because the generated graphml files do not have any imbedded layout, the graph will be displayed with all the nodes on top of each other:



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):

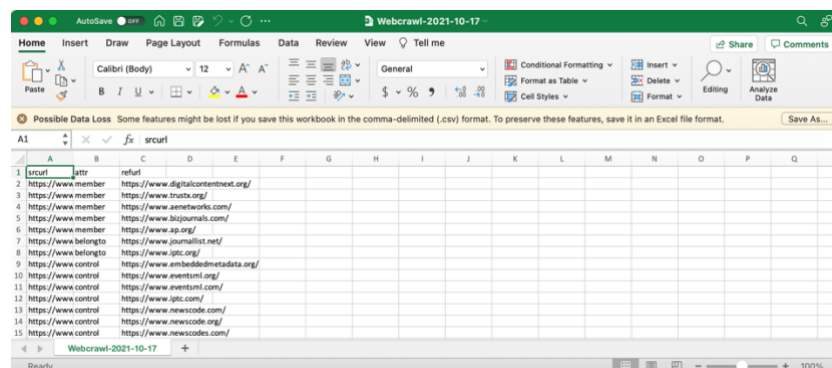


Use the selection tool to select these nodes and edges. You can obtain a better layout using the “Selection (Partial)” layout tool:

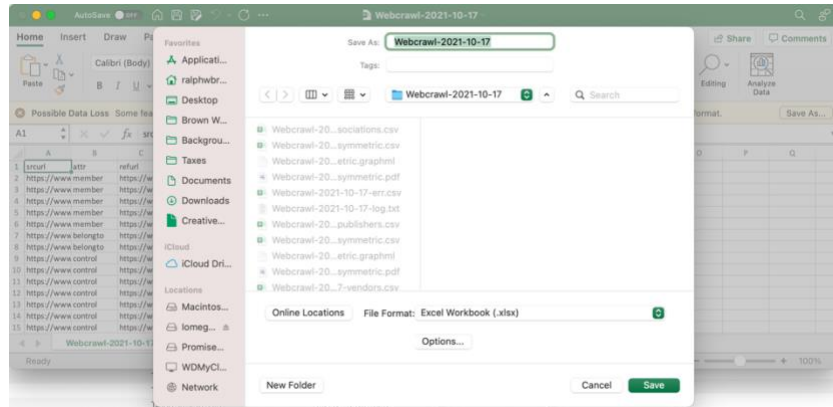


The “Component” layout tool has options for organizing the layout of the social graph.

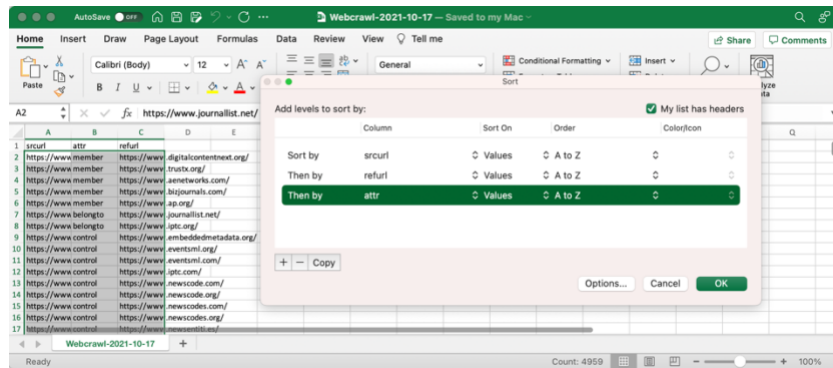
To create the Excel Pivot Table, open the file “Webcrawl-YYYY-MM-DD.csv” with Excel. It will notify you regarding Possible Data Loss.



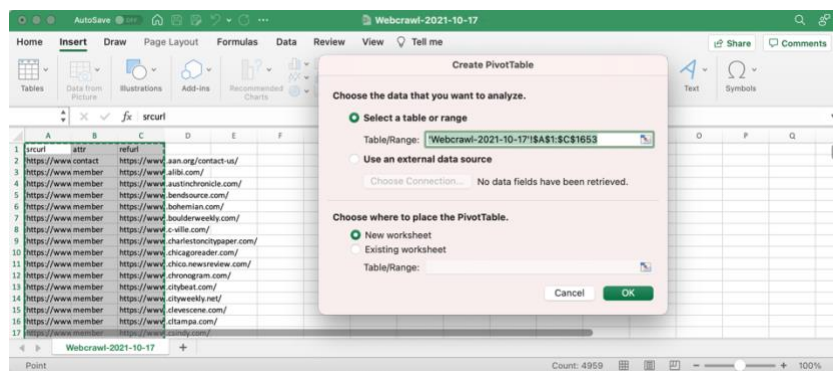
Save the file as an Excel Workbook “Webcrawl-YYYY-MM-DD.xlsx”.



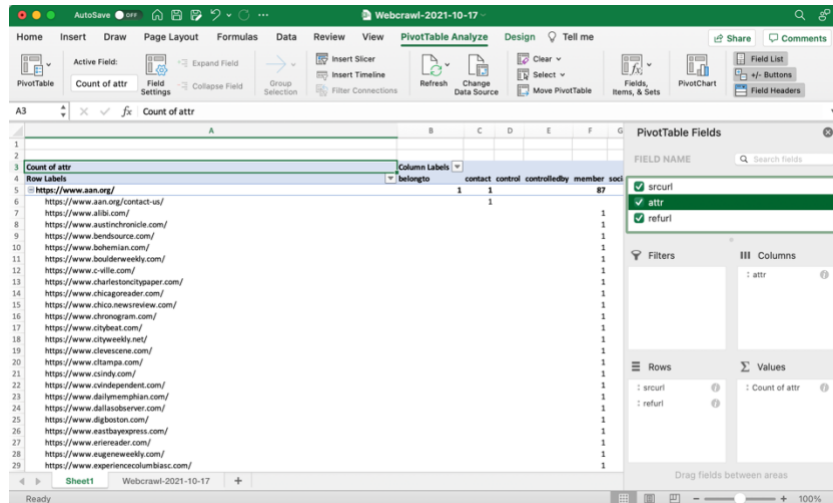
Select all the contents of this worksheet and custom sort the contents, first by srcurl, then by refurl, and then by attr.



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



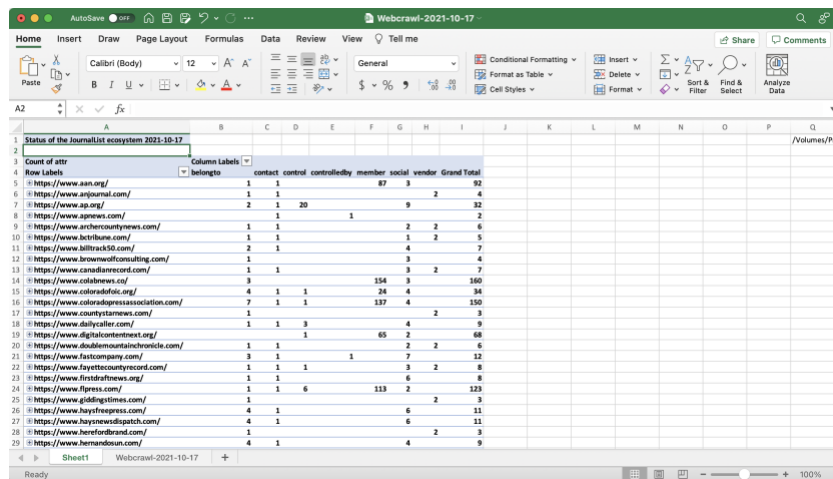
Add srcurl to the Rows of the PivotTable Fields and refurl below this. Add attr to the Columns of the PivotTable Fields and to the Sum Values.



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))



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

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

	contact	control	controlledby	member	social	vendor	Grand Total
1	1			87	3		92
2	1					2	4
3	1	20		9			32
4	1		1				2
5	1			2	2		6
6	1			3	2	5	
7	1			4		7	
8	1			3		4	
9	1			154	3	2	160
10	1	1		24	4		34
11	1	1		137	4		150
12	1					2	3
13	1	3		4		9	
14	1			65	2	68	
15	1			2	2	6	
16	1		1			7	12
17	1	1		3	2	8	
18	1			6		8	
19	1	6		113	2	123	
20						2	3
21						6	11
22						6	11
23						2	3
24						4	9

Copy this cell to the two adjacent cells

	contact	control	controlledby	member	social	vendor	Grand Total	
1	1			87	3		92	https://www.aan.org/
2	1					2	4	https://www.aajournal.com/
3	1	20		9			32	https://www.ap.org/
4	1		1				2	https://www.apnews.com/
5	1			2	2		6	https://www.architecturalnews.com/
6	1			3	2	5		https://www.tribune.com/
7	1			4		7		https://www.bitrack30.com/
8	1			3		4		https://www.brownwoolconsulting.com/
9	1			154	3	2	160	https://www.canadianrecord.com/
10	1	1		24	4		34	https://www.colabnews.ca/
11	1	1		137	4		150	https://www.colabnews.ca/
12	1					2	3	https://www.colabnews.ca/
13	1	3		4		9		https://www.colabnews.ca/
14	1			65	2	68		https://www.colabnews.ca/
15	1			2	2	6		https://www.colabnews.ca/
16	1		1			7	12	https://www.colabnews.ca/
17	1	1		3	2	8		https://www.colabnews.ca/
18	1			6		8		https://www.colabnews.ca/
19	1	6		113	2	123		https://www.colabnews.ca/
20						2	3	https://www.colabnews.ca/
21						6	11	https://www.colabnews.ca/
22						6	11	https://www.colabnews.ca/
23						2	3	https://www.colabnews.ca/
24						4	9	https://www.colabnews.ca/

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
---------------------------------	---------------	-----------------------------------

	contact	control	controlledby	member	social	vendor	Grand Total	List of existing trust.txt urls	List of Attrs	List of referenced trust.txt urls
1	1			87	3		92	https://www.aan.org/	contact	https://www.aan.org/contact-us/
2	1					2	4	https://www.aajournal.com/	member	https://www.aajournal.com/member
3	1	20		9			32	https://www.ap.org/	social	https://www.austinchronicle.com/
4	1		1				2	https://www.apnews.com/	belongs	https://www.bendsource.com/
5	1			2	2		6	https://www.architecturalnews.com/	vendor	https://www.bendsource.com/
6	1			3	2	5		https://www.tribune.com/	control	https://www.bendsource.com/
7	1			4		7		https://www.bitrack30.com/	controlledby	https://www.bendsource.com/
8	1			3		4		https://www.brownwoolconsulting.com/	social	https://www.bendsource.com/
9	1			154	3	2	160	https://www.canadianrecord.com/		https://www.bendsource.com/
10	1	1		24	4		34	https://www.colabnews.ca/		https://www.bendsource.com/
11	1	1		137	4		150	https://www.colabnews.ca/		https://www.bendsource.com/
12	1					2	3	https://www.colabnews.ca/		https://www.bendsource.com/
13	1	3		4		9		https://www.colabnews.ca/		https://www.bendsource.com/
14	1			65	2	68		https://www.colabnews.ca/		https://www.bendsource.com/
15	1			2	2	6		https://www.colabnews.ca/		https://www.bendsource.com/
16	1		1			7	12	https://www.colabnews.ca/		https://www.bendsource.com/
17	1	1		3	2	8		https://www.colabnews.ca/		https://www.bendsource.com/
18	1			6		8		https://www.colabnews.ca/		https://www.bendsource.com/
19	1	6		113	2	123		https://www.colabnews.ca/		https://www.bendsource.com/
20						2	3	https://www.colabnews.ca/		https://www.bendsource.com/
21						6	11	https://www.colabnews.ca/		https://www.bendsource.com/
22						6	11	https://www.colabnews.ca/		https://www.bendsource.com/
23						2	3	https://www.colabnews.ca/		https://www.bendsource.com/
24						4	9	https://www.colabnews.ca/		https://www.bendsource.com/

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

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

The screenshot shows an Excel spreadsheet with the following data:

Row Labels	belongs to	contact	control	controlledby	member	social	vendor	Grand Total
https://www.aan.org/	1	1			87	3		92
https://www.aajournal.com/	1	1				2	4	
https://www.ap.org/	2	1	20		9			32
https://www.apnews.com/	1	1		1				2
https://www.archercountynews.com/	1	1			2	2		6
https://www.bctrb.com/	1	1			1	2	5	
https://www.bitrack50.com/	2	1			4			7
https://www.brownmoffconsulting.com/	1	1			3			4
https://www.canadianrecord.com/	1	1			3	2	7	
https://www.cdnnews.ca/	3				154	3		160
https://www.coladefile.org/	4	1	1		24	4		34
https://www.colodapressassociation.com/	7	1	1		137	4		150
https://www.countrytimes.com/	1					2	3	
https://www.dailycaller.com/	1	1	3		4			9
https://www.digitalsentinel.net/	1	1			65	2		68
https://www.dublinmountainchronicle.com/	1	1			2	2	6	
https://www.fastcompany.com/	3	1		1	7			12
https://www.fayettecountyprecord.com/	1	1	1		3	2	8	
https://www.firstdraftnews.org/	1	1			6			8
https://www.foxpress.com/	1	1	6		113	2		123
https://www.gldigitaltimes.com/	1					2	3	
https://www.hayfreepress.com/	4	1			6		11	
https://www.haynewsdispatch.com/	4	1			6		11	
https://www.heraldtribune.com/	1				2	3		
https://www.heraldtribune.com/	4	1			6	2	9	

Below the pivot table, there are two lists of URLs:

List of existing trust.txt urls	List of Attribs	List of referenced trust.txt urls
https://www.aan.org/	contact	https://www.aan.org/contact-us/
https://www.aajournal.com/	member	https://www.aajournal.com/member
https://www.ap.org/	social	https://www.austinchronicle.com/
https://www.apnews.com/	belongs to	https://www.bendsource.com/
https://www.archercountynews.com/	vendor	https://www.bohemian.com/
https://www.bctrb.com/	control	https://www.boulderweekly.com/
https://www.bitrack50.com/	controlledby	https://www.c-ville.com/
https://www.brownmoffconsulting.com/		https://www.charlestoncitypaper.com/
https://www.canadianrecord.com/		https://www.chicagoreader.com/
https://www.cdnnews.ca/		https://www.chico.newswire.co
https://www.coladefile.org/		https://www.chronogram.com/
https://www.colodapressassociation.com/		https://www.chybest.com/
https://www.countrytimes.com/		https://www.civeweb.com/
https://www.dailycaller.com/		https://www.chicago.com/
https://www.digitalsentinel.net/		https://www.chicago.com/
https://www.dublinmountainchronicle.com/		https://www.cindy.com/
https://www.fastcompany.com/		https://www.cxdpendent.com/
https://www.fayettecountyprecord.com/		https://www.dailymemphian.com/
https://www.firstdraftnews.org/		https://www.dallasobserver.com/
https://www.foxpress.com/		https://www.digitalsentinel.net/
https://www.gldigitaltimes.com/		https://www.digitalsentinel.net/
https://www.hayfreepress.com/		https://www.digitalsentinel.net/
https://www.haynewsdispatch.com/		https://www.digitalsentinel.net/
https://www.heraldtribune.com/		https://www.digitalsentinel.net/
https://www.heraldtribune.com/		https://www.digitalsentinel.net/

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.

The screenshot shows the same Excel spreadsheet as the previous image, but with the following updates:

- The pivot table data remains the same.
- The "List of existing trust.txt urls" now includes 25 URLs.
- The "List of referenced trust.txt urls" now includes 47 URLs.
- The "List of attributes" in Column N now includes 197 attributes.

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://journalist.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: Afiliias

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 <<<