



Platforms, Experts, Tools: Specialised Cyber-Activists Network

ONLINE USER GUIDE

**For using automated technologies in
monitoring hate speech content**



Project funded by the European Union's Rights,
Equality and Citizenship Programme (2014-2020)

About the Project

The EU-funded project sCAN – Platforms, Experts, Tools: Specialised Cyber-Activists Network (2018-2020), coordinated by Licra (International League Against Racism and Antisemitism), aims at gathering expertise, tools, methodology and knowledge on cyber hate and developing transnational comprehensive practices for identifying, analysing, reporting and counteracting online hate speech. This project draws on the results of successful European projects already realised, for example the project “Research, Report, Remove: Countering Cyber-Hate phenomena” and “Facing Facts”, and strives to continue, emphasize and strengthen the initiatives developed by civil society for counteracting hate speech.

Through cross-European cooperation, the project partners are enhancing and (further) intensifying their fruitful collaboration. The sCAN project partners are contributing to selecting and providing relevant automated monitoring tools to improve the detection of hateful content. Another key aspect of sCAN is the strengthening of the monitoring actions (e.g. the monitoring exercises) set up by the European Commission. The project partners are also jointly gathering knowledge and findings to better identify, explain and understand trends of cyber hate at a transnational level. Furthermore, this project aims to develop cross-European capacity by providing e-learning courses for cyber-activists, moderators and tutors through the Facing Facts Online platform.

sCAN is implemented by ten different European partners, namely ZARA – Zivilcourage und Anti-Rassismus-Arbeit from Austria, CEJI – A Jewish contribution to an inclusive Europe from Belgium, Human Rights House Zagreb from Croatia, Romea from Czech Republic, Licra – International League Against Racism and Antisemitism and Respect Zone from France, jugendschutz.net from Germany, CESIE from Italy, Latvian Centre For Human Rights from Latvia and the University of Ljubljana, Faculty of Social Sciences from Slovenia.

The sCAN project is funded by the European Commission Directorate – General for Justice and Consumers, within the framework of the Rights, Equality and Citizenship (REC) Programme of the European Union.

Legal Disclaimer

This user guide was funded by the European Union’s Rights, Equality and Citizenship Programme (2014-2020).

The content of the user guide represents the views of the authors only and is the sole responsibility of the scan project consortium. The European Commission does not accept any responsibility for use that may be made of the information it contains.



Project funded by the European Union’s Rights, Equality and Citizenship Programme (2014-2020)

Table of contents

About the Project	2
Introduction	4
Automated monitoring Crawlers: How to detect hate speech on web 1.0 and social networks?	6
TAGS v 6.1	6
PROS AND CONS	10
HTTrack	11
PROS AND CONS	14
SociScraper	15
PROS AND CONS	17
CrowdTangle	19
PROS AND CONS	20
Algorithms and AI: SCAN and Factmata, a partnership for strengthening impact of automated tools in monitoring	22
Presentation	22
Partnership and methodology	22
PROS AND CONS	23
Conclusion	24
APPENDICES	25
Appendix 1: List of criteria	25
Appendix 2: A Hate Ontology for better understanding of key definitions	25

Introduction

Automated technologies that provide civil society organisations with a better environment in conducting monitoring projects are at the heart of a new complementary way to tackle online hate speech. This user guide has the purpose of presenting and explaining how a selection of existing automated tools that are straightforward for non-governmental organisations, Human rights and antiracist activists can be used for monitoring online hateful content.

There are a range of automated monitoring tools for the detection of online hate speech as for example web spidering, crawlers, artificial intelligence (AI). Social media tend to develop AI tools and algorithms as a new technology to moderate online content in an efficient way.

Nonetheless, these various ways are not easily accessible to civil society organisations and activists: consideration must be given to human capital, materials, resources and language challenges. Most NGOs do not have an IT desk or coding skills. Another key-element is the level of privacy of each platform. This level has direct consequences on possibilities for using automated technologies. The more public and open a social network is (as for example Twitter), the more effective is using automated monitoring tools.

It is also important to bear in mind that automated monitoring tools cannot be considered as the only efficient way to tackle online hate speech. Evading hate speech detection is possible not only by corrupting and coding text, but also by attacking a machine learning system itself. Moreover, human contribution in terms of knowledge, adaptation capacity and analysis are still crucial in monitoring hate speech.

In the framework of the sCAN project, jugendschutz.net and Licra have developed a common methodology in order to test a selection of automated tools.

This testing exercise has been organised in two different campaigns: the first one dedicated to crawlers and the second one dedicated to the AI developed by a partner company, Factmata.

The main objective of this work is to provide objective elements regarding the accuracy and relevance of the selected tools in order to integrate them in the SCAN consortium monitoring task.

- The first campaign has been organised at the beginning of the project for two months, in September and October 2018. This campaign was focused on testing several crawlers on platforms part of the Code of Conduct with the integration of relevant key words selected for the hate ontologies report. The SCAN consortium has developed hate ontologies in nine different languages to reflect national context and usual practices by users spreading hate speech. In order to select these keywords, each organisation part of the project had to identify users' trends – in terms of words, theories, rhetoric, etc. - in expressing hate towards individuals and groups. Due to the very quick evolution of hate speech vocabulary, the sCAN consortium plays a significant role in data collection. A non-exhaustive list of criteria has been defined regarding the choice of automated tools before their testing, in

two phases: price of the crawler/software, training, skills needed, manual assistance, results regarding hate speech categories, time, bugs and issues, pros and cons, linguistic settings (see annex).

- The second campaign has been launched from December 2018 until the end of this project. For this campaign, a partnership has been developed with the Company Factmata which proposed a technology mixing artificial intelligence, algorithm and expert knowledge to deal with hate speech, propaganda and fake news. jugendschutz.net and Licra have contributed as part of the community of users for fact-checking content for quality with the help of their AI algorithm. For this second testing phase, due to the algorithm, the test has been made with English content. Nonetheless, thanks to the partnership developed with Factmata, next steps would be to train the tool in other languages.
- In addition to these two testing campaigns, it was also crucial to share and disseminate online tools developed by civil society organisations and networks. As a consequence, an online tutorial training has been proposed to the consortium in order to use the INACH database from the 2nd monitoring exercise organised in May and June 2019.

The sCAN project focuses on available and inexpensive tools with a comprehensive interface in order to improve the monitoring process and collection of data. In this report, we will present available tools designed to counter hate speech and to improve hate speech removal's prospects

Automated monitoring Crawlers: How to detect hate speech on web 1.0 and social networks?

The following automated monitoring tools have been chosen for the first testing campaign of the project. They have been selected in order to match the expectations of NGOs and activists wanting a first-hand tool for detecting online content without technical expertise. The following list is of course non-exhaustive.

TAGS v 6.1

TAGS v 6.1 is a crawler dedicated to Twitter. The crawler allows users to set up and run an automated collection of search results on Twitter. It scrapes data by keywords/expressions, which are recorded in a Google Spreadsheet. The excel sheet offers information related to the account, tweet, date, retweet and sometimes other type of information if the user made it public (e.g. URL, retweet, language, location of the user, account name, etc.).



PREREQUISITES:

For using this tool, you will need:



1. To create/use a Gmail alias account with space capacity,
2. To create/use a Twitter alias account linked to this Gmail account (same mail address for both accounts),
3. To open both accounts in your web browser (Chrome, Firefox or Safari).

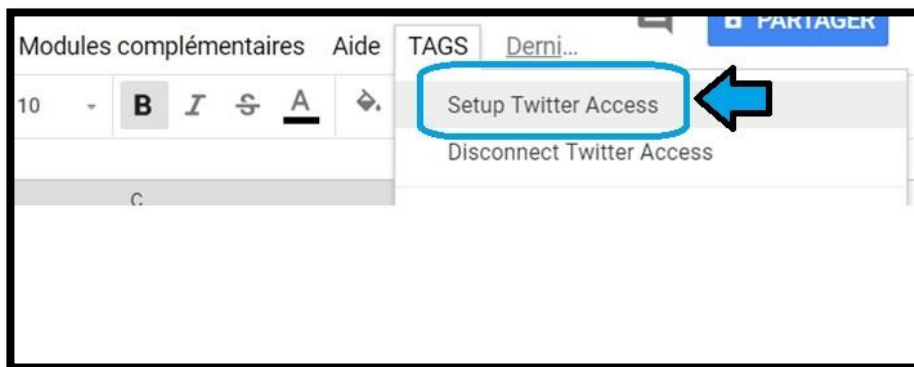
STEP 1:

1. Click on the [following link](#)

Notice: TAGS is a free Google Sheet template which lets you setup and run an automated collection of search results from Twitter thanks to key words.

STEP 2:

1. Click on  and then .
2. Then click on “create a Google Excel sheet”,
3. The template will be automatically generated,
4. Select the tab “TAGS” then “Setup Twitter Access” to allow the script collecting data through your twitter account.



Notice: Your web browser may request a validation for the operation (security access).

STEP 3:

1. Fill the form

Insert the key word here
(you can use 'OR', 'AND',
#, expressions, terms)

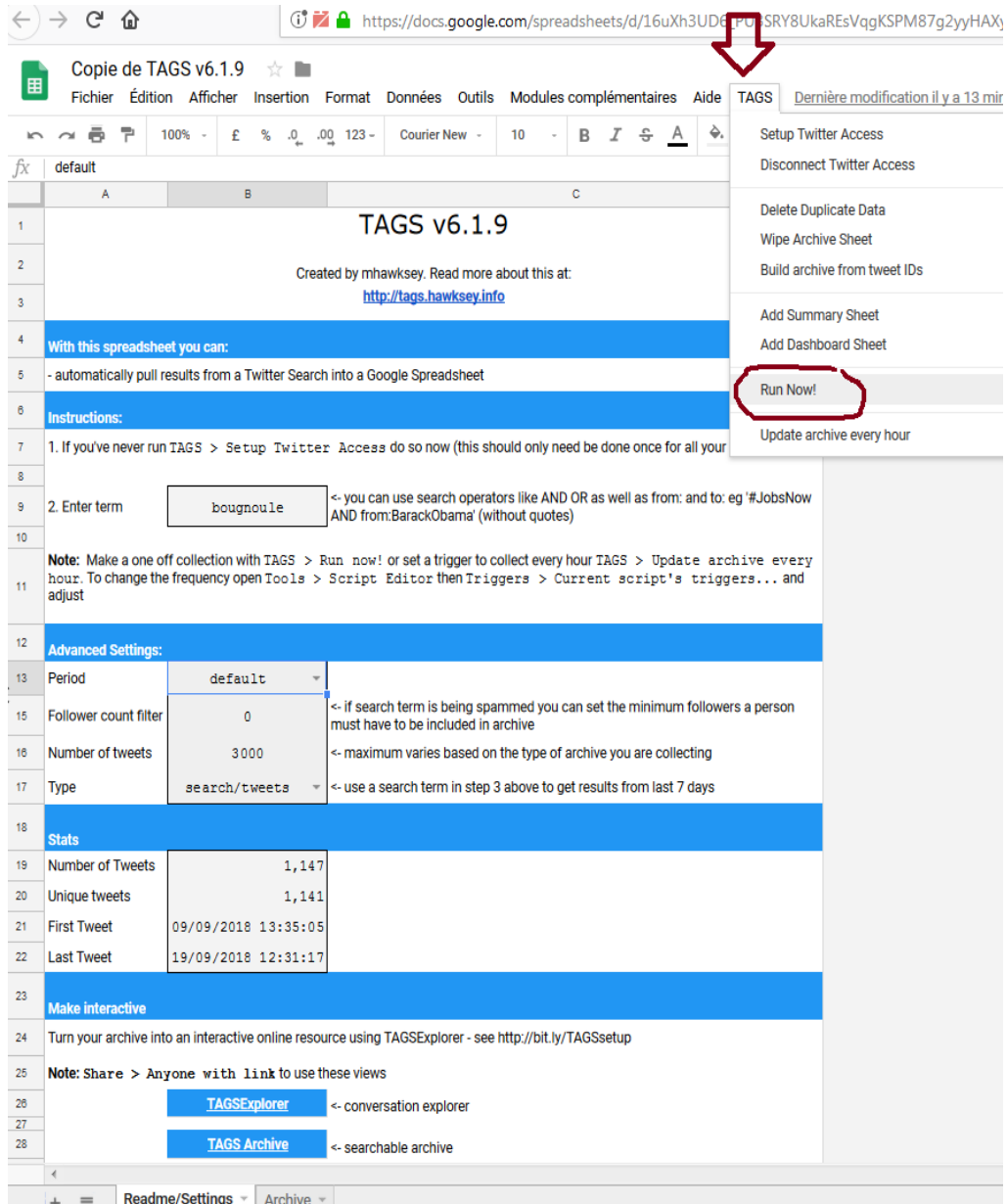
Select your parameters:
date (one week maxi-
mum), number of follow-
ers, number of tweets
and type of content

The screenshot shows a Google Sheet titled "Cople de TAGS v6.1.9.1". The spreadsheet content includes:

- Header: TAGS v6.1.9.1
- Created by: mhawksey. Read more about this at: <http://tags.hawksey.info>
- Instructions:
 - 1. If you've never run TAGS > Setup Twitter Access do so now (this should only need to be done once for all your TAGS sheets)
- Form fields:
 - Enter term:
- Note: Make a one off collection with TAGS > Run now! or set a trigger to collect every hour TAGS > Update archive every hour. To change the frequency open Tools > Script Editor then Triggers > Current script's triggers... and adjust
- Advanced Settings:
 - Period: default
 - Follower count filter: 0
 - Number of tweets: 3000
 - Type: search/tweets
- Stats:
 - Number of Tweets: 0
 - Unique tweets: 0
 - First Tweet: 30/12/1899 00:00:00
 - Last Tweet: 30/12/1899 00:00:00
- Make interactive:
 - Turn your archive into an interactive online resource using TAGExplorer - see <http://bit.ly/TAGSsetup>
 - Note: Share > Anyone with link to use these views
 - TAGExplorer (conversation explorer)
 - TAGS Archive (searchable archive)

STEP 4:

1. Once the form is fully completed, select again the tab 'TAGS' and 'Run Now!' to start the script scraping tweets on Twitter



The screenshot shows a Google Spreadsheet interface. The title bar reads 'Copie de TAGS v6.1.9'. The menu bar includes 'Fichier', 'Édition', 'Afficher', 'Insertion', 'Format', 'Données', 'Outils', 'Modules complémentaires', 'Aide', and 'TAGS'. The 'TAGS' menu is open, showing options: 'Setup Twitter Access', 'Disconnect Twitter Access', 'Delete Duplicate Data', 'Wipe Archive Sheet', 'Build archive from tweet IDs', 'Add Summary Sheet', 'Add Dashboard Sheet', 'Run Now!' (highlighted with a red circle), and 'Update archive every hour'. A red arrow points to the 'TAGS' menu.

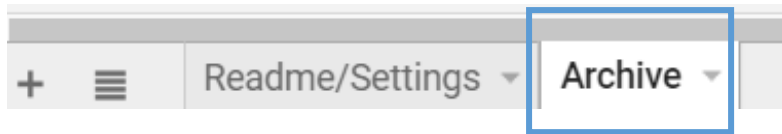
The spreadsheet content includes:

- Row 1: 'TAGS v6.1.9'
- Row 2: 'Created by mhawkey. Read more about this at: <http://tags.hawkey.info>'
- Row 4: 'With this spreadsheet you can:'
- Row 5: '- automatically pull results from a Twitter Search into a Google Spreadsheet'
- Row 6: 'Instructions:'
- Row 7: '1. If you've never run TAGS > Setup Twitter Access do so now (this should only need to be done once for all your accounts)'
- Row 9: '2. Enter term' with a text box containing 'bougnoule' and a note: '<- you can use search operators like AND OR as well as from: and to: eg '#JobsNow AND from:BarackObama' (without quotes)'
- Row 11: 'Note: Make a one off collection with TAGS > Run now! or set a trigger to collect every hour TAGS > Update archive every hour. To change the frequency open Tools > Script Editor then Triggers > Current script's triggers... and adjust'
- Row 12: 'Advanced Settings:'
- Row 13: 'Period' dropdown set to 'default'
- Row 15: 'Follower count filter' set to '0' with a note: '<- if search term is being spammed you can set the minimum followers a person must have to be included in archive'
- Row 16: 'Number of tweets' set to '3000' with a note: '<- maximum varies based on the type of archive you are collecting'
- Row 17: 'Type' dropdown set to 'search/tweets' with a note: '<- use a search term in step 3 above to get results from last 7 days'
- Row 18: 'Stats'
- Row 19: 'Number of Tweets' 1,147
- Row 20: 'Unique tweets' 1,141
- Row 21: 'First Tweet' 09/09/2018 13:35:05
- Row 22: 'Last Tweet' 19/09/2018 12:31:17
- Row 23: 'Make interactive'
- Row 24: 'Turn your archive into an interactive online resource using TAGSExplorer - see <http://bit.ly/TAGSsetup>'
- Row 25: 'Note: Share > Anyone with link to use these views'
- Row 26: 'TAGSExplorer' button with note: '<- conversation explorer'
- Row 27: 'TAGS Archive' button with note: '<- searchable archive'

The bottom of the spreadsheet shows a footer with '+', 'Readme/Settings', and 'Archive'.

STEP 5:

1. Turn on 'Archive' tab below the form



Create as many excel sheets as key words you would like to monitor.
You can download the excel sheets in order to start sorting all content (regarding several criteria: number of followers, of retweets, language used, etc.)

PROS AND CONS

CRITERIA	PROS	CONS
Language	It can work with any language since the crawling is based on a selection of key-words.	Some key words are written the same in different languages. There is no efficient language selection tool (you have to be more specific for some keywords to crawl relevant content).
Time needed for data collection	The tool crawls the content quickly.	When extracting the sheets, it is important to develop a methodology for sorting all data.
Level of human assistance	The tool offers analysis, additional information and filter options.	When extracting the sheets, it is important to develop a methodology for sorting all data.
Training + skills	The tool is easy to use.	It is important to define and develop a methodology for sorting all data.
Technical issues	N/A	N/A
Accuracy of data monitoring	<ul style="list-style-type: none">• Easy to use,• No downloading needed,• Relevant tools for key words,• Options available for developing a methodology for selecting relevant data.	<ul style="list-style-type: none">• Only available for Twitter.• Data stored on online shared files

HTTrack

HTTrack is a free software allowing users to mirror websites from the Internet to a local directory, building recursively all directories, getting HTML, images, and other files from the server to your computer for offline browsing and store the entire data on your computer. Keywords can be used for copy and paste the sub-domains.



STEP 1

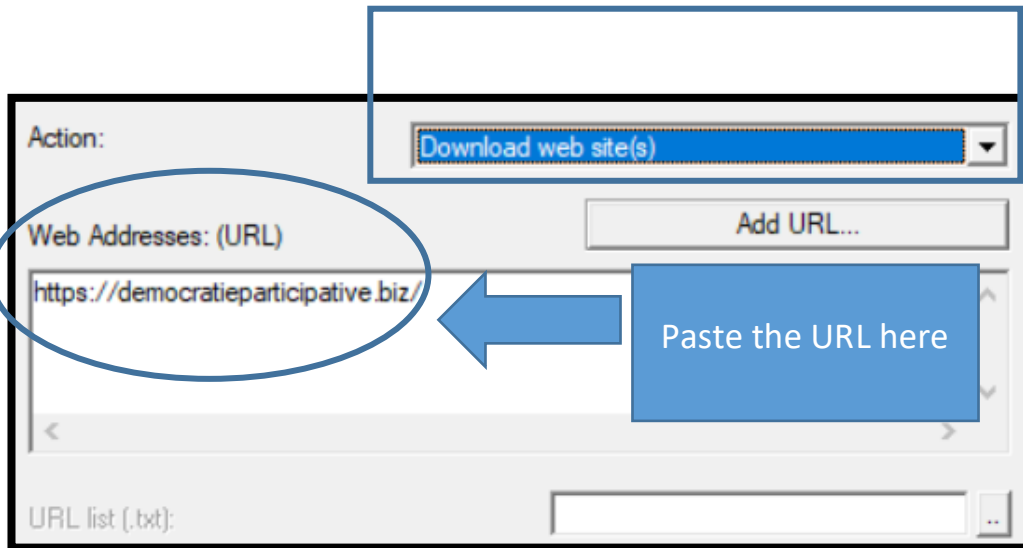
1. Install HTTrack by clicking on the [following link](#)
2. Launch it,
3. Create a project name
4. Choose the folder in which files will be recorded,
5. Press 'next'.

The screenshot shows the 'New project' dialog box in HTTrack. It has the following fields:

- New project name:** A text box containing 'sCAN - Testing Campaign 1'. A blue arrow points to this field.
- Project category:** A dropdown menu that is currently empty.
- Info:** A text area containing the text 'New project'.
- Base path:** A text box containing 'C:\My Web Sites'. A blue arrow points to this field.

STEP 2

1. Copy the URL (website selected) you want to scrape the content,
2. Paste the URL into the box as defined below,
3. Make sure “Download web site(s)” is selected.



STEP 3

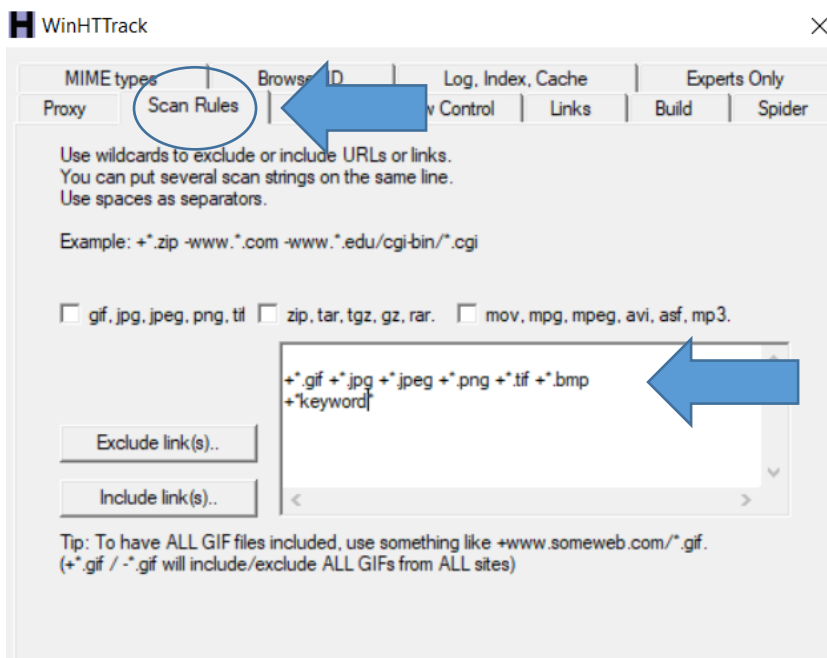
1. After you indicated which website has to be scanned, you will have the opportunity to sort and filter extracted content by format or keywords.
2. Press “Set options” below the URL’s box.

Preferences and mirror options:

Set options...

STEP 4

1. Select the tab "Scan rules",
2. Fill the box with which type of files will be extracted from the website (picture, text, or videos) and add the relevant keywords,
3. Press "Ok" to record settings,
4. Press "Next",
5. Press "Finish",
6. HTTrack will start the downloading process,
7. You can check the target folder in which you will see downloaded materials.

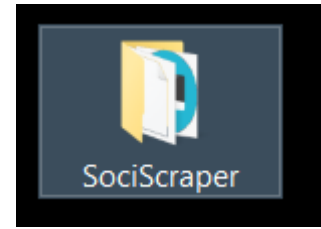


PROS AND CONS

CRITERIA	PROS	CONS
Language	It can work with any language since the program collects content of websites selected.	
Time needed for data collection	<p>The time of crawling depends on the websites selected (it depends on the sub-domains).</p> <p>The crawling is automatic: working on another task in the meantime is possible.</p>	It is crucial to first identify and target relevant websites as well as relevant keywords.
Level of human assistance	The level of human assistance depends on the quality and relevance of the keywords and the amount of data collected.	<p>For accessing the content that has been stored, it is necessary to click through numerous sub-folders and sort the content by hand.</p> <p>After the data collection, it is no longer possible to search or filter the content.</p> <p>The analysis is manual.</p>
Training + skills	The tool is easy to use.	The most important information for new users is how and where to set the keyword-option before data collection.
Technical issues		During the testing campaign, both partners have observed that some of the files could not be opened due to broken file names/unknown file extensions
Accuracy of data monitoring	<ul style="list-style-type: none"> • Crawl every data available on the site + possibility of selection of the types of files it should mirror (e.g. images) and keywords. • Every sub-domain or piece of content (such as images) is stored in a separate sub-folder • No need of internet access and data protected. • It is possible to keep track of changes since there is a copy of the website at a given point in time 	<ul style="list-style-type: none"> • Need to identify websites • Need to select very relevant keywords • Technically: structure of folders/subfolders is complex and time consuming.

SociScraper

It is a free software developed by individual users. It scrapes data from different social media platforms as for example Twitter, Instagram and YouTube with various search settings and sorting options which allow users to extract data by keywords: images, tweets, video URLs, number of followers, likes and views. Nonetheless, there is no context information provided.



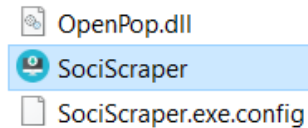
PREREQUISITES:

For using this tool, you will need:

1. To create/use an Instagram alias account
2. To use a proxy address (to know your proxy click on the [following link](#))

STEP 1:

1. Install sociScraper by clicking on the [following link](#)
2. Extract the .RAR folder from the download on your desktop: for extracting a RAR file, use the free and open source 7-Zip
3. Open the folder and click on 'SociScraper'.



STEP 2

1. In order to allow the software scraping content with your social media accounts, fill the box with your ID, password and proxy address which can be found out there: <http://www.whatismyproxy.com>

- When it's done, do not forget to click on 'Save' at the top of the box.

ON INSTAGRAM

1 Filters by accounts having specific keywords

2 Folder in which files will be downloaded

3 Choose 'Hashtag' to scrape by keywords

4 Define account criteria targeted

5 "Save results to": Folder in which files will be downloaded

ON YOUTUBE

1. Once the tab 'Youtube' is selected, keep 'Videos urls by keywords',
2. Fill the box with specific keywords,
3. Define criteria and limits on Youtube accounts,
4. Press 'Start' to explore Youtube.

The screenshot shows the SocioScaper web interface. At the top, there is a navigation bar with tabs for 'Account Mangement', 'Instagram', 'Pinterest', 'Twitter', 'Tumblr', 'GPlus', 'Youtube', and 'Resource & Tools'. The 'Youtube' tab is highlighted with a blue circle. Below the navigation bar is the 'Scraper' section. It features a dropdown menu set to 'Videos urls by keywords'. Below this is a text area labeled 'Enter keyword (one per line)' which is currently empty. To the right of the text area are two small icons: a green checkmark and a yellow warning sign. Below the text area are several configuration options: 'Threads Number' set to 10, a checkbox for 'Limit user scrape' which is unchecked, 'Limit user scrape' set to 1000, a checkbox for 'Ignore comment contains url' which is unchecked, and 'Min views' set to 100. At the bottom of the scraper section is a 'Save results to:' field which is empty. Below the scraper section, there is a 'Total: 0' indicator and two buttons: 'Start' and 'Abort'.

PROS AND CONS

CRITERIA	PROS	CONS
Language	It can work with any language since the crawling is based on a selection of keywords.	Some key words are written the same in different language. There is no efficient language selection tool (you must be more specific for some keywords to crawl relevant content).
Time needed for data collection	The tool crawls the content quickly. Several keywords can be added at the same time.	It is crucial to first identify and target relevant keywords.
Level of human assistance	The tool crawls relevant texts and images.	The output of SocioScaper does not offer a URL showing the source. Therefore, a high amount of time is needed to

		sort and analyse the output and identify the source (URL)
Training + skills	The tool is easy to use.	It is nonetheless important to know how to use the output: <ul style="list-style-type: none"> - Google reverse-image search to find the URLs of the images and the context in which they were posted, - Import texts in an Excel-Sheet and reverse-search content using a search engine or Twitters own search function.
Technical issues	N/A	When searching for more than one keywords at once, the tool stopped scraping at approx. 700 Tweets. When searching for only one keyword, it worked well and sometimes found over 10.000 Tweets per keyword.
Accuracy of data monitoring	<ul style="list-style-type: none"> • Easy to use, • Relevant tools for key words, • One tool for several platforms, • Use of several tools at once. 	<ul style="list-style-type: none"> • No links provided and no context information, • Need to use a reverse-search → not useful for tweets but for pictures, images posted in Instagram. • Post-analyze of the data by hand.

CrowdTangle

It is actually divided into two separate platforms, the app dashboard and the Chrome Extension, which complement each other extremely well. The Crowdtangle extension allows to search CrowdTangle's database for any content on Facebook, Twitter, Instagram and Reddit that have linked to that URL. The extension lets you see which Pages or accounts have actually shared the specific link in question.



PREREQUISITES:

For using this tool, you will need:

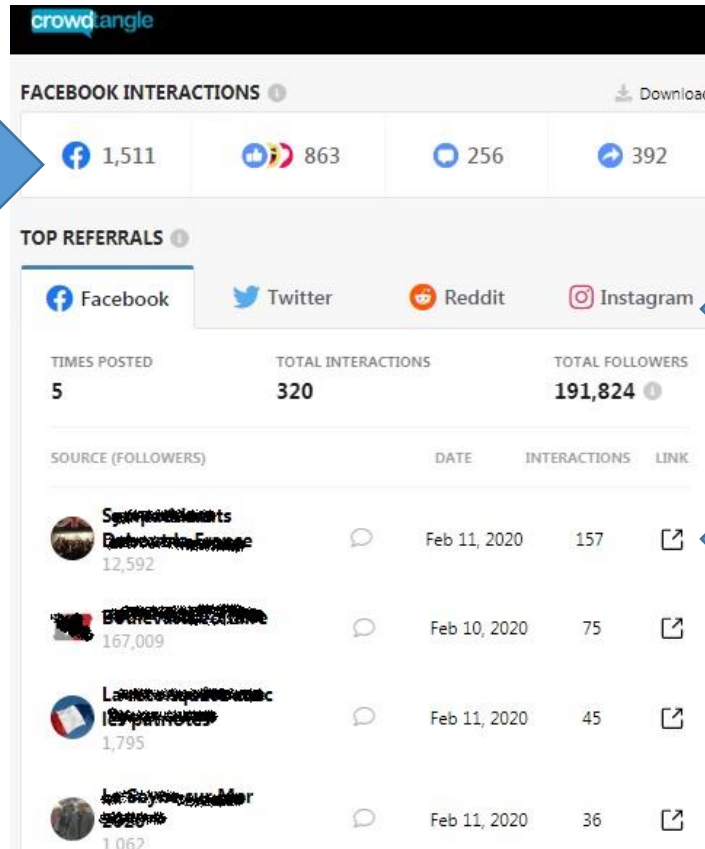
1. To create/use a Facebook alias account,
2. To create/use a Twitter alias account,
3. To create/use an Instagram alias account,
4. Since it is a Chrome extension, you will need Google Chrome.

STEP 1

1. Download the extension "CrowdTangle Link Checker",
2. Log into Facebook to identify yourself.

STEP 2

1. Select relevant URL in your chrome browser (e.g. blogs articles, tweets, websites, etc.)
2. Click on the CrowdTangle icon (upper-right).
3. CrowdTangle will then search its database for posts that link that URLs on Facebook, Twitter, Instagram and Reddit.



Number and type of interactions per platform

Click on each platform to get all information/data

Click on each link and you will access to each public page which sharing the information.

PROS AND CONS

CRITERIA	PROS	CONS
Language	It can work with any language since the extension is based on selected URL.	
Time needed for data collection	The tool targets the content shared instantly.	In order to keep track on your data collection, you will have to use your own database for storing all data.
Level of human assistance	The tool is very helpful for Facebook and Instagram since it helps to focus on public groups, pages and profiles sharing hateful content.	You will always need to monitor content on groups, pages and profiles found.
Training + skills	The tool is easy to use. No specific skills	It is important to have very relevant URLs for the extension (relevant list of websites and blogs).
Technical issues	N/A	Need to log in several times to the Facebook account

Accuracy of data monitoring

- Easy to use,
 - No downloading needed,
 - Relevant tool for nearly private platforms as Facebook, Reddit and Instagram.
- Use your own database system for collecting data,
 - Develop a list of relevant websites and blogs.
 - Only finds instances of links that are included in its own database – there might be more links on the monitored websites.

Algorithms and AI: SCAN and Factmata, a partnership for strengthening impact of automated tools in monitoring

Presentation

Factmata has developed a technology mixing artificial intelligence, algorithm and expert knowledge to deal with hate speech and fake news. Based in London, Factmata proposes an anti-fake-news AI platform and services by providing a scoring system for the content across the web, enhancing quality and credibility of textual content (e.g. articles, comments and user interactions) published through traditional mass media. The company works mainly with newspaper companies such as The New York Times, Bloomberg or The Guardian and “analyse[s] millions of new URLs daily, digging into individual articles and their actual content on a sentence by sentence level, to assess risk¹.”

Partnership and methodology

Regarding hate speech content, the Factmata API scores content to give “a deep understanding of the quality, safety and credibility of any piece of content on the web”. Criteria are: “insults”, “obscenity”, “toxicity”, “stereotyping”, and “threat”, “identity hate” as well as “sexism” and against “any particular gender”. As for social media, the start-up needs also human background back-stopping in order to improve AI detections and results.

jugendschutz.net, Licra and Factmata have launched their partnership in November 2018: these organisations are involved in the “Factmata Communities” (e.g. team of experts to help training the AI in a fair and accurate way).

During several testing sessions of the AI, jugendschutz.net and Licra have participated to annotation sessions regarding criteria of hate speech, threats, insults and obscenity. The tool has been trained in a North American context. Therefore, both European organisations had to provide a European experience regarding how to define hate speech: they contributed to integrate country-specific and European transnational criteria of hate speech trends.

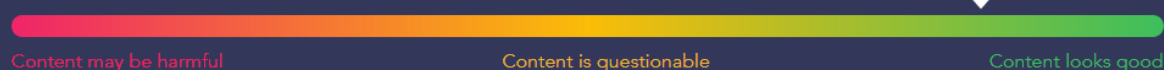
Several rounds of English content annotations have been organised in order to improve the model for detecting hateful content.

STEP 1

- 1) Click on: <https://try.factmata.com/>
- 2) Insert websites, blogs, articles links,
- 3) The review is ready and helps to identify the elements of hate speech.

¹ Website: <https://factmata.com/>

We're working hard to detect and reduce harmful online content. Please review these findings and give your feedback below to help our system continually improve.



This page does not appear to contain:

Clickbait	19% likely	Article headlines which at the expense of being informative, are designed to entice readers into clicking the accompanying link.
Political bias	16% likely	Strongly biased political language.
Toxicity	2% likely	Demeaning and abusive language in general.
Racism	1% likely	Demeaning and abusive language targeted towards a particular ethnicity, usually with stereotypes.
Obscenity	0% likely	Obscene or profane language.
Sexism	0% likely	Demeaning and abusive language targeted towards a particular gender, usually with stereotypes.
Insults	0% likely	Scornful remarks directed towards an individual.
Threats	0% likely	Expressing a wish or intention for pain, injury, or violence against an individual or group.

PROS AND CONS

CRITERIA	PROS	CONS
Language	Upcoming development(s): it would be possible to integrate new languages.	Only English content works for now.
Time needed for data collection	The tool rates the content instantly.	It is crucial to identify and select websites, blogs, etc.
Level of human assistance	The tool is very helpful for websites and blogs.	
Training + skills	The tool is easy to use. No specific skills requested.	
Technical issues	N/A	N/A
Accuracy of data monitoring	<ul style="list-style-type: none"> • Easy to use, • No downloading needed, • Relevant tool for web 1.0 	Some types of European context hatred are not detected (e.g. denial of Holocaust).

Conclusion

This user guide aims at focusing on available and inexpensive tools with a comprehensive interface in order to improve the monitoring process and collection of data. A practical file has been developed for each tool tested by jugendschutz.net and Licra. In addition, one of the main steps of this work was to determine the advantages and drawbacks of each tool selected. These results have been established regarding a common list of criteria (see Appendix 1: List of criteria).

This selection is of course non-exhaustive. It is also crucial to consider that new companies have developed new tools since the beginning of the SCAN project. Based on new testing campaigns, this selection will need to be updated on a regular basis in upcoming civil society projects in order to improve capacities in term of monitoring online hate speech.

Regarding the results of all the tools tested, it appears that human contribution in terms of knowledge and analysis is still crucial in monitoring online hate speech.

APPENDICES

Appendix 1: List of criteria

- Is this crawler/software compatible with the budget of each organisation?
- Does this crawler/software require a training to be used? If yes, what kind of training is required: guidelines? Tutorials? Coding? IT skills?
- To what extend does the tool require a manual/human assistance?
- Is the crawler/software relevant regarding the data collection?
- How long does it take to analyse and to sort each data collection compiled by each tool tested?
- Did a bug or an issue occur during the testing period? If yes, what kind of bug or issue?
- What are the pros and cons of each tool?
- Does the linguistic setting have an impact on the testing of each tool?

Appendix 2: A Hate Ontology for better understanding of key definitions

With this publication, the sCAN project partners, namely LICRA (France), Human Rights House Zagreb (Croatia), jugendschutz.net (Germany), CESIE (Italy), Zara – Zivilcourage und Anti-Rassismus-Arbeit (Austria), Spletno oko (Slovenia), ROMEA (Czech Republic), and the Latvian Centre for Human Rights (Latvia), bring their contribution to the further expansion of the body of knowledge and literature on online hate speech, to enable researchers, cyber activists, and the civil society representatives to better recognise and contrast the phenomenon of cyber hate.

<http://scan-project.eu/wp-content/uploads/scan-hate-ontology.pdf>