



MediaFutures

Deliverable 4.3

Data, tools and infrastructure for experimental support

**Pablo Aragón (Eurecat), Julian Vicens (Eurecat),
Elena Simperl (King's College London) and
Vicky Hallam (Open Data Insitute)**



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951962.

Disclaimer

This report was written as part of the MediaFutures project under EC grant agreement 951962. The information, documentation and figures available in this deliverable were written by the MediaFutures project consortium and do not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

How to quote this document

Aragón, P., Vicens, J., Simperl, E., Hallam, V. (2021). *Data, tools and infrastructure for experimental support. MediaFutures H2020 project*



This deliverable is licensed under a Creative Commons Attribution 4.0 International License

Table of Contents

Executive Summary	8
1 Introduction	9
2 Technical infrastructure	10
3 Datasets	11
3.1 COVID-19 Fact-checkers Dataset	12
3.2 The CoronaVirusFacts/DatosCoronaVirus Alliance Database	13
3.3 CoAID	14
3.4 COVID19FN	15
3.5 GDELT	16
3.6 Webhose's free datasets	17
3.7 COVID19 Infodemics Observatory	18
3.8 CMU-MisCov19	19
3.9 COVID-19-TweetIDs	20
3.10 Coronavirus (COVID-19) Tweets Dataset	21
3.11 Institutional and news media tweet dataset for COVID-19 social science research	22
3.12 COVID-19 Reddit Algo-Tracker	23
3.13 WMF COVID-19	24
3.14 Coronavirus en YouTube	24
3.15 CORD-19: The Covid-19 Open Research Dataset	25
3.16 COVID-19 Data Repository	26
3.17 Data on COVID-19	27
3.18 COVID-19 World Survey Data API	28
3.19 Data for the Open COVID-19 Data Working Group	29
3.20 CCCSL: CSH Covid-19 Control Strategies List	30
3.21 Health Intervention Tracking for COVID-19 (HIT-COVID) Data	31
3.22 Mozilla COVID dataset	32

3.23 The CoVidAffect dataset	32
3.24 COVID-19 Mobility Monitoring project	33
3.25 #Data4COVID19	34
4 Tools	34
4.1 Digital methods	35
4.1.1 DMI Instagram Scraper	35
4.1.2 Hyphe	36
4.1.3 YouTube Data Tools	36
4.1.4 Tumblr Tool	37
4.1.5 Hydrator	37
4.1.6 DMI TCAT	37
4.1.7 Reddit Tools	38
4.1.8 Telegram Tools	38
4.1.9 Open Refine	38
4.1.10 Palladio	39
4.1.11 Gephi	40
4.1.12 Voyant	41
4.1.13 RAWGraphs	42
4.2 Data Science	43
4.2.1 Scrappy	43
4.2.2 NLTK	43
4.2.3 Gensim	44
4.2.4 scikit-learn	45
4.2.5 NetworkX	46
4.2.6 Plotly	46
4.2.7 Matplotlib	46
4.3 Digital Art	47
4.3.1 Processing	47
4.3.2 openFrameworks	48

4.3.3 PureData	48
4.3.4 vvvv	49
4.3.5 Cinder	49
4.3.6 Magenta	50
4.3.7 D3.js	50
5 Training	51
5.1 Anonymisation is for everyone (ODI)	51
5.2 Applying Machine Learning and AI Techniques to Data (ODI)	51
5.3 Introduction to data ethics and the data ethics canvas (ODI)	52
5.4 Open data in a day (ODI)	52
5.5 Strategic Data Skills (ODI)	52
5.6 Datopolis (ODI)	52
5.7 Open data essentials (ODI)	52
5.8 Finding stories in data (ODI)	52
5.9 Annual event: ODI Summit (ODI)	52
5.10 Weekly lectures series: ODI Fridays (ODI)	52
5.11 Art-tech collaborations: best practices (IRCAM)	52
5.12 Public funding opportunities and KAILA tool (Zabala)	53
5.13 Social Innovation (Zabala)	53
5.14 Data journalism (LUISS)	53
6 Conclusions	53
7 References	54
8 Abbreviation List	57
9 More information about this document	57

List of Figures

Figure 1: Frontend of the infrastructure for experiment support in MediaFutures

Figure 2: COVID-19 Fact-checkers Dataset (Ryerson University Social Media Lab; The International Federation of Medical Students' Associations, 2020)

Figure 3: The CoronaVirusFacts/DatosCoronaVirus Alliance Database

Figure 4: Frequency of hashtags in tweets about fake and true news articles in CoAID (Cui & Lee, 2020)

Figure 5: COVID19FN

Figure 6: GDELT dataset

Figure 7: Webhose's free datasets

Figure 8: Dashboard of the data from the COVID19 Infodemics Observatory (Gallotti et al., 2020)

Figure 9: CMU-MisCov19 (Memon, Shahan & Carley, 2020)

Figure 10: Hashtags containing the substrings "wuhan", "covid" and "coronavirus" usage over time in COVID-19-TweetIDs (Chen., Lerman, & Ferrara, 2020)

Figure 11: Coronavirus (COVID-19) Tweets Dataset (Lamsal, 2020)

Figure 12: Institutional and news media tweet dataset for COVID-19 social science research (Yu, 2020).

Figure 13: COVID-19 Reddit Algo-Tracker

Figure 14: Edits per day in 263 Wikipedia projects

Figure 15: Evolution of the number of videos published about Covid-19 on YouTube in which Spain is mentioned between January and April 2020 (Orduña-Malea, Font-Julián, & Ontalba-Ruipérez, 2020)

Figure 16: CORD-19: The Covid-19 Open Research Dataset

Figure 17: Dashboard based on the COVID-19 Data Repository (Dong, Du & Gardner, 2020)

Figure 18: Web Explorer of the COVID-19 data maintained by Our World in Data (Hasell et al., 2020)

Figure 19: COVID-19 World Survey Data API (Barkay et al, 2020)

Figure 20: Dashboard for the data of the Open COVID-19 Data Working Group (Open COVID-19 Data Working Group, 2020)

Figure 21: Geographical coverage of the CCCSL and total number of recorded NPIs that were implemented in each country to control the spread of COVID-19 (Desvars-Larrive et al., 2020)

Figure 22: World map of sample PHSM (quarantine and isolation) implementation intensity since January 1, 2020 based on the Health Intervention Tracking for COVID-19 (HIT-COVID) Data

Figure 23: Firefox Desktop DAU deviation in France

Figure 24: Distribution of reported valence and arousal values In the CoVidAffect dataset (Bailon, 2020)

Figure 25: Scatterplot of the number of users of the COVID-19 Mobility Monitoring project assigned to each Italian province against the resident population reported by the Italian census in each province, as a fraction of the totals (Pepe, 2020)

Figure 26: #Data4COVID19

Figure 27: Hyphe, a curation-oriented approach to web crawling for the social sciences (Jacomy, Girard, Ooghe-Tabanou & Venturini, 2016)

Figure 28: RankFlow visualization of a query using YouTube Data Tools (Rieder., Matamoros-Fernández & Coromina, 2018)

Figure 29: Screenshots of Hydrator (Documenting the Now, 2020)

Figure 30: Digital Methods Initiative Twitter Capture and Analysis Toolset - DMI TCAT (Borra & Rieder, 2014)

Figure 31: Screenshot of Google Refine (Huynh, 2011)

Figure 32: Palladio, humanities thinking about data visualization (Ceserani, 2015)

Figure 33: Gephi, an open source software for exploring and manipulating networks (Bastian, Heymann, & Jacomy, 2009)

Figure 34: Screenshot of Voyant Tools

Figure 35: Charts provided by RAWGraphs, grouped by data model (Mauri, Elli, Caviglia, Ubaldi, & Azzi 2017)

Figure 36: Simple pipeline architecture for an information extraction system with NLTK (Bird, Klein & Loper, 2009)

Figure 37: Visualization of a topic model with Gensim (Rehurek & Sojka, 2010)

Figure 38: Examples of machine learning models built with Scikit-learn (Varoquaux et al., 2015)

Figure 39: Chart generated with NetworkX (Hagberg, Schult & Swart, 2012)

Figure 40: Chart generated with Matplotlib (Hunter, 2007)

Figure 41: Chart generated with Processing (Reas & Fry, 2015)

Figure 42: Chart generated with openFrameworks (Levin & Dorsey)

Figure 43: Patchable object generated with PureData (Puckette, 1997)

Figure 44: Hybrid development environment with vvvv (Bohnacker, Gross, Laub & Lazzeroni, 2012)

Figure 45: Images in Cinder <https://libcinder.org/docs/guides/cinder-images/index.html>

Figure 46: Plug-ins in Magenta Studio (Roberts et al., 2019)

Figure 47: Interactive visualizations built with D3.js (Bostock, Ogievetsky & Heer, 2011)

List of Tables

Table 1: Suggested datasets for addressing challenges of the MediaFutures 1st Open Call

Table 2: Platforms for digital methods

Table 3: Python libraries for data science

Executive Summary

MediaFutures will fund data-driven projects to conduct experiments that will require technical partners to make available a wide catalogue of open data resources and free open source technologies for data exploitation. To this end, we have deployed a technical infrastructure that allows the easy sharing of this type of resources. The infrastructure has been designed to engage participants with data from different sources (fact-checking and news, online social media, scientific articles, and repositories about statistics, interventions and behavioural traces) together with tools to collect, clean, analyse and visualize data powered by popular techniques from diverse research disciplines like social network analysis or natural language processing.

In addition to these technical capabilities, we also offer a range of training resources and courses, addressing data science and AI topics alongside more general topics relevant to the media value chain ecosystem, such as social innovation, entrepreneurship, and funding opportunities.

In this deliverable, we describe this infrastructure for experiment support that has been deployed on Github, a platform for collaborative software development. The catalogue is structured as follows:

- **Datasets** to help participants address the main challenges of the 1st Open Call. These datasets have been generated by institutions all over the world (universities, research centres, foundations and public institutions) that have granted open-access to allow data re-use, a core principle in MediaFutures¹.
- **Tools**, under free libre open source licenses, suggested by Mediafutures mentors² for startups and artists to collect new data, analyze and visualize datasets in order to create their projects on data, technology and arts. Tools are categorized as:
 - *Digital methods*: platforms for collecting, analysing and visualizing online data
 - *Data science*: Python packages for data science
 - *Digital art*: technologies for creative and artistic purposes
- **Training** resources and activities on technical and non-technical aspects.

While the resources of the catalogue presented in this deliverable strongly focus on the challenges of 1st Open Call, new needs and challenges are expected to emerge from participants. Therefore, the mentors of WP4 will provide support and update the infrastructure with additional datasets to be identified, new data analysis methods to make sense of data, and training activities about complementary topics.

¹ <https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information>

² https://ec.europa.eu/info/departments/informatics/open-source-software-strategy_en

1 Introduction

The team of mentors of WP4 in MediaFutures will offer two levels of support to startups and artists:

- Programme-wide support including an online catalogue of open data, software and other resources, training, and perks.
- Bespoke support to set up in-depth studies and experiments in human-centric AI, citizen participation, and human-data interaction to advance the field as a whole and provide case studies for the MediaFutures toolkit.

The general support consists of:

1. **An online catalogue** with curated openly available datasets, tools and training in data science and AI, described in this deliverable. This is meant as a living catalogue and we plan to update it using a mix of crowdsourcing and editorial work done in WP4.
 - a. For the first open call, WP4 has compiled the most extensive collections of data resources on misinformation and Covid-19, which are described in plain English to potential applicants.
 - b. To provide MediaFutures applicants with a diverse set of tools to collect new data, analyze and visualize datasets for their projects, we classified them into digital methods, data science and digital art.
 - c. A core set of training resources spanning from courses to eLearning to workshops. We will extend this with a list of external training resources endorsed by MediaFutures partners, and with a series of live webinars of specialised topics.
2. **Training delivery**
 - a. For each cohort, we plan to provide a training calendar of courses (online and otherwise), workshops and webinars delivered by the MediaFutures team, as well as other organisations we partner with.
 - b. This deliverable gives insight into the range of resources and activities explored, which will be expanded during the project.
3. **Perks**
 - a. As of January 2021, each participant will receive:
 - i. \$15,000 in credits for Amazon Web Services
 - ii. [HubSpot for Startups](#)
 - iii. [ODI membership](#)
 - b. As with everything else in this document, this list is just a snapshot of the support available at this moment and will be updated as the programme unfolds.

In addition to the programme-wide support, we will work together with startups and artists to provide bespoke support, according to a two-tiers approach that is mindful of the complexity of the issues addressed and of the resources available in WP4:

- **For startups and startup/artist collaborations:**
 - *Start phase*: analysis of support needs together with mentor and advisor to devise individual training plans.
 - *Build phase*: experiments support for a selection of candidates following an action research methodology. This methodology will allow us to apply the experimental

results to the ongoing projects in a collaborative way, following the main action research steps: identify concerns and research questions, collecting and analysing data, reporting results and taking informed actions based on evidence. The selection will very much depend on what the startup and startup/artist is planning to do. Standard data science and AI capabilities should be covered by new staff or contractors paid from the grant. More explorative, interdisciplinary challenges could take advantage of the unique expertise of KCL/EUT in the form of experiments and case studies, led by KCL/EUT researchers. We anticipate this would apply, purely by the nature of the work, more to startup/artist collaborations rather than startup pilots alone, though startups that plan extensive public engagement or participatory approaches, or those focusing on fairness, explainability and interpretability of tech methods would qualify as well.

- **For artists:**
 - *Build phase*: similar to 1b + the analysis of support needs at the start of the programme.

While the programme-wide support will be provided to all participants upon joining MediaFutures and advancing through the start-build-exhibit stages of our innovation framework, the bespoke support is expected to cover around 10 participants over the three cohorts. In the following we elaborate on the general support, including the infrastructure and training.

2 Technical infrastructure

The experiments to be carried out in the projects funded and supported in MediaFutures will require a wide catalogue of open data resources and free libre open source technologies for data exploitation. To this end, a technological infrastructure has been deployed that will allow to share these data assets with project participants. The infrastructure is hosted on a Github account³, the main Internet platform for collaborative software development.

First, we have forked repositories by third parties corresponding to technological tools⁴ recommended by the team of mentors from WP4 (see section 4). Mentors will collaborate with participants drawing on extensive experience of applied research methods at KCL and EUT, both in socio-technical areas such as digital humanities, human-AI interaction as well as traditional machine learning and AI experimentation and evaluation. Second, a new repository has been created to develop a [GitHub page](#) that will serve as a frontend for the infrastructure for experiment support in MediaFutures (see Figure 1). The page is structured in three main sections: datasets, tools, and training. In each one of them, we have added a list of resources of interest that we have mapped for the 1st Open call and included a short description, access links and metadata of interest. All this information is detailed in the following sections.

³ <https://github.com/mediafutureseu>

⁴ <https://github.com/mediafutureseu?tab=repositories>



Figure 1: Frontend of the infrastructure for experiment support in MediaFutures.

3 Datasets

The infrastructure for experimental support contains a catalogue of datasets with different open data assets that might be of interest for addressing the challenges of MediaFutures open calls. In this deliverable, we specify the datasets of the 1st Open Call about coronavirus and misinformation. As shown in Table 1, we have identified datasets about COVID-19 from fact-checking and news, online social media (Twitter, Reddit, Wikipedia and Youtube), scientific articles, and repositories about statistics, interventions and behavioural traces. We recall that participants are free to use these or other datasets.

Table 1: Suggested datasets for addressing challenges of the MediaFutures 1st Open Call.

Category	Dataset
Fact-checking	COVID-19 Fact-checkers Dataset
	The CoronaVirusFacts/DatosCoronaVirus Alliance Database
News	CoAID
	COVID19FN
	GDELT
	Webhose's free datasets

Twitter	COVID19 Infodemics Observatory
	CMU-MisCov19
	COVID-19-TweetIDs
	Coronavirus (COVID-19) Tweets Dataset
	Institutional and news media tweet dataset for COVID-19 social science research
Reddit	COVID-19 Reddit Algo-Tracker
Wikipedia	WMF COVID-19
YouTube	Coronavirus en YouTube
Scientific articles	CORD-19: The Covid-19 Open Research Dataset
Statistics	COVID-19 Data Repository
	Data on COVID-19
	COVID-19 World Survey Data API
	Data for the Open COVID-19 Data Working Group
Interventions	CCCSL: CSH Covid-19 Control Strategies List
	Health Intervention Tracking for COVID-19 (HIT-COVID) Data
Behavioural traces	Mozilla COVID dataset
	The CoVidAffect dataset
	COVID-19 Mobility Monitoring project
Miscellanea	#Data4COVID19

We list them below including the official description, the organization that provided them and a link to the resource.

3.1 COVID-19 Fact-checkers Dataset

- **Provider:** Social Media Lab - Ryerson University
- **Link:** <https://doi.org/10.5683/SP2/IMISPE>
- **Description:** The COVID-19 Fact Checkers Dataset is a comprehensive list of over 200 active fact-checking organizations and groups that verify COVID-19 misinformation to study the proliferation of COVID-19 misinformation and to map fact-checking activities around the world in partnership with the World Health Organization (WHO). It was created to provide

the public with a better understanding of the COVID-19 fact-checking ecosystem and is intended to be used by policymakers and others to make evidence-based decisions in combating COVID-19 misinformation.

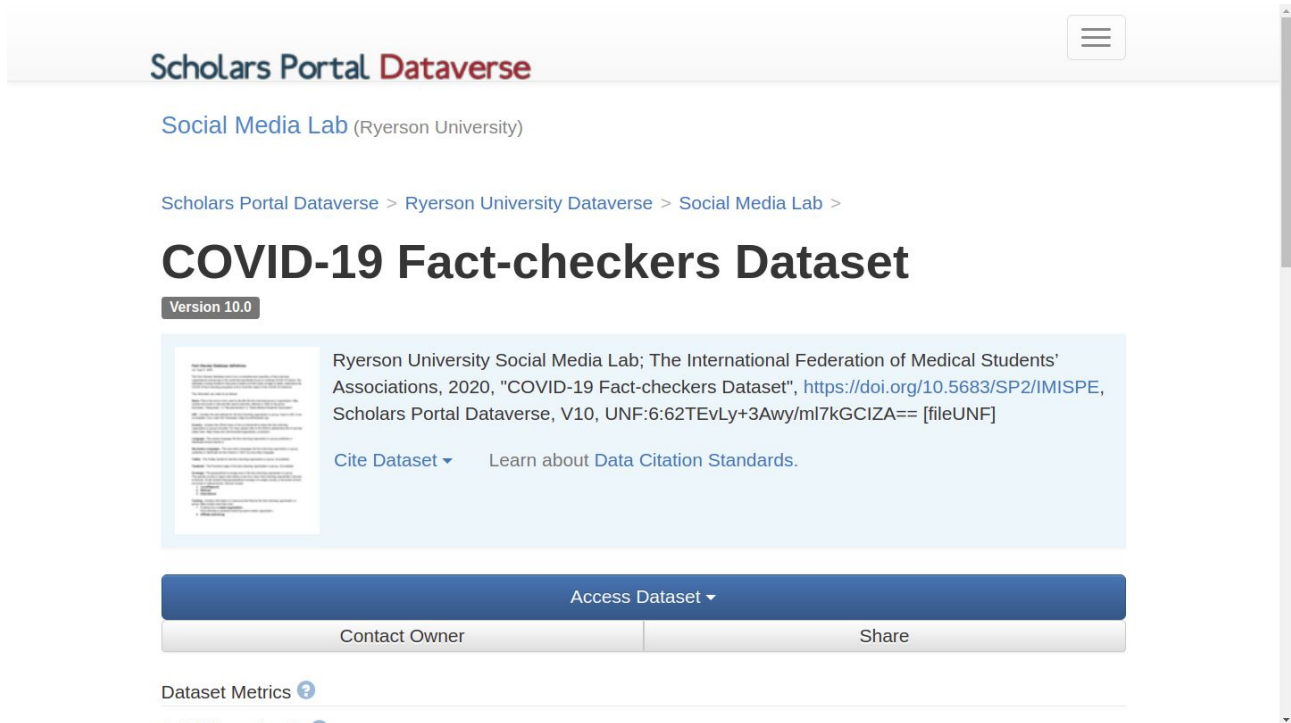


Figure 2: COVID-19 Fact-checkers Dataset (Ryerson University Social Media Lab; The International Federation of Medical Students' Associations, 2020).

3.2 The CoronaVirusFacts/DatosCoronaVirus Alliance Database

- **Provider:** Poynter Institute
- **Link:** <https://www.poynter.org/ifcn-covid-19-misinformation/>
- **Description:** Database that gathers all of the falsehoods that have been detected by the CoronaVirusFacts/DatosCoronaVirus alliance. This database unites fact-checkers in more than 70 countries and includes articles published in at least 40 languages.

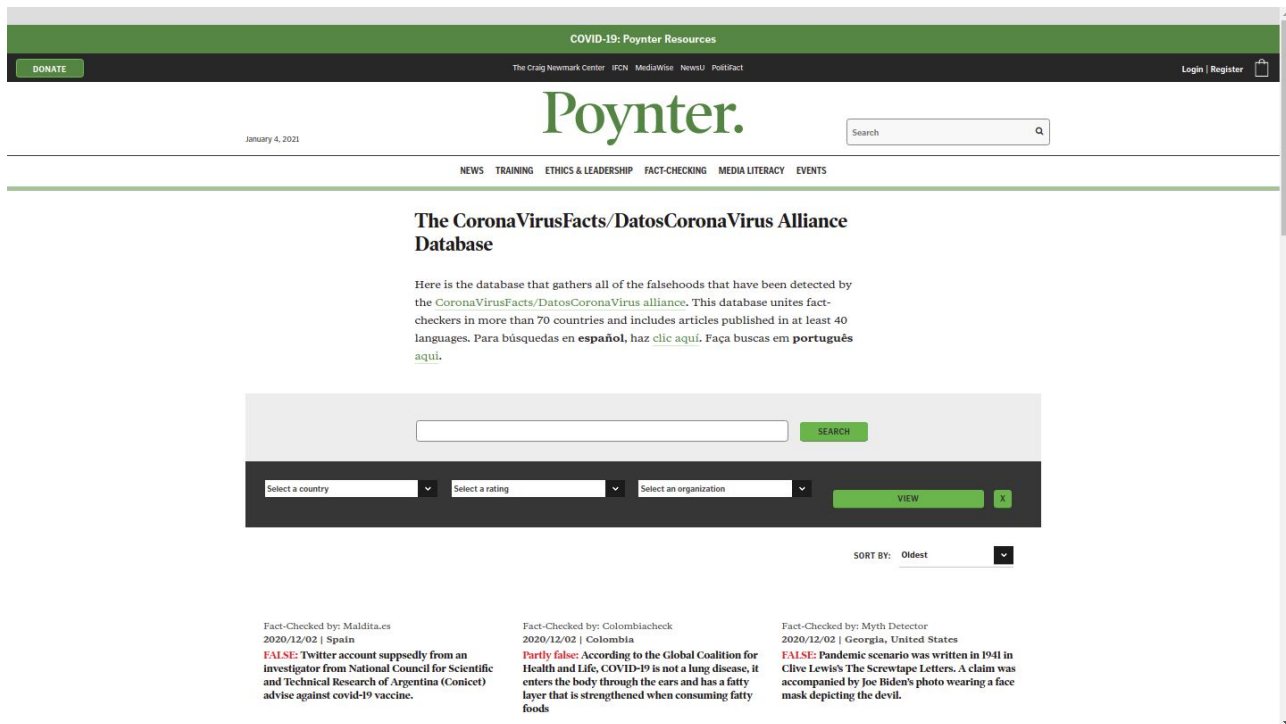


Figure 3: The CoronaVirusFacts/DatosCoronaVirus Alliance Database.

3.3 CoAID

- **Provider:** The Pennsylvania State University
- **Link:** <https://github.com/cuilimeng/CoAID>
- **Description:** Diverse COVID-19 healthcare misinformation dataset, including fake news on websites and social platforms, along with users' social engagement about such news. It contains 4,251 news, 296,000 related user engagements, 926 social platform posts about COVID-19, and ground truth labels.

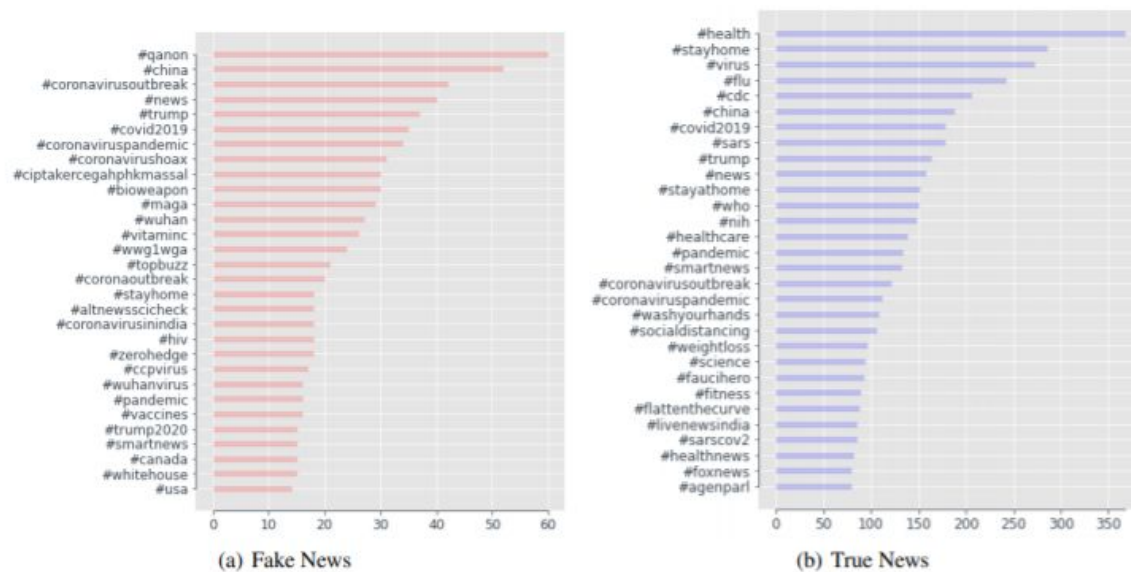


Figure 4: Frequency of hashtags in tweets about fake and true news articles in CoAID (Cui & Lee, 2020).

3.4 COVID19FN

- **Provider:** Sardar Vallabhbhai National Institute of Technology
- **Link:** <http://dx.doi.org/10.17632/b96v5hmfv6.3>
- **Description:** Dataset that comprises labelled news articles of misinformation spread during the Infodemic. It contains approximately 2800 news articles, real and fake, scraped from Poynter and other fact-checking sites. It also contains information such as source URL, publish date and origin country of the news article. Potential applications of this dataset would be to explore various research areas such as classification of intention, study of spatial and temporal features, and linguistic indications that can provide further insight and help mitigate its effect as much as possible.

[Back to dataset](#)

	Title
0	A video shows a fortune teller predicting the coronavirus pandemic in December on Spanish TV.
1	Internet sensation and the world's cutest baby Anahita Hashemzadeh is suffering COVID-19.
2	A video has been viewed hundreds of thousands of times on Facebook and YouTube in March and April 2020 alongside a claim it shows a Koran recording the end of the world.
3	Treasury is depositing Kshs 45, 000 to the mobile wallet of Nairobi residents in Kenya.
4	Hungarian authorities are capturing men 50 or over to make them spend quarantine in government facilities.
5	A Facebook user claims Melinda Gates divorced her husband, Bill Gates, for wanting to destroy Africa.
6	News photo from stay-at-home protest was doctored to add Confederate flag.
7	A video has been viewed tens of thousands of times on Twitter, Facebook and YouTube alongside a claim that it shows the US Federal Bureau of Investigation.
8	Claim that the Washington Post confirmed that coronavirus patient zero was a worker from the China Laboratory.
9	The US \$1,200 coronavirus relief checks this year are "just an advance on your next tax return." Next year, you're automatically going to owe the IRS.
10	The field hospital built in a football stadium in S�o Paulo (Brazil) to receive COVID-19 patients is empty.
11	Ecuadorians dump their dead by COVID-19 into the sea.
12	A photograph shared thousands of times on Facebook purports to show the blister-covered hand of a patient suffering from a new disease.
13	Coronavirus in poultry products has been confirmed by the Bihar Health Department.
14	Multiple Facebook posts claim an anti-viral injection that was being developed in the Philippines in April 2020 is a cure for COVID-19.

< 1 OF 1 FILES >

>

FILE INFORMATION

COVID19FN.csv

File extension	csv
File size	15 MB
Uploaded	13-06-2020
License	CC BY 4.0

[Download file](#)

[Cite this file](#)

Figure 5: COVID19FN.

3.5 GDELT

- **Provider:** Google Jigsaw
- **Link:** <https://www.gdeltproject.org>
- **Description:** The GDELT Project monitors the world broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day, creating a free open platform for computing on the entire world.

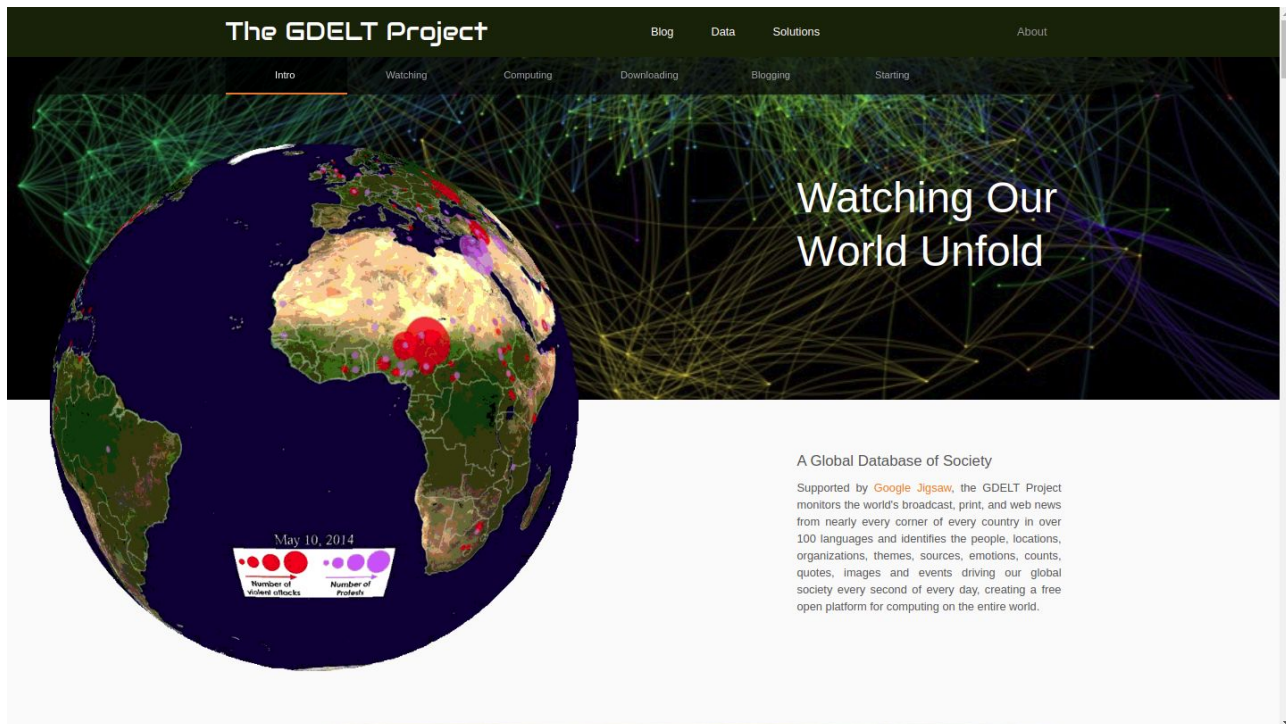
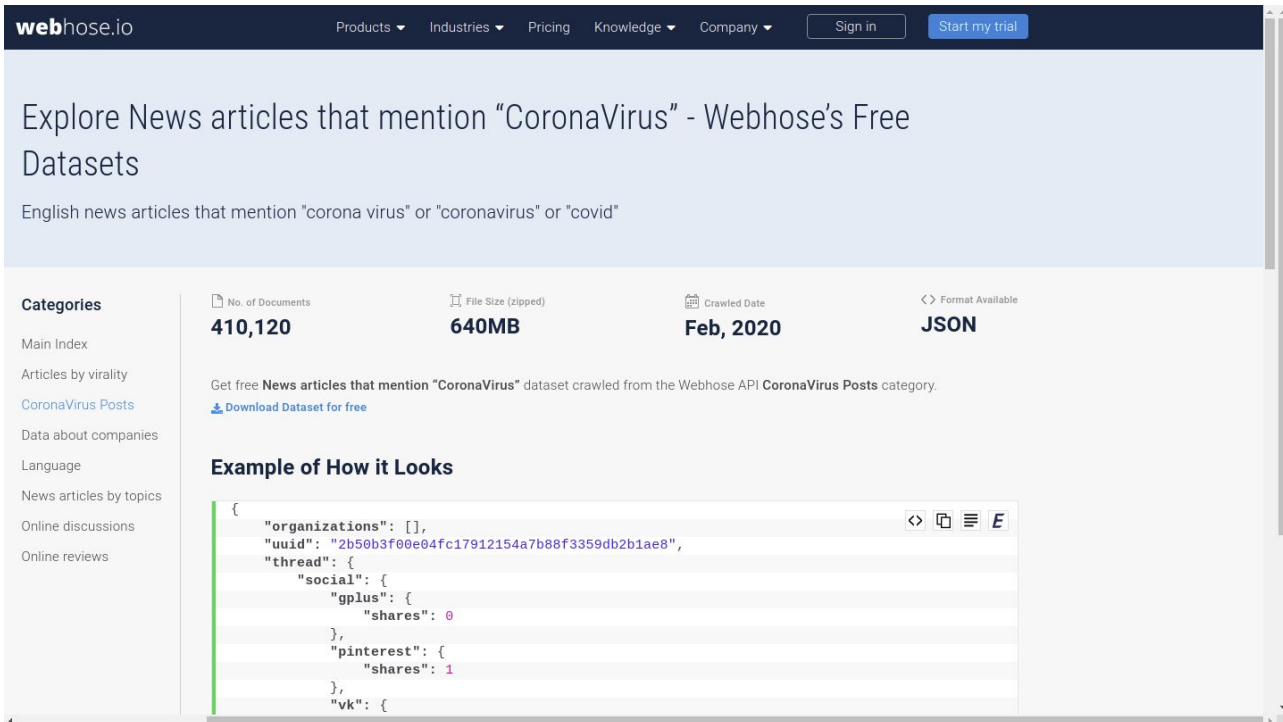


Figure 6: GDELT dataset.

3.6 Webhose's free datasets

- **Provider:** Webhose
- **Link:** <https://webhose.io/free-datasets>
- **Description:** News articles, blog posts and online discussions that mention "CoronaVirus" crawled on February 2020.



webhose.io Products Industries Pricing Knowledge Company Sign in Start my trial

Explore News articles that mention "CoronaVirus" - Webhose's Free Datasets

English news articles that mention "corona virus" or "coronavirus" or "covid"

Categories	No. of Documents	File Size (zipped)	Crawled Date	Format Available
Main Index	410,120	640MB	Feb, 2020	JSON

Get free News articles that mention "CoronaVirus" dataset crawled from the Webhose API CoronaVirus Posts category.

[Download Dataset for free](#)

Example of How it Looks

```
{
  "organizations": [],
  "uuid": "2b59b3f0e04fc17912154a7b88f3359db2b1ae8",
  "thread": {
    "social": {
      "gplus": {
        "shares": 0
      },
      "pinterest": {
        "shares": 1
      },
      "vk": {
```

Figure 7: Webhose's free datasets.

3.7 COVID19 Infodemics Observatory

- **Provider:** CoMuNe Lab - Fondazione Bruno Kessler
- **Link:** <https://osf.io/n6upx>
- **Description:** Results from the analysis of infodemics due to unreliable content in online social media. Specifically, public posts on Twitter, analyzed with state-of-the-art machine learning techniques for: (1) population emotional state; (2) bot/human classification; (3) news reliability.

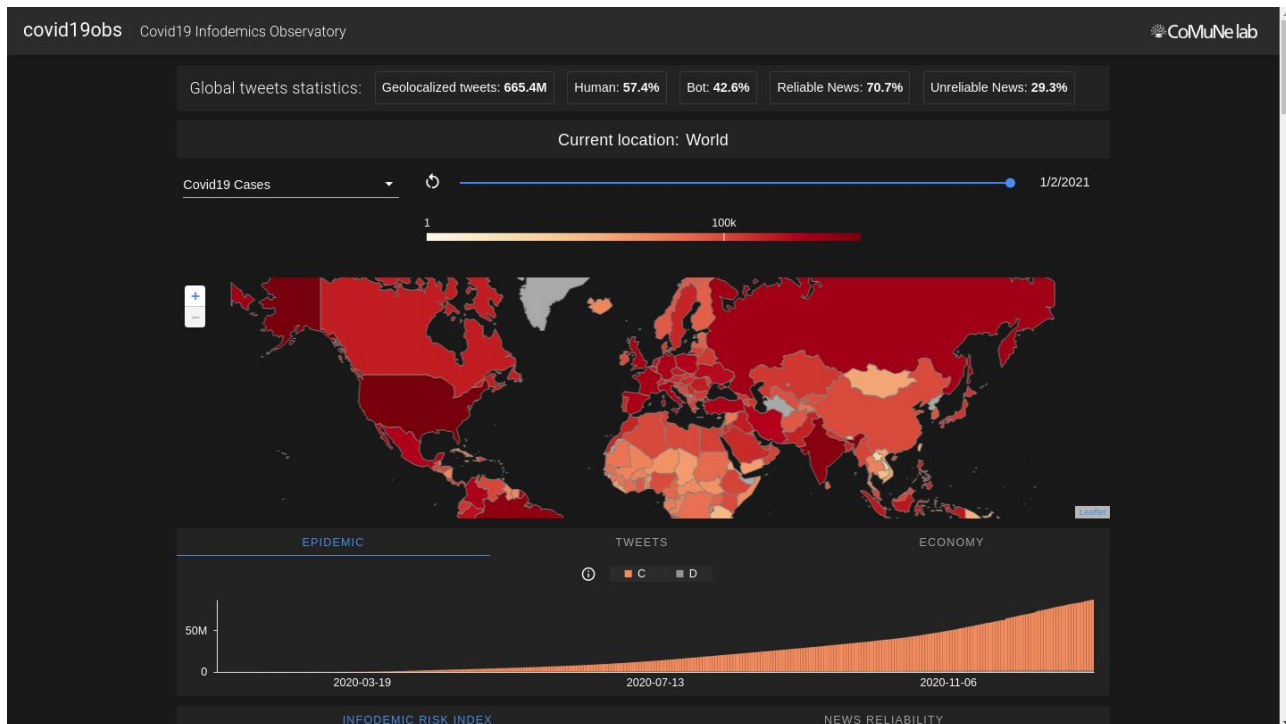


Figure 8: Dashboard of the data from the COVID19 Infodemics Observatory (Gallotti et al., 2020).

3.8 CMU-MisCov19

- **Provider:** Carnegie Mellon University
- **Link:** <https://zenodo.org/record/4024154#.YAr4EOhKhPZ>
- **Description:** Twitter misinformation dataset called "CMU-MisCov19" with 4573 annotated tweets over 17 themes around the COVID-19 discourse. It also includes an annotation codebook for the different COVID-19 themes on Twitter, along with their descriptions and examples, for the community to use for collecting further annotations. Further details related to the dataset, and our analysis based on this dataset can be found at Memon & Carley (2020b). In adherence to Twitter's terms and conditions, full tweet JSONs are not provided but a ".csv" file with the tweet IDs so that the tweets can be rehydrated. The dataset also provides the annotations, and the date of creation for each tweet for the reproduction of the results of our analyses.



The screenshot shows the Zenodo dataset page for "CMU-MisCov19: A Novel Twitter Dataset for Characterizing COVID-19 Misinformation". The page includes a search bar, login/sign-up buttons, and a header with the date "September 19, 2020". The dataset is labeled as "Dataset" and "Open Access". It has 1,222 views and 344 downloads. The dataset is indexed in OpenAIRE. The publication date is September 19, 2020, and the DOI is 10.5281/zenodo.4024154. The keywords are covid, coronavirus, misinformation, twitter, covid-19, network analysis, sociolinguistics, and dataset. The meeting is the 5th International Workshop on Mining Actionable Insights from Social Networks (MAISoN) at CIKM 2020, Online, 19-20 October 2020.

September 19, 2020

CMU-MisCov19: A Novel Twitter Dataset for Characterizing COVID-19 Misinformation

Memon, Shahan Ali; Carley, Kathleen M.

From conspiracy theories to fake cures and fake treatments, COVID-19 has become a hot-bed for the spread of misinformation online. It is more important than ever to identify methods to debunk and correct false information online. Detection and characterization of misinformation requires an availability of annotated datasets. Most of the published COVID-19 Twitter datasets are generic, lack annotations or labels, employ automated annotations using transfer learning or semi-supervised methods, or are not specifically designed for misinformation. Annotated datasets are either only focused on "fake news", are small in size, or have less diversity in terms of classes.

Here, we present a novel Twitter misinformation dataset called "CMU-MisCov19" with 4573 annotated tweets over 17 themes around the COVID-19 discourse. We also present our annotation codebook for the different COVID-19 themes on Twitter, along with their descriptions and examples, for the community to use for collecting further annotations. Further details related to the dataset, and our analysis based on this dataset can be found at <https://arxiv.org/abs/2008.00791>. In adherence to the Twitter's terms and conditions, we do not provide the full tweet JSONs but provide a ".csv" file with the tweet IDs so that the tweets can be rehydrated. We also provide the annotations, and the date of creation for each tweet for the reproduction of the results of our analyses.

Note: If for any reason, you are not able to rehydrate all the tweets, reach out to Shahan Ali Memon at (shahan@nyu.edu).

If you use this data, please cite our paper as follows:

"Shahan Ali Memon and Kathleen M. Carley. Characterizing COVID-19 Misinformation Communities Using a Novel Twitter Dataset, In Proceedings of The 5th International Workshop on Mining Actionable Insights from Social Networks (MAISoN 2020), co-located with CIKM, virtual event due to COVID-19, 2020."

If you use this dataset, please cite our recently accepted paper on "Characterizing COVID-19 Misinformation

1,222 views

344 downloads

See more details...

Indexed in

OpenAIRE

Publication date: September 19, 2020

DOI: 10.5281/zenodo.4024154

Keyword(s): covid, coronavirus, misinformation, twitter, covid-19, network analysis, sociolinguistics, dataset

Meeting: 5th International Workshop on Mining Actionable Insights from Social Networks (MAISoN) at CIKM 2020, Online, 19-20 October 2020

Figure 9: CMU-MisCov19 (Memon & Carley, 2020).

3.9 COVID-19-TweetIDs

- **Provider:** University of Southern California
- **Link:** <https://github.com/echen102/COVID-19-TweetIDs>
- **Description:** Ongoing collection of tweets IDs associated with the novel coronavirus COVID-19 (SARS-CoV-2), which commenced on January 28, 2020. The Twitter's search API was used to gather historical Tweets from the preceding 7 days, leading to the first Tweets in our dataset dating back to January 21, 2020. Twitter's streaming API was leveraged to follow specified accounts and also collect in real-time tweets that mention specific keywords. To comply with Twitter's Terms of Service, only the Tweet IDs of the collected Tweets are publicly released. The data is released for non-commercial research use.

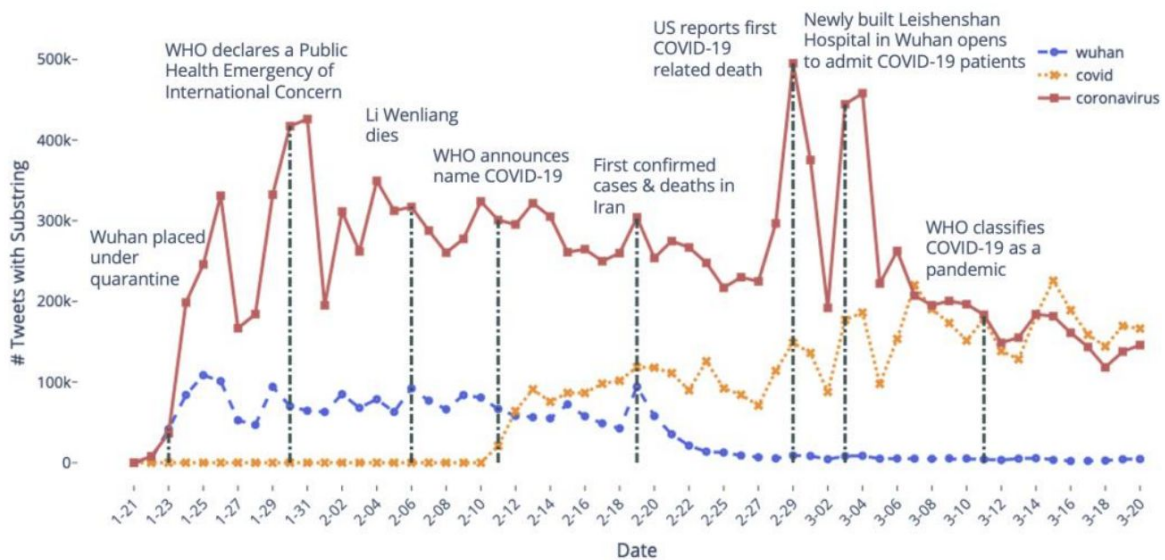


Figure 10: Hashtags containing the substrings “wuhan”, “covid” and “coronavirus” usage over time in COVID-19-TweetIDs (Chen., Lerman, & Ferrara, 2020).

3.10 Coronavirus (COVID-19) Tweets Dataset

- **Provider:** Jawaharlal Nehru University
- **Link:** <https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset#files>
- **Description:** This dataset includes CSV files that contain IDs and sentiment scores of the tweets related to the COVID-19 pandemic. The tweets have been collected by an on-going project deployed at <https://live.rlamsal.com.np>. The model monitors the real-time Twitter feed for coronavirus-related tweets using 90+ different keywords and hashtags that are commonly used while referencing the pandemic. This dataset has been wholly re-designed on March 20, 2020, to comply with the content redistribution policy set by Twitter.



IEEE.org | IEEE Xplore Digital Library | IEEE Standards | IEEE Spectrum | More Sites

IEEE DataPort™ DATASETS COMPETITIONS SUBMIT A DATASET Q SEARCH IEEE

Datasets

Open Access

@CORONAVIRUS (COVID-19) TWEETS DATASET

coronavirus tweets dataset.

Citation Author(s): Rabindra Lamsal (School of Computer and Systems Sciences, JNU)

Submitted by: Rabindra Lamsal

Last updated: Mon, 01/04/2021 - 01:21

DOI: 10.21227/781w-ef42

Data Format: CSV

Links: Live Twitter Sentiment
Author's Homepage

License: Creative Commons Attribution

92156 Views

Categories: COVID-19
Machine Learning

Keywords: Corona Tweets Dataset, COVID-19 Tweets Dataset, Corona Tweets, COVID-19 Tweets, Corona Twitter Sentiment, COVID-19 Twitter Sentiment, SARS-CoV-2 Tweets Dataset, SARS-CoV-2 Twitter Sentiment, Coronavirus English Tweets Dataset, COVID-19 English Tweets Dataset

16 ratings - Please login to submit your rating.

ACCESS DATASET CITE SHARE/EMBED

ABSTRACT DATASET FILES

Figure 11: Coronavirus (COVID-19) Tweets Dataset (Lamsal, 2020).

3.11 Institutional and news media tweet dataset for COVID-19 social science research

- **Provider:** Universitat Autònoma de Barcelona
- **Link:** <https://github.com/narcisoyu/>
- **Description:** Open access data repository for institutional/news media tweet dataset in the time of COVID-19 pandemic

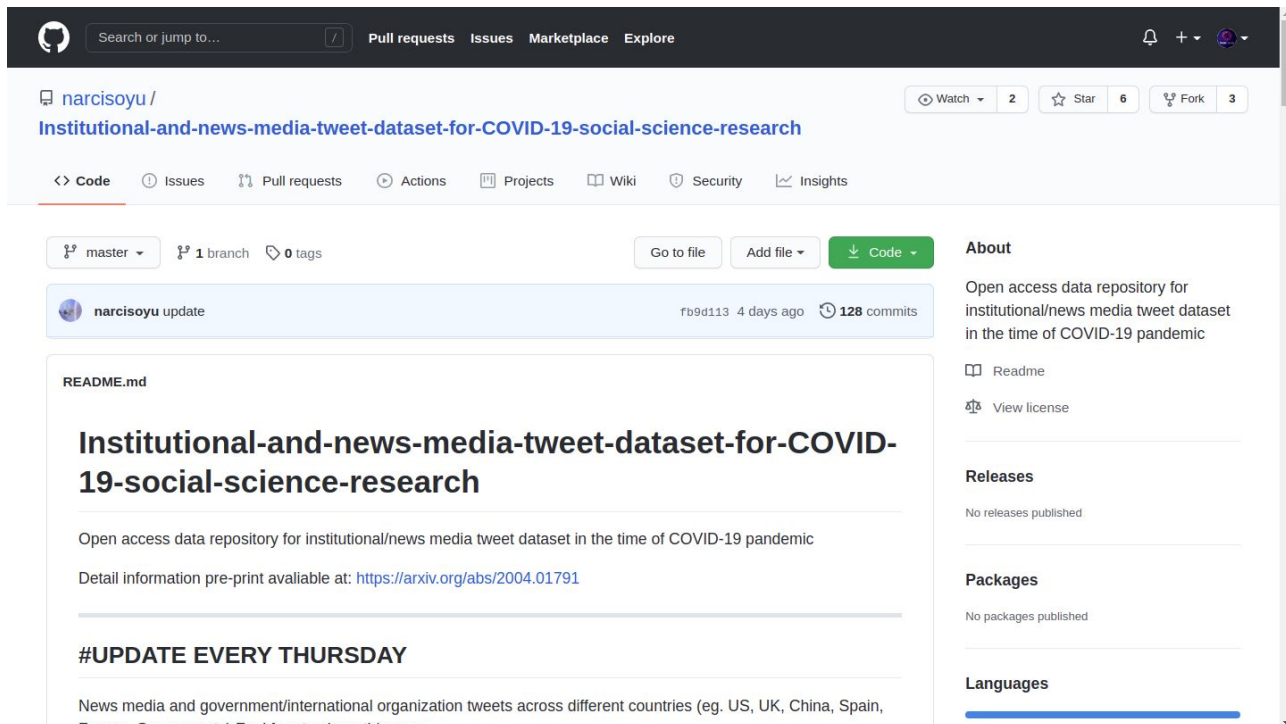


Figure 12: Institutional and news media tweet dataset for COVID-19 social science research (Yu, 2020).

3.12 COVID-19 Reddit Algo-Tracker

- **Provider:** Cornell University
- **Link:** <https://github.com/natematias/covid-algotracker>
- **Description:** COVID-19 content being promoted by reddit algorithms

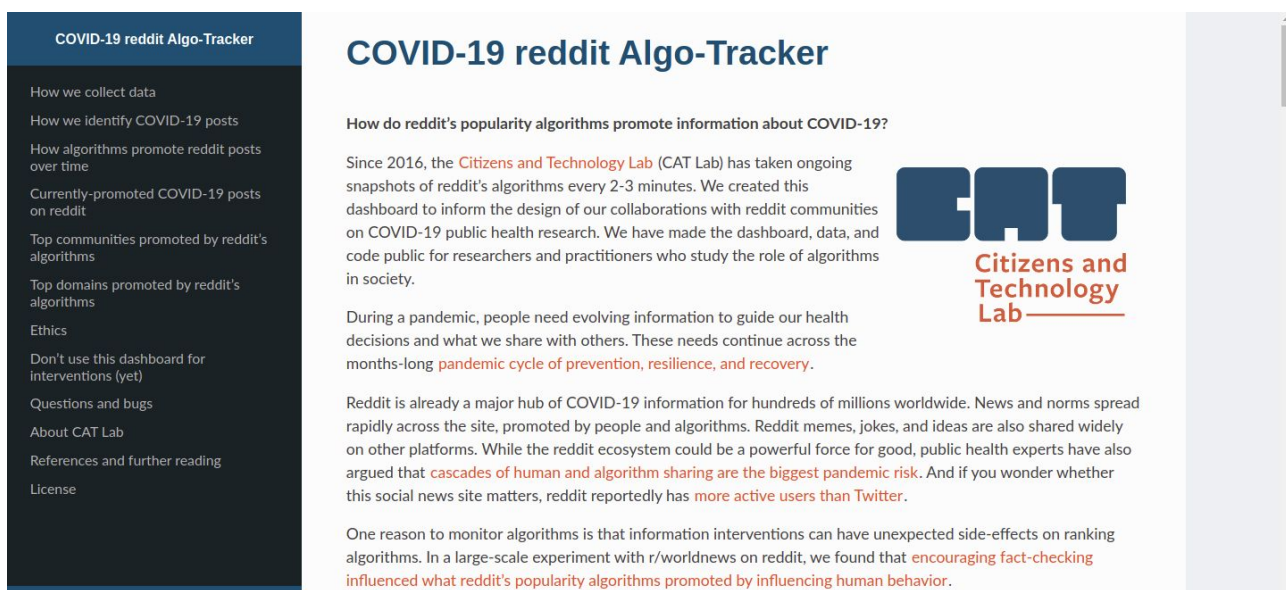


Figure 13: COVID-19 Reddit Algo-Tracker.

3.13 WMF COVID-19

- **Provider:** Wikimedia Foundation
- **Link:** <https://covid-data.wmflabs.org>
- **Description:** COVID-19 related content across Wikipedia projects

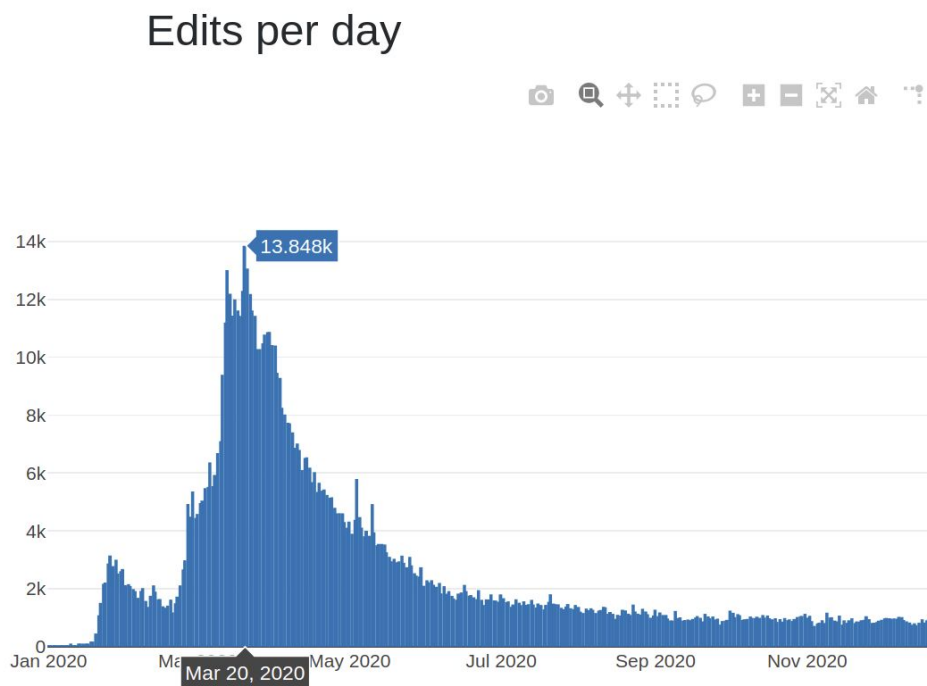


Figure 14: Edits per day in 263 Wikipedia projects.

3.14 Coronavirus en YouTube

- **Provider:** Universitat Politècnica de València
- **Link:** <https://doi.org/10.4995/Dataset/10251/143671>
- **Description:** This dataset contains, on the one hand, the initial sample of 73,268 videos recovered from YouTube in response to specific queries related to covid-19 and Spain and, on the other hand, the final sample of 39-702 videos in which the term coronavirus, covid-19 or SARS-CoV-2 appears explicitly in the title or description of the videos.

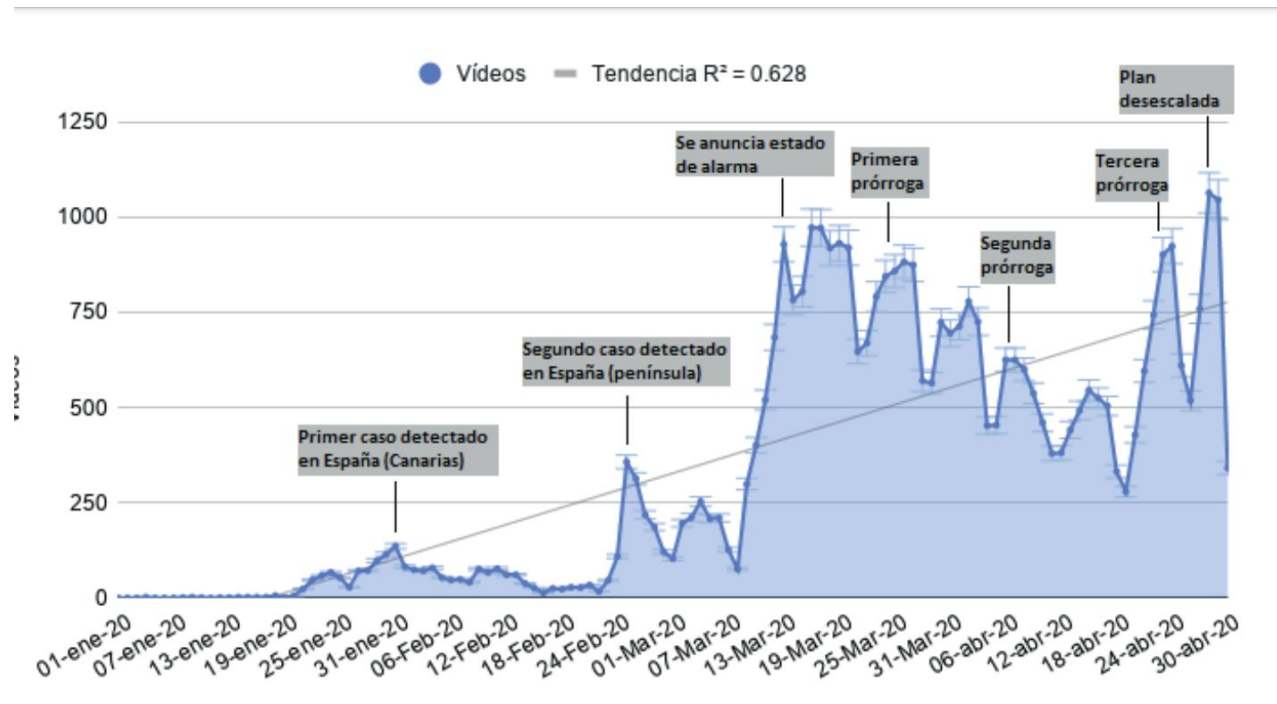
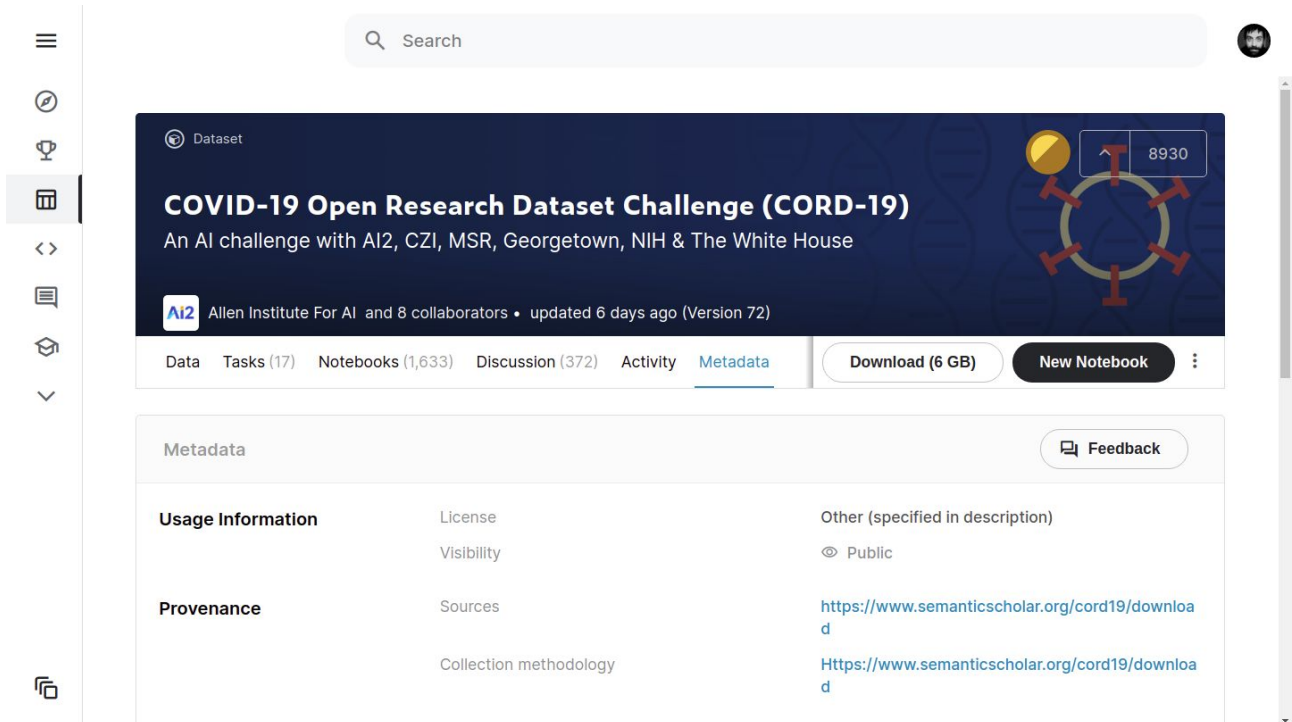


Figure 15: Evolution of the number of videos published about Covid-19 on YouTube in which Spain is mentioned between January and April 2020 (Orduña-Malea, Font-Julián, & Ontalba-Ruipérez, 2020).

3.15 CORD-19: The Covid-19 Open Research Dataset

- **Provider:** Allen Institute for AI
- **Link:** <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>
- **Description:** CORD-19 is a resource of over 200,000 scholarly articles, including over 100,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in new coronavirus literature, making it difficult for the medical research community to keep up.



Dataset

COVID-19 Open Research Dataset Challenge (CORD-19)
An AI challenge with AI2, CZI, MSR, Georgetown, NIH & The White House

AI2 Allen Institute For AI and 8 collaborators • updated 6 days ago (Version 72)

Data Tasks (17) Notebooks (1,633) Discussion (372) Activity Metadata Download (6 GB) New Notebook

Metadata

Usage Information	License	Other (specified in description)
	Visibility	Public
Provenance	Sources	https://www.semanticscholar.org/cord19/download
	Collection methodology	https://www.semanticscholar.org/cord19/download

Feedback

Figure 16: CORD-19: The Covid-19 Open Research Dataset.

3.16 COVID-19 Data Repository

- **Provider:** Johns Hopkins University
- **Link:** https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data
- **Description:** Data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Also, Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL).

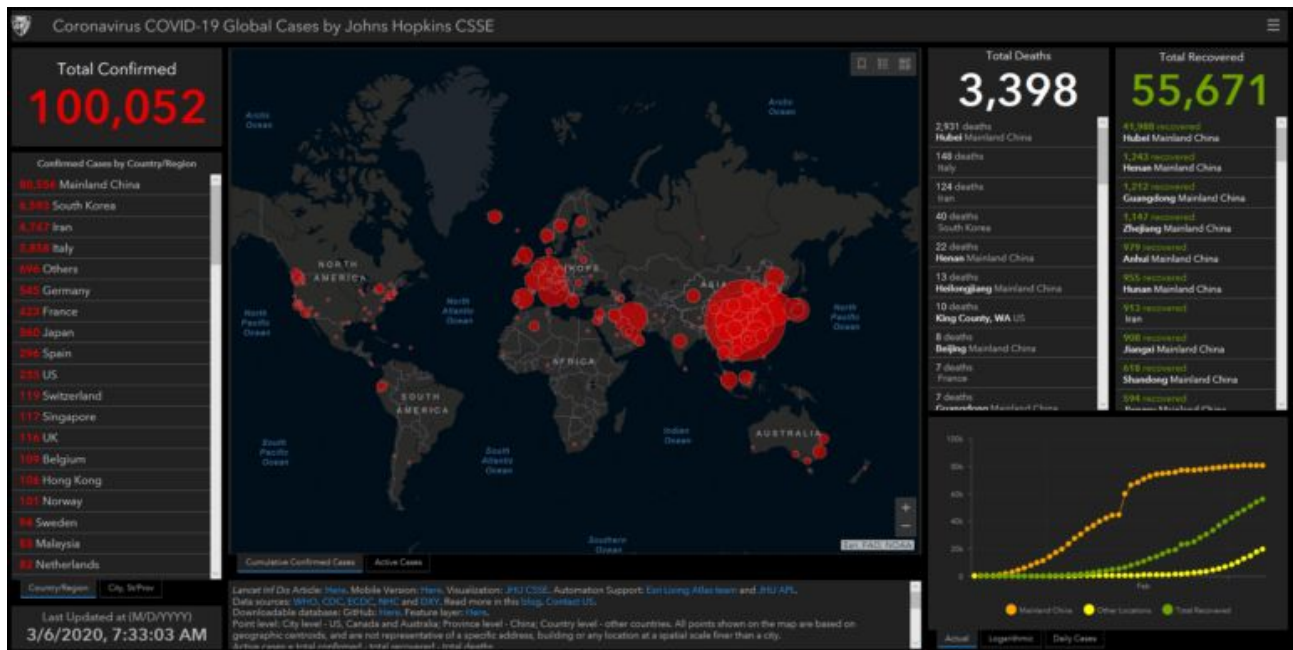


Figure 17: Dashboard based on the COVID-19 Data Repository (Dong, Du & Gardner, 2020).

3.17 Data on COVID-19

- **Provider:** Our World in Data
- **Link:** <https://github.com/owid/covid-19-data/tree/master/public/data>
- **Description:** Complete COVID-19 dataset is a collection of the COVID-19 data maintained by Our World in Data. It is updated daily and includes data on confirmed cases, deaths, hospitalizations, and testing, as well as other variables of potential interest.

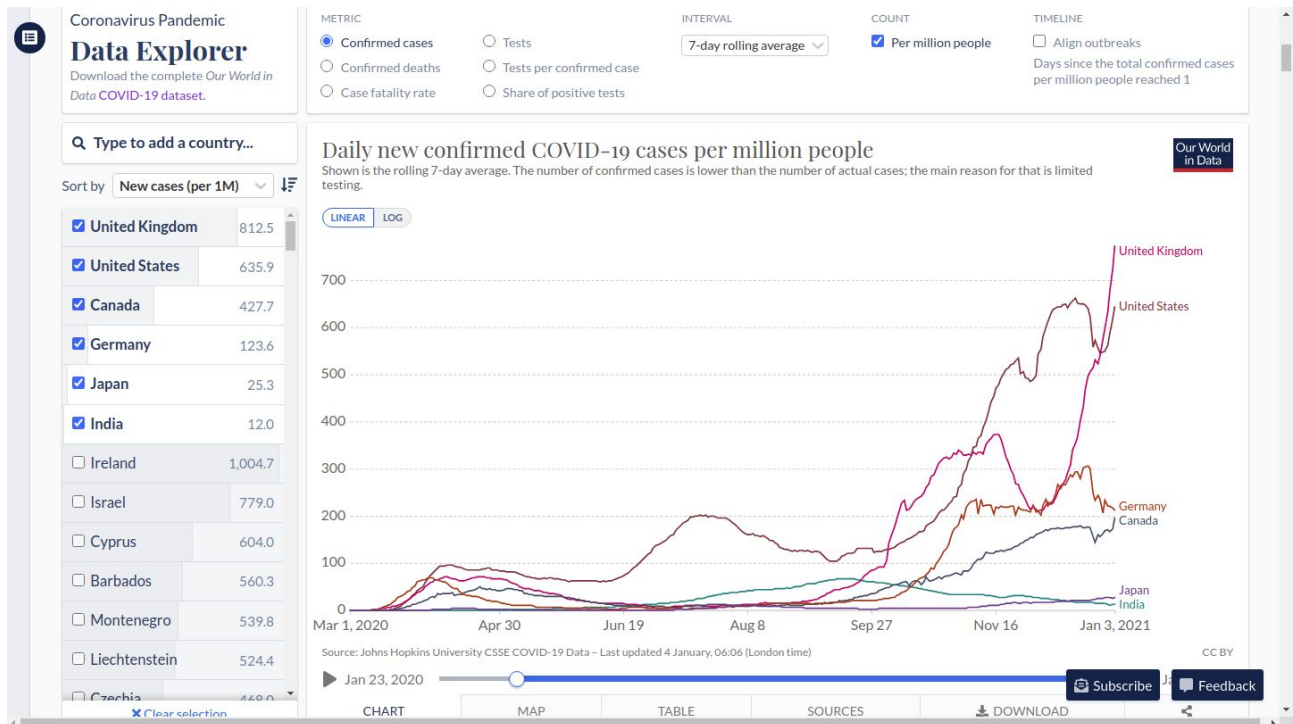


Figure 18: Web Explorer of the COVID-19 data maintained by Our World in Data (Hasell et al., 2020).

3.18 COVID-19 World Survey Data API

- **Provider:** University of Maryland
- **Link:** <https://covidmap.umd.edu/api.html>
- **Description:** API for accessing the daily global Facebook symptoms survey data



COVID-19 World Survey Data

API

An open API from the University of Maryland

About

This is an API for accessing the daily global Facebook symptoms survey data. The details of our methodology and disclaimer can be checked [here](#).

For updates, including data corrections, new aggregates, and any other changes to the API, we encourage you to [subscribe](#) to our COVID-19 API mailing list.

Figure 19: COVID-19 World Survey Data API (Barkay et al, 2020).

3.19 Data for the Open COVID-19 Data Working Group

- **Provider:** University of Washington
- **Link:** <https://github.com/beoutbreakprepared/nCoV2019>
- **Description:** Location for summaries and analysis of data related to n-CoV 2019, first reported in Wuhan, China.

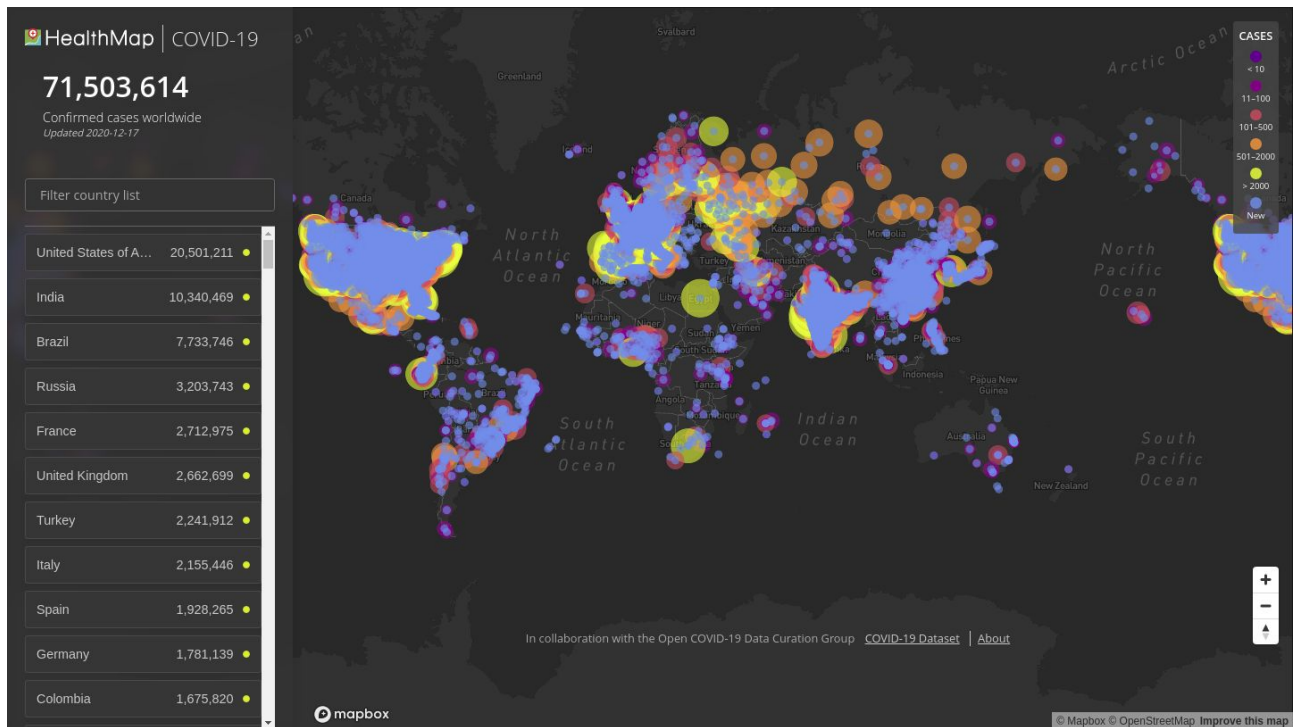


Figure 20: Dashboard for the data of the Open COVID-19 Data Working Group (Open COVID-19 Data Working Group, 2020).

3.20 CCCSL: CSH Covid-19 Control Strategies List

- **Provider:** Complexity Science Hub
- **Link:** <https://github.com/amel-github/covid19-interventionmeasures>
- **Description:** A wide range of different public sources were used to populate, update and curate our dataset, including official government sources, peer-reviewed and non-peer-reviewed scientific papers, webpages of public health institutions (WHO, CDC, and ECDC), press releases, newspaper articles, and government communication through social media.

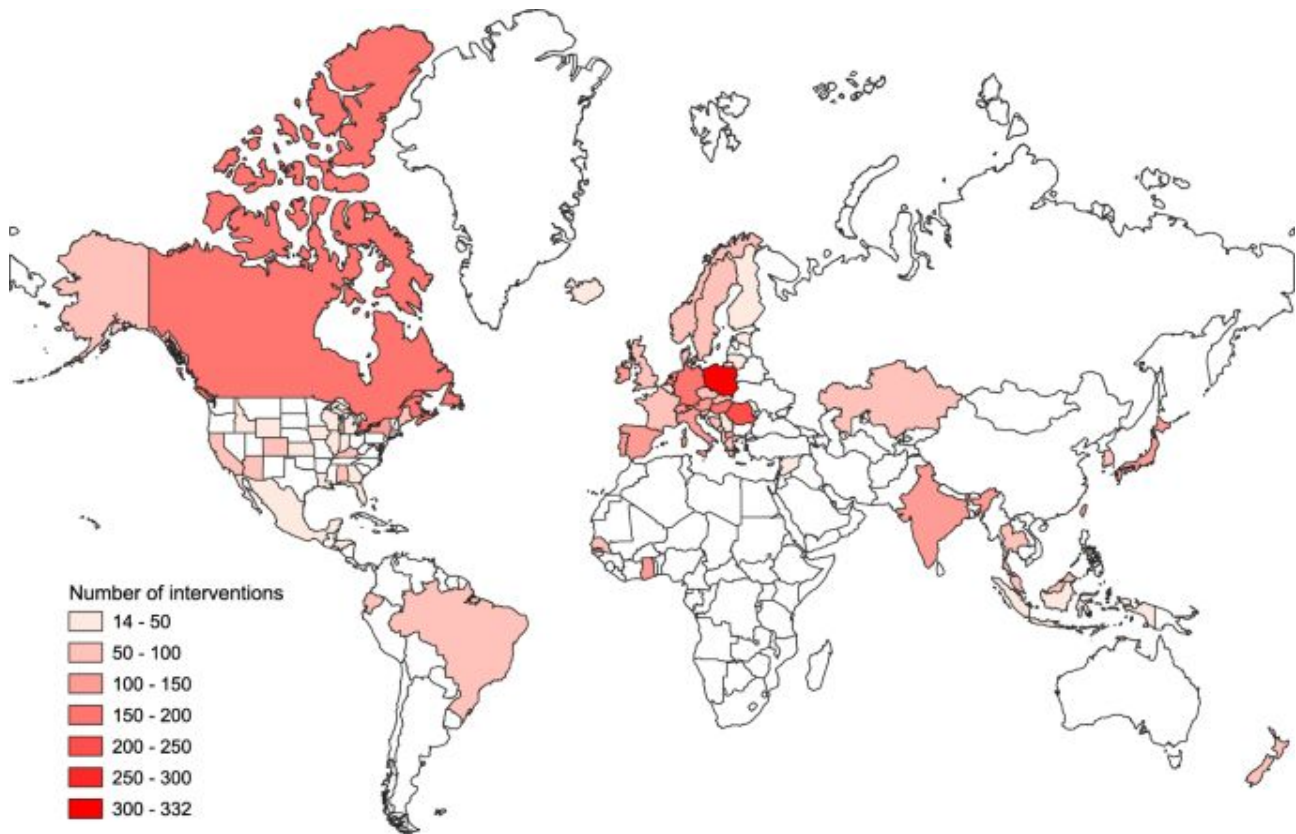


Figure 21: Geographical coverage of the CCCSL and total number of recorded NPIs that were implemented in each country to control the spread of COVID-19 (Desvars-Larrive et al., 2020).

3.21 Health Intervention Tracking for COVID-19 (HIT-COVID) Data

- **Provider:** Boston University and Johns Hopkins University
- **Link:** <https://github.com/HopkinsIDD/hit-covid>
- **Description:** The Health Intervention Tracking for COVID-19 (HIT-COVID) project tracks the implementation and relaxation of public health and social measures (PHSMs) taken by governments to slow transmission of SARS-COV-2 globally. Hundreds of volunteer data contributors were trained, provided with standardized field definitions and access to an online forum for asking questions and sharing ideas.

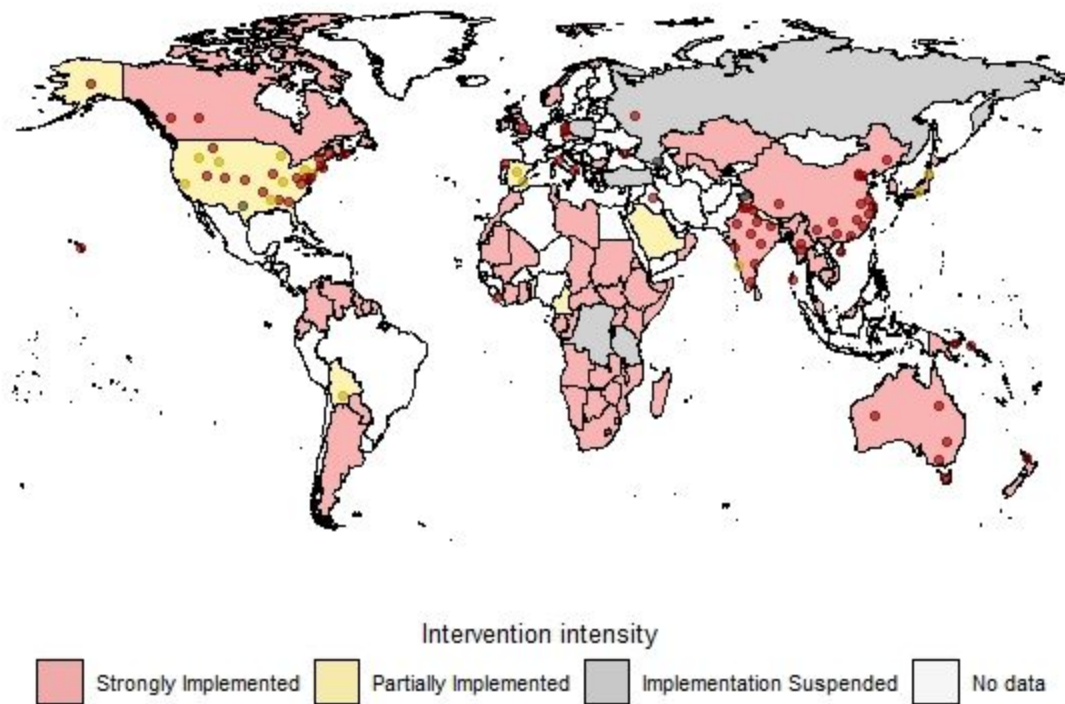


Figure 22: World map of sample PHSM (quarantine and isolation) implementation intensity since January 1, 2020 based on the Health Intervention Tracking for COVID-19 (HIT-COVID) Data.

3.22 Mozilla COVID dataset

- **Provider:** Mozilla Foundation
- **Link:** <https://blog.mozilla.org/data/2020/03/30/opening-data-to-understand-social-distancing>
- **Description:** Data about user browsing in Mozilla Firefox to understand social distancing

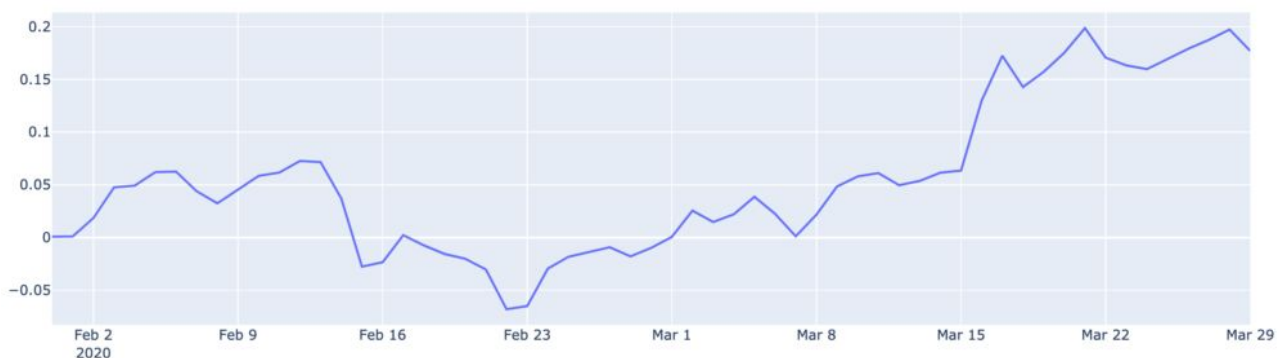


Figure 23: Firefox Desktop DAU deviation in France.

3.23 The CoVidAffect dataset

- **Provider:** CoVidAffect project

- **Link:** <https://doi.org/10.5281/zenodo.4024141>
- **Description:** Data from CoVidAfect, a nationwide citizen science project aimed to provide longitudinal data of mood changes following the COVID-19 outbreak in the spanish territory

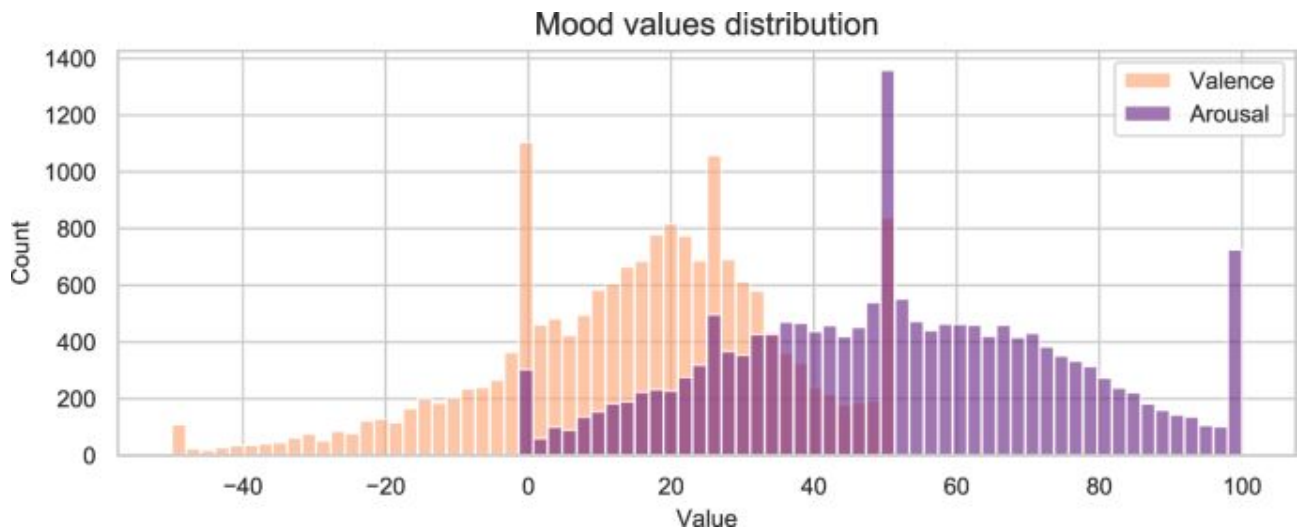


Figure 24: Distribution of reported valence and arousal values In the CoVidAfect dataset (Bailon, 2020).

3.24 COVID-19 Mobility Monitoring project

- **Provider:** ISI Foundation and Cuebiq
- **Link:** <https://covid19mm.github.io/data.html>
- **Description:** Data from COVID-19 Mobility Monitoring project that analyses anonymized location data to understand the effect of mobility restrictions and behavioral changes on the current international COVID-19 outbreak.

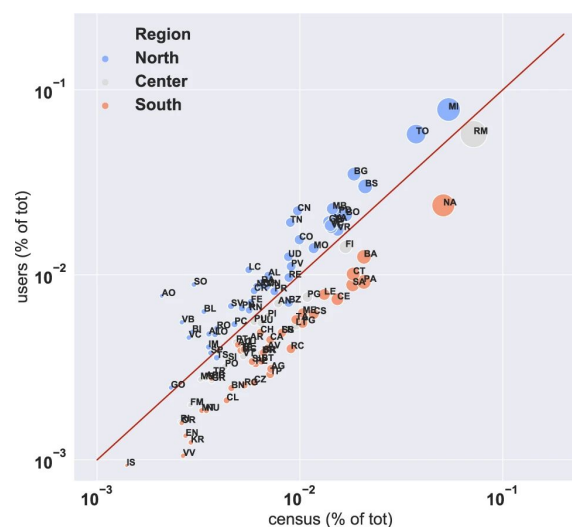


Figure 25: Scatterplot of the number of users of the COVID-19 Mobility Monitoring project assigned to each Italian province against the resident population reported by the Italian census in each province, as a fraction of the totals (Pepe, 2020).

3.25 #Data4COVID19

- **Provider:** The GovLab
- **Link:** <https://data4covid19.org>
- **Description:** A series of projects to identify, collect, and analyze the value data can provide to the ongoing COVID-19 pandemic

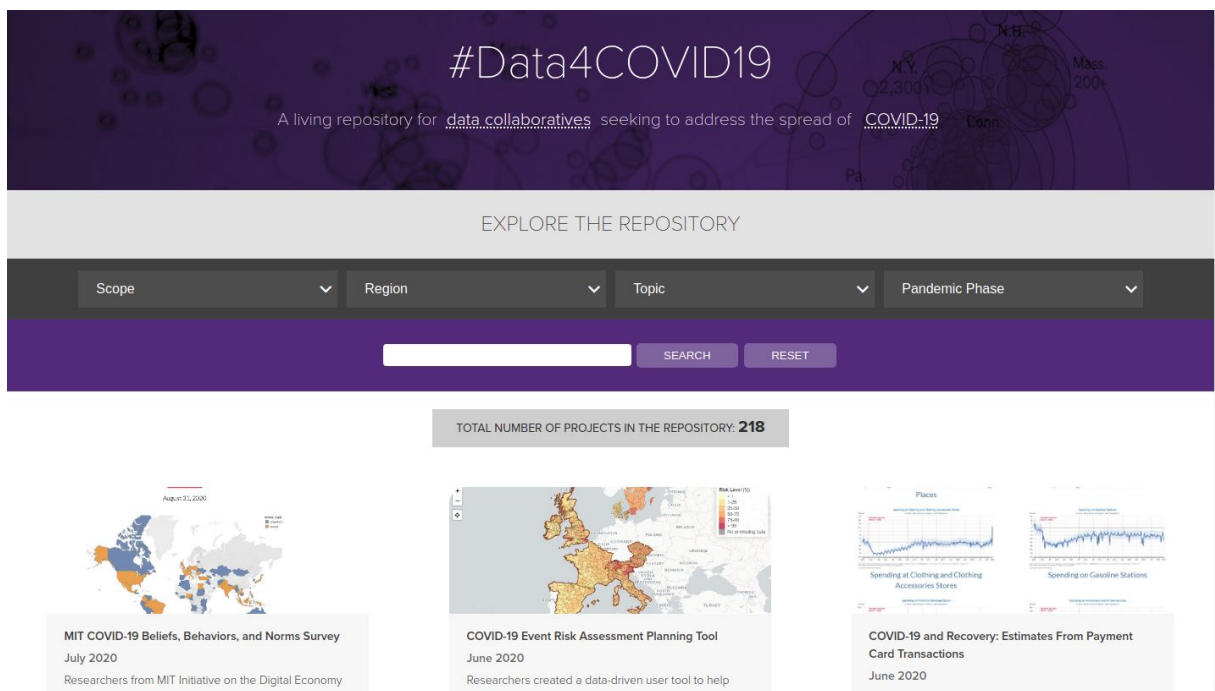


Figure 26: #Data4COVID19.

4 Tools

The participants of MediaFutures, both startups and artists, will need technological tools to carry out their data-driven projects. The team of mentors from WP4 has identified tools grouped into three distinct categories:

- **Digital Methods:** Platforms with user-friendly interfaces (most do not require programming skills) to collect, analyze and visualize data from the Internet.
<https://mediafutureseu.github.io/digitalmethods>
- **Data Science:** Python libraries for advanced data mining techniques.
<https://mediafutureseu.github.io/datascience>
- **Digital Art:** Technologies for creative and artistic purposes.
<https://mediafutureseu.github.io/digitalart>

In the same way that the list of datasets was based on open data resources, this list only includes open source tools. However, we recall that participants are free to use these or other tools.

4.1 Digital methods

The selected platforms for digital methods were implemented by individual researchers, organizations and academic institutions like the Digital Methods Initiative (University of Amsterdam)⁵, the Médialab (Sciences Po)⁶ or the DensityDesign research studio (Politecnico di Milano)⁷ for the study of societal change and cultural condition with online data (Rogers, 2013). As shown in Table 2, we grouped these platforms into 4 categories: data collection, data cleaning, data analysis and data visualization.

Table 2: Platforms for digital methods.

Category	Tool	Data source / type
Data collection	DMI Instagram Scraper	Instagram
	Hyphe	Web
	YouTube Data Tools	Youtube
	Tumblr Tool	Tumblr
	Hydrator	Twitter
	DMI TCAT	Twitter
	Reddit Tools	Reddit
	Telegram Tools	Telegram
Data cleaning	Open Refine	Tabular
Data analysis (and visualization)	Palladio	Complex, temporal and spatial
	Gephi	Networked
	Voyant	Text
Data visualization	RAWGraphs	Tabular

We list them below with a short description, indicating the provider and link to each tool too.

4.1.1 DMI Instagram Scraper

- **Provider:** Digital Methods Initiative
- **Link:** <https://auth.issuecrawler.net>

⁵ <https://wiki.digitalmethods.net/Dmi/DmiAbout>

⁶ <https://medialab.sciencespo.fr>

⁷ <https://densitydesign.org>

- **Description:** GUI for Instaloader to scrape users and hashtags with on Instagram

4.1.2 Hyphe

- **Provider:** Médialab - Sciences Po
- **Link:** <http://hyphe.medialab.sciences-po.fr/>
- **Description:** Websites crawler with built-in exploration and control web interface

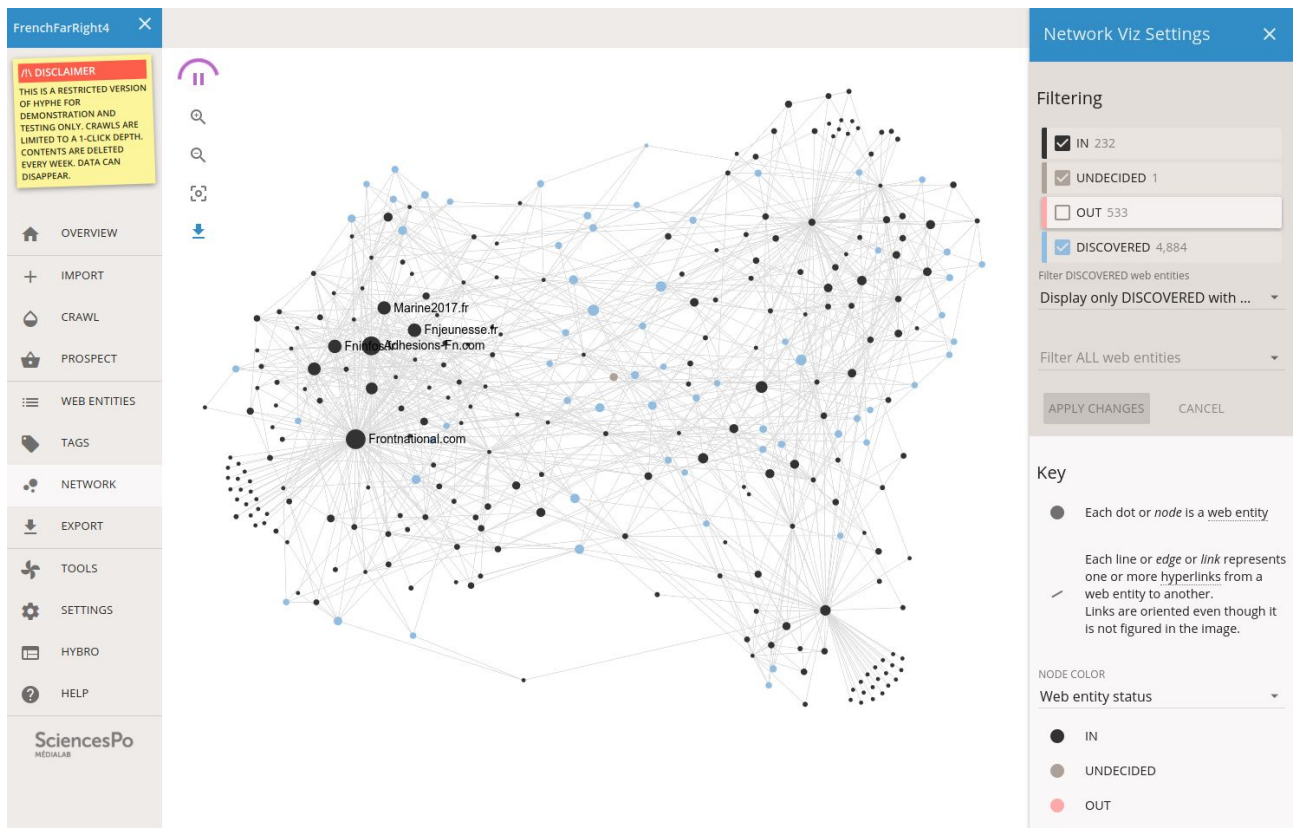


Figure 27: Hyphe, a curation-oriented approach to web crawling for the social sciences (Jacomy, Girard, Ooghe-Tabanou & Venturini, 2016).

4.1.3 YouTube Data Tools

- **Provider:** Digital Methods Initiative
- **Link:** <https://tools.digitalmethods.net/netvizz/youtube/>
- **Description:** Collection of simple tools for extracting data from the YouTube platform via the API v3

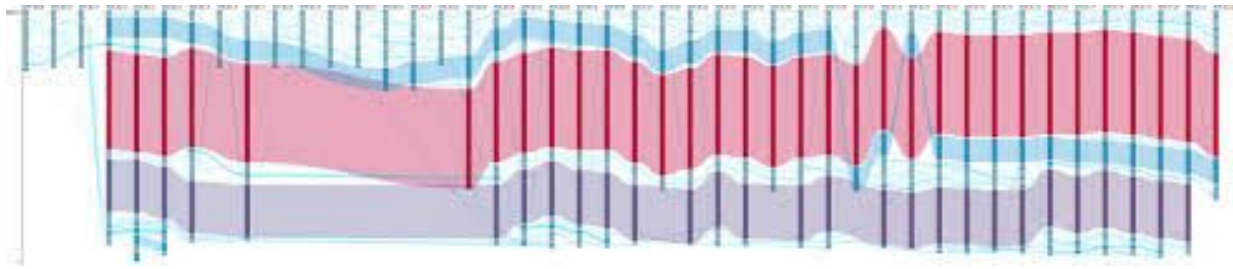


Figure 28: RankFlow visualization of a query using YouTube Data Tools (Rieder., Matamoros-Fernández & Coromina, 2018).

4.1.4 Tumblr Tool

- **Provider:** Bernhard Rieder
- **Link:** <http://labs.polsys.net/tools/tumblr/>
- **Description:** Script that extracts data from tumblr to create co-tag networks and tabular post stats

4.1.5 Hydrator

- **Provider:** Documenting the Now
- **Link:** <https://github.com/DocNow/hydrator>
- **Description:** Electron based desktop application for hydrating Twitter ID datasets

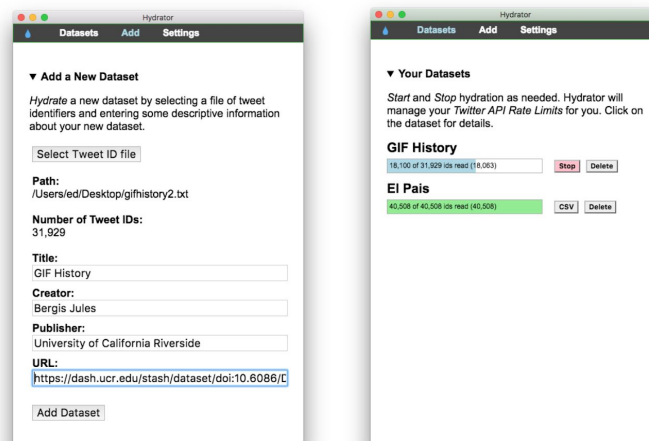


Figure 29: Screenshots of Hydrator (Documenting the Now, 2020).

4.1.6 DMI TCAT

- **Provider:** Digital Methods Initiative
- **Link:** <https://github.com/digitalmethodsinitiative/dmi-tcat/wiki>

- **Description:** Toolset to retrieve and collect tweets from Twitter and to analyze them in various ways

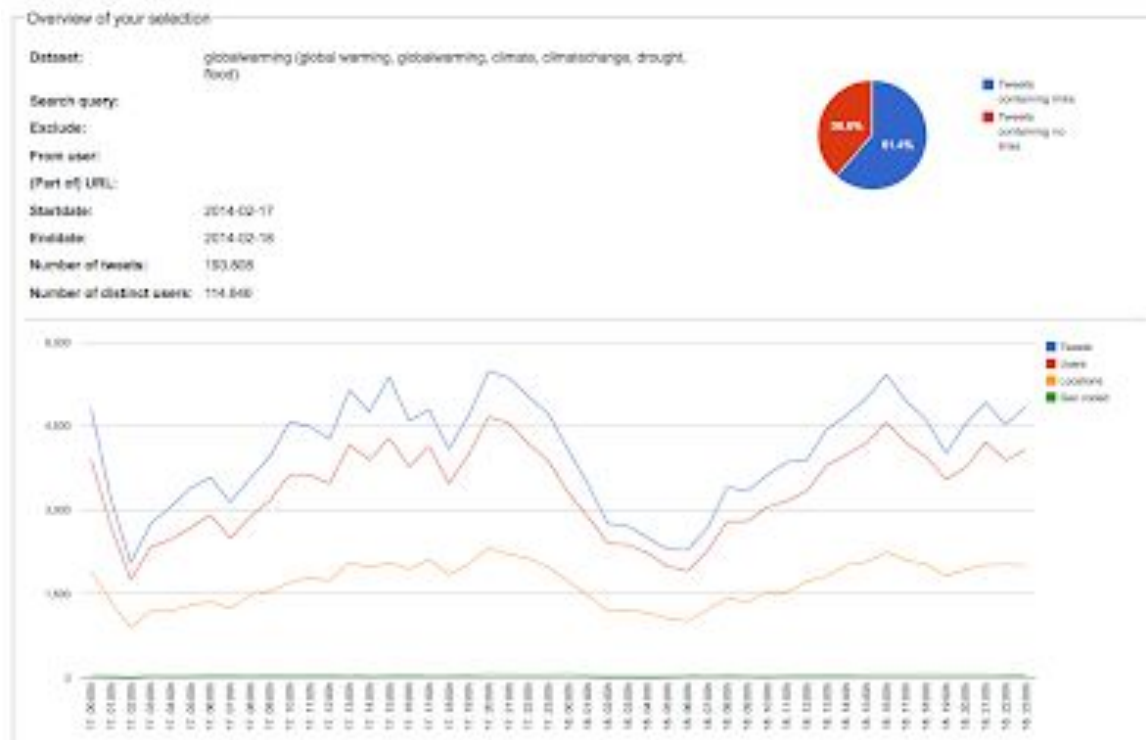


Figure 30: Digital Methods Initiative Twitter Capture and Analysis Toolset - DMI TCAT (Borra & Rieder, 2014).

4.1.7 Reddit Tools

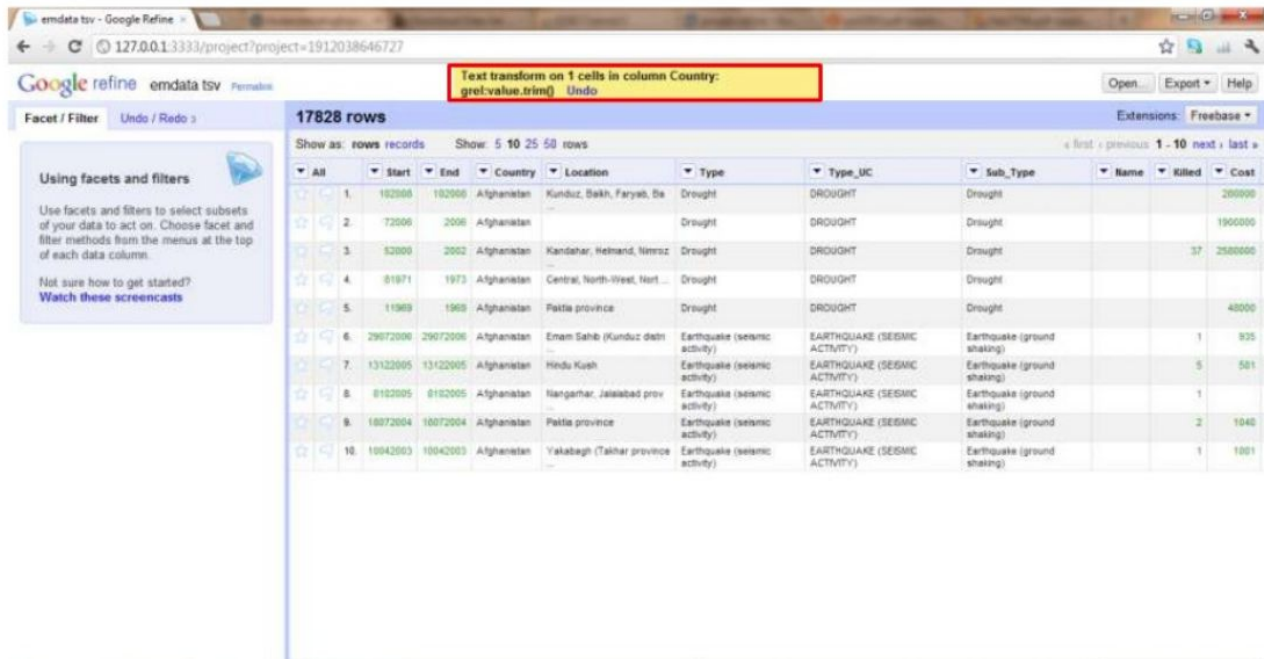
- **Provider:** Bernhard Rieder
- **Link:** <https://github.com/bernorieder/reddit-tools>
- **Description:** Collection of PHP command line scripts to grab data from Reddit and transform it into CSV files

4.1.8 Telegram Tools

- **Provider:** Bernhard Rieder
- **Link:** https://github.com/bernorieder/telegram_tools
- **Description:** Collection of tools for data retrieval from Telegram

4.1.9 Open Refine

- **Provider:** Metaweb Technologies
- **Link:** <https://openrefine.org/>
- **Description:** Free, open source power tool for working with messy data and improving it



emdata.tsv - Google Refine

127.0.0.1:3333/project?project=1912038646727

Google refine emdata.tsv

Text transform on 1 cells in column Country:
grel:value.trim() Undo

Facet / Filter Undo / Redo 3

17828 rows

Show as: rows records Show: 5 10 25 50 rows

Extensions: Freebase

Using facets and filters
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.
Not sure how to get started?
Watch these screencasts

	All	Start	End	Country	Location	Type	Type_UC	Sub_Type	Name	Killed	Cost
1.	182008	182008		Afghanistan	Kunduz, Balkh, Faryab, Ba...	Drought	DROUGHT	Drought			200000
2.	72006	2006		Afghanistan		Drought	DROUGHT	Drought			1900000
3.	52000	2002		Afghanistan	Kandahar, Helmand, Nimroz	Drought	DROUGHT	Drought		37	2500000
4.	81871	1973		Afghanistan	Central, North-West, Nort...	Drought	DROUGHT	Drought			
5.	11969	1969		Afghanistan	Paktia province	Drought	DROUGHT	Drought			40000
6.	29072006	29072006		Afghanistan	Emam Sahib (Kunduz distr...	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)		1	835
7.	13122005	13122005		Afghanistan	Hindu Kush	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)		5	581
8.	8182005	8182005		Afghanistan	Nangarhar, Jalaalabad prov...	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)		1	
9.	18072004	18072004		Afghanistan	Paktia province	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)		2	1040
10.	10042003	10042003		Afghanistan	Yakabagh (Tajikhar province)	Earthquake (seismic activity)	EARTHQUAKE (SEISMIC ACTIVITY)	Earthquake (ground shaking)		1	1001

Figure 31: Screenshot of Google Refine (Huynh, 2011).

4.1.10 Palladio

- **Provider:** Humanities + Design a Research Lab at Stanford University
- **Link:** <https://hdlab.stanford.edu/palladio>
- **Description:** Application for cleaning and explore complex, temporal and spatial data

P Data Map Graph **Table** Gallery v. 1.0.0 [Download](#)

Name ^	Arrival Point	Place	Place of Death	Pic	Date of Birth
Winaretta Singer	Monaco	London, Yonkers, Monaco	London	http://2.bp.blogspot.com/_dHMUkWjxiWM/Sxv9UdvfICI/AAAAAAACDQw/100px/Winaretta%20Singer.jpg	1893-11-24
Grimaldi Princesse Marie Caroline Gilbert de Lametz	Monaco	Monaco, Coulommiers	Monaco	http://upload.wikimedia.org/wikipedia/commons/f/fd/Caroline_giberti.jpg	1804-01-25
Sarah Bernhardt	Monaco	Paris, Monaco	Paris	http://upload.wikimedia.org/wikipedia/commons/f/f1/Sarah_Bernhardt.jpg	1859-03-26
Sara Murphy	Monaco	Arlington, Monaco, Cincinnati	Arlington	http://lbp.blogspot.com/-NBQR8TQOaqs/TeLtpXrllI/AAAAAAAMzMI/vjy0v4vXnQs/sara.jpg	1870-01-10
Roland Bonaparte	Monaco	Paris, Monaco	Paris	http://upload.wikimedia.org/wikipedia/en/4/43/Prince_Roland_Bonaparte_3641567004_c517894947_o.jpg	1924-4-14
Rene Leon	Monaco	Paris, Monaco		http://upload.wikimedia.org/wikipedia/commons/a/ab/WALDST3.jpg	
Raoul Gunsbourg	Monaco	Monaco, Bucharest	Monaco	http://www.opera.mc/graphx/side_historique_3.jpg	1955-5-31
Pierre Polovtsoff	Monaco	Monaco, St. Petersburg		http://upload.wikimedia.org/wikipedia/commons/a/ab/WALDST3.jpg	
Pierre Auguste Daval	Monaco	Monaco, France		http://upload.wikimedia.org/wikipedia/commons/a/ab/WALDST3.jpg	
Pablo Picasso	Monaco	Malaga, Monaco, Mougins	Mougins	http://upload.wikimedia.org/wikipedia/en/4/41/Portrait_of_Pablo_Picasso%2C_1908-1909%2C_anonymous_photographer%2C_Mus%C3%A9e_Picasso%2C_Paris.jpg	1973-4-8
Napoleon Langlois	Monaco	France, Monaco		http://upload.wikimedia.org/wikipedia/commons/a/ab/WALDST3.jpg	
Mata Hari	Monaco	Leeuwarden, Paris, Monaco	Paris	http://upload.wikimedia.org/wikipedia/commons/4/41/Mata_Hari_2.jpg	1917-10-15
Marie Therese Blanc	Monaco	Monaco, Courthezon		http://upload.wikimedia.org/wikipedia/commons/d/d4/JaneAustenSilhouette.png	
Marie Juliette Louvet	Monaco	Monaco, Paris, Pierrefeu	Paris	http://upload.wikimedia.org/wikipedia/commons/d/d4/JaneAustenSilhouette.png	1930-9-24
Marie Félix Blanc	Monaco	Saint Cloud, Monaco, Paris	Saint Cloud	http://upload.wikimedia.org/wikipedia/commons/4/44/Marie_Charlotte_Blanc.jpg	1882-8-1
Marie Blanc	Monaco	Moutiers, Monaco, Friedrichsdorf	Moutiers	http://upload.wikimedia.org/wikipedia/commons/d/db/Marie_Charlotte_Hensel_%26%C3%A9pouse_Blanc%29.jpg	1881-7-25
Magdeleine-Victoire Huguelin	Monaco	Monaco		http://upload.wikimedia.org/wikipedia/commons/d/d4/JaneAustenSilhouette.png	1852-01-01
Ludwig Jacobi	Monaco	Germany, Monaco		http://upload.wikimedia.org/wikipedia/commons/a/ab/WALDST3.jpg	
Louise Blanc	Monaco	Paris, Monaco	Paris	http://www.marcelproust.it/magg/radziwill_principessa.jpg	1911
Louis-Philippe	Monaco	Claremont, Monaco, Paris	Claremont	http://upload.wikimedia.org/wikipedia/commons/f/f3/Louis-Philippe_de_Bourbon.jpg	1850-08-26
Louis Blanc	Monaco	Courthezon, Monaco		http://upload.wikimedia.org/wikipedia/commons/a/ab/WALDST3.jpg	1852-01-01
Leopold II	Monaco	Brussels, Laeken, Monaco	Laeken	http://upload.wikimedia.org/wikipedia/commons/3/39/Leopold_II_garter_knight.jpg	1909-12-17

Facet Timeline Timespan You have no active filters

Figure 32: Palladio, humanities thinking about data visualization (Ceserani, 2015).

4.1.11 Gephi

- **Provider:** Gephi.org
- **Link:** <https://gephi.org>
- **Description:** Interactive visualization and exploration software for all kinds of networks and complex systems, dynamic and hierarchical graph

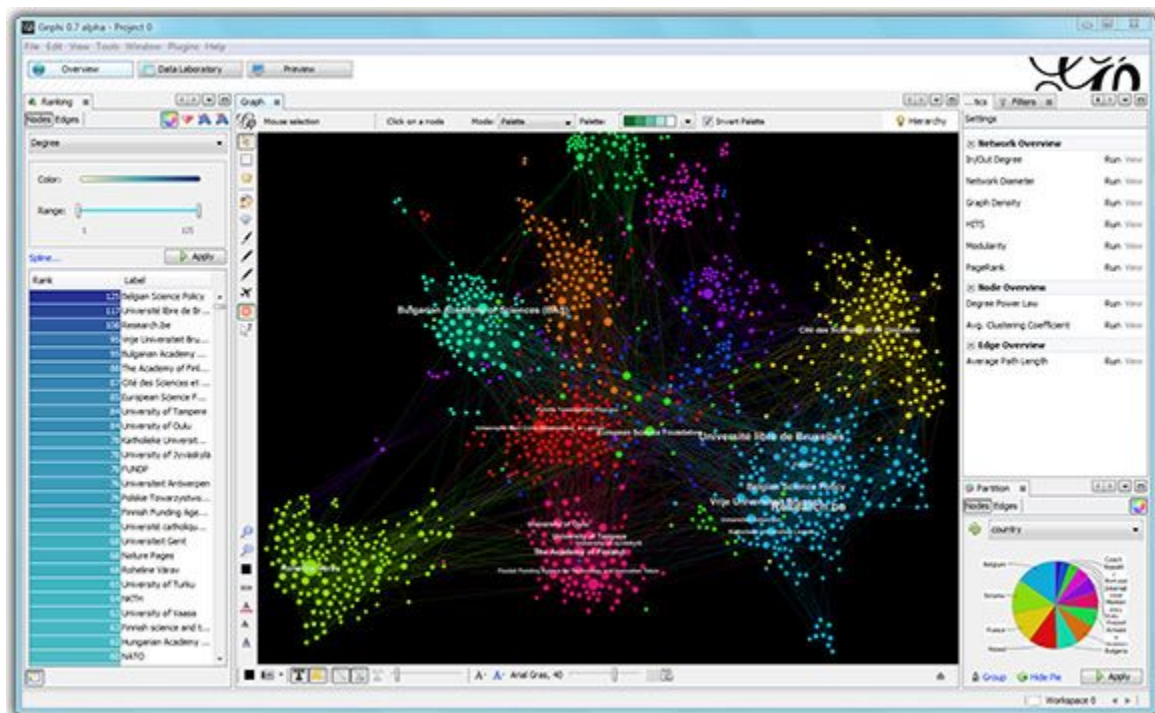


Figure 33: Gephi, an open source software for exploring and manipulating networks (Bastian, Heymann, & Jacomy, 2009).

4.1.12 Voyant

Web-based text analysis, reading and visualization environment

- **Provider:** Voyant Tools
- **Link:** <https://voyant-tools.org>
- **Description:** Websites crawler with built-in exploration and control web interface

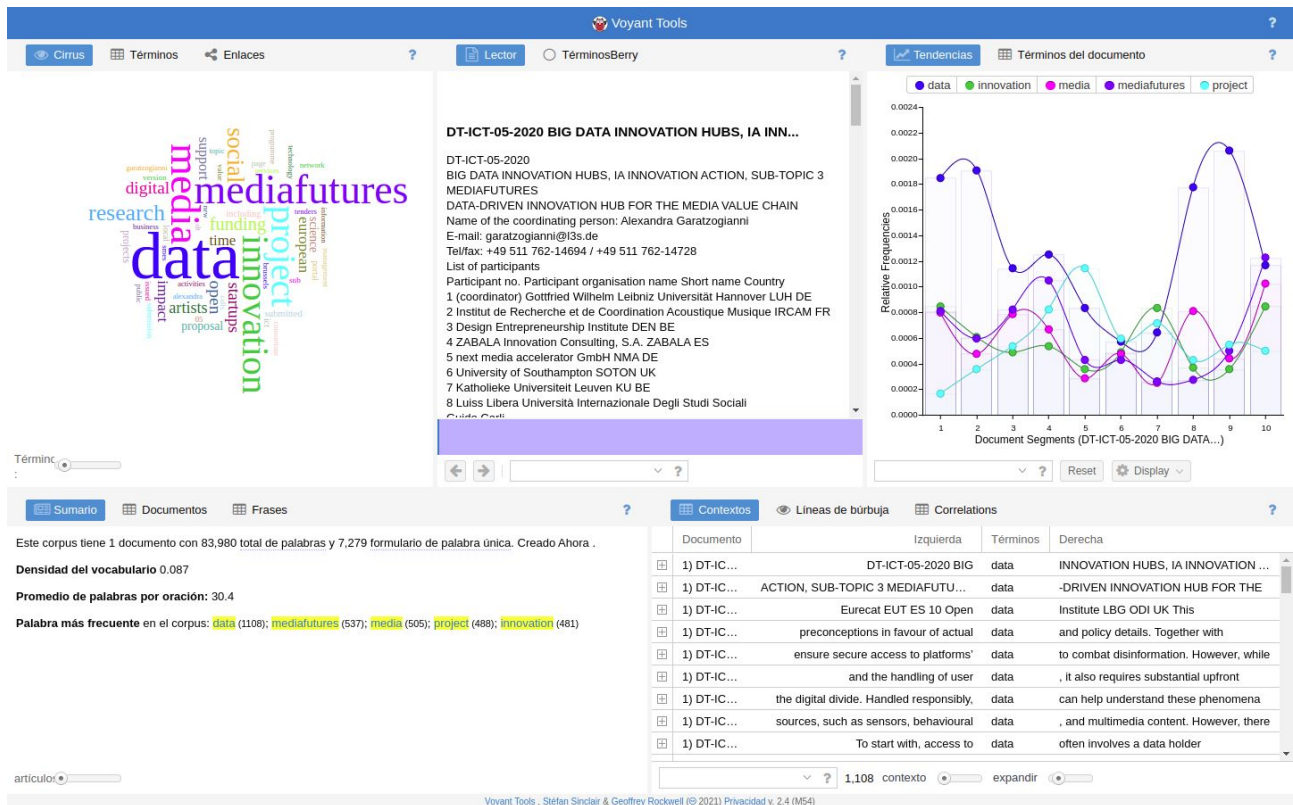


Figure 34: Screenshot of Voyant Tools.

4.1.13 RAWGraphs

- **Provider:** DensityDesign
- **Link:** <https://rawgraphs.io>
- **Description:** Open web tool to create custom vector-based visualizations on top of D3.js library

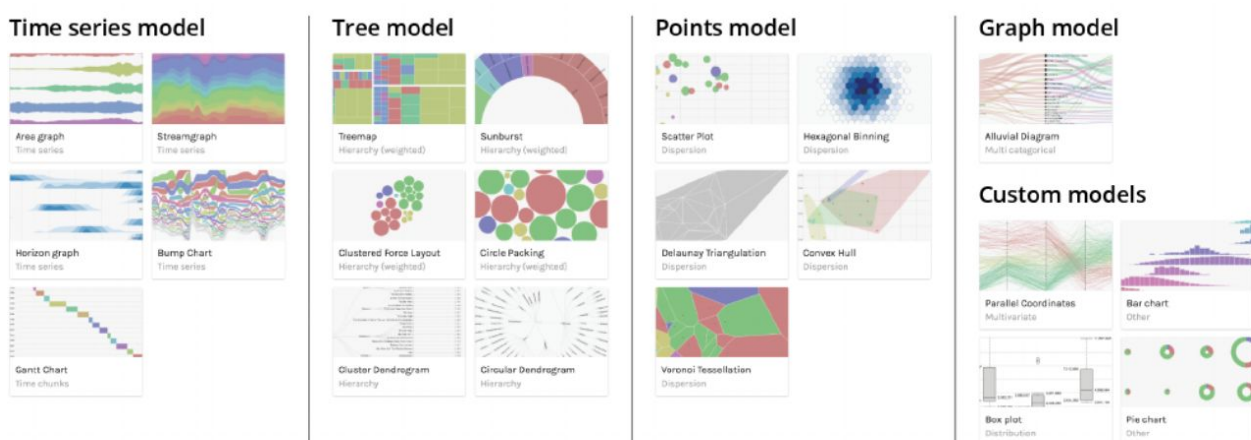


Figure 35: Charts provided by RAWGraphs, grouped by data model (Mauri, Elli, Caviglia, Ubaldi, & Azzi 2017).

4.2 Data Science

The selected Python libraries are essential tools of data scientists for applying machine learning techniques on data from the Web (Isoni, 2016). As shown in Table 3, we grouped them into 3 categories: data collection, data analysis and data visualization.

Table 3: Python libraries for data science.

Category	Tool	Data source / type
Data collection	Scrapy	Web
Data analysis	NLTK	Text
	Gensim	Text
	Scikit learn	Any
	NetworkX	Networked
Data visualization	Plotly	Any
	Matplotlib	Any

We list them below with a short description, indicating the provider and link to each tool too.

4.2.1 Scrapy

- **Provider:** Scrapinghub
- **Link:** <https://scrapy.org>
- **Description:** Fast high-level web crawling & scraping framework for Python

4.2.2 NLTK

- **Provider:** Steven Bird and Liling Tan
- **Link:** <http://www.nltk.org>
- **Description:** Suite of open source Python modules, data sets, and tutorials supporting research and development in Natural Language Processing

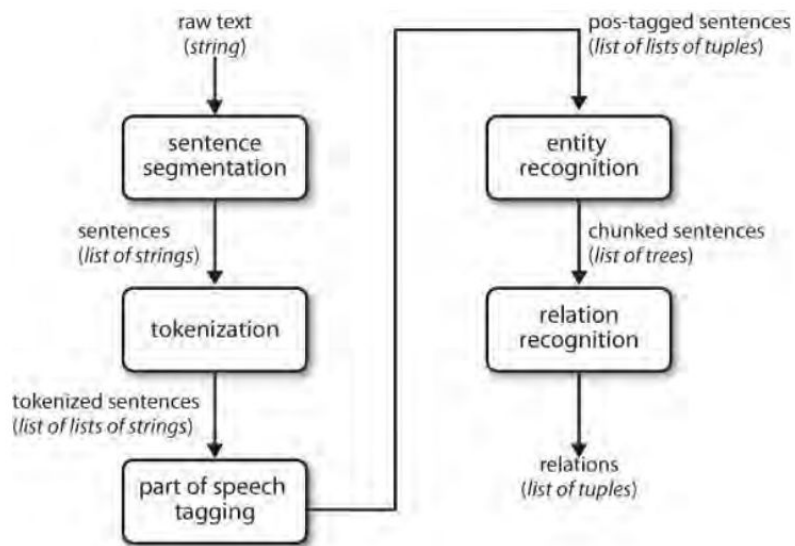


Figure 36: Simple pipeline architecture for an information extraction system with NLTK (Bird, Klein & Loper, 2009).

4.2.3 Gensim

- **Provider:** RARE Technologies
- **Link:** <https://radimrehurek.com/gensim>
- **Description:** Python library for topic modelling, document indexing and similarity retrieval with large corpora

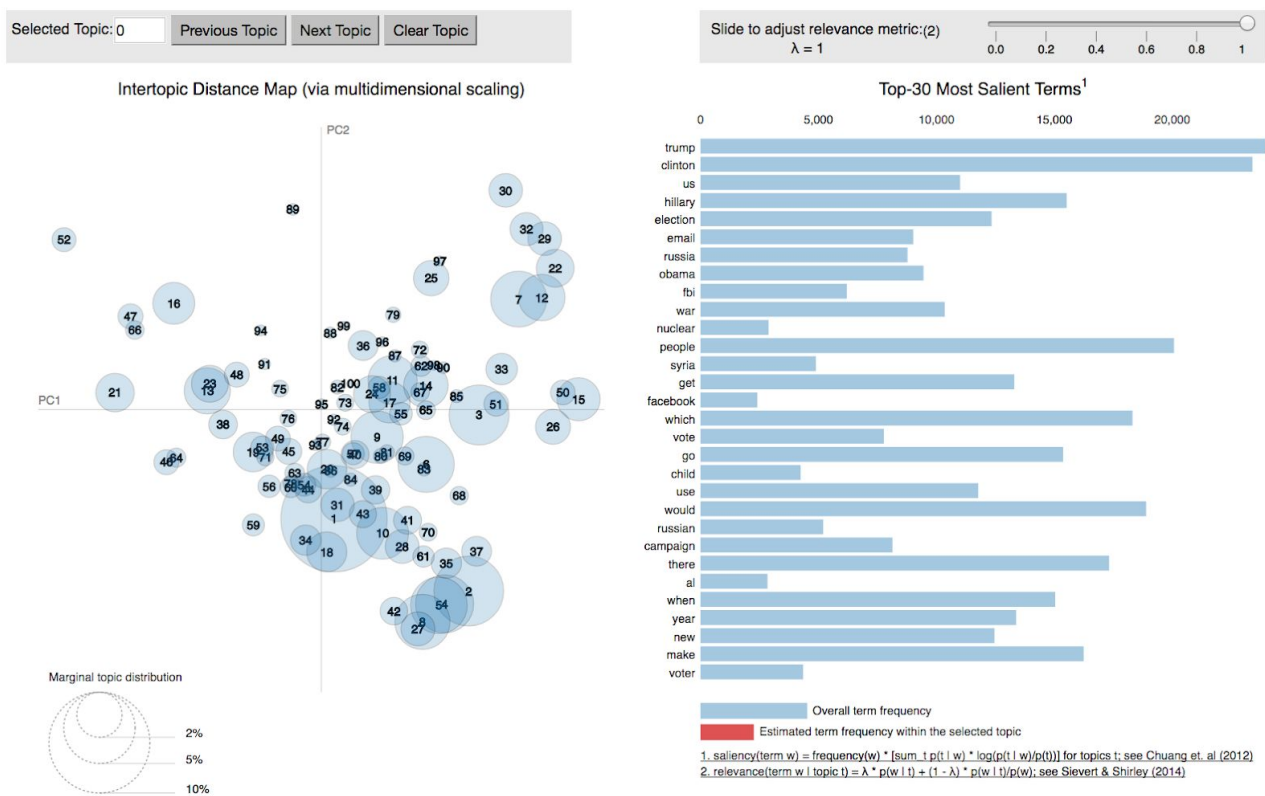
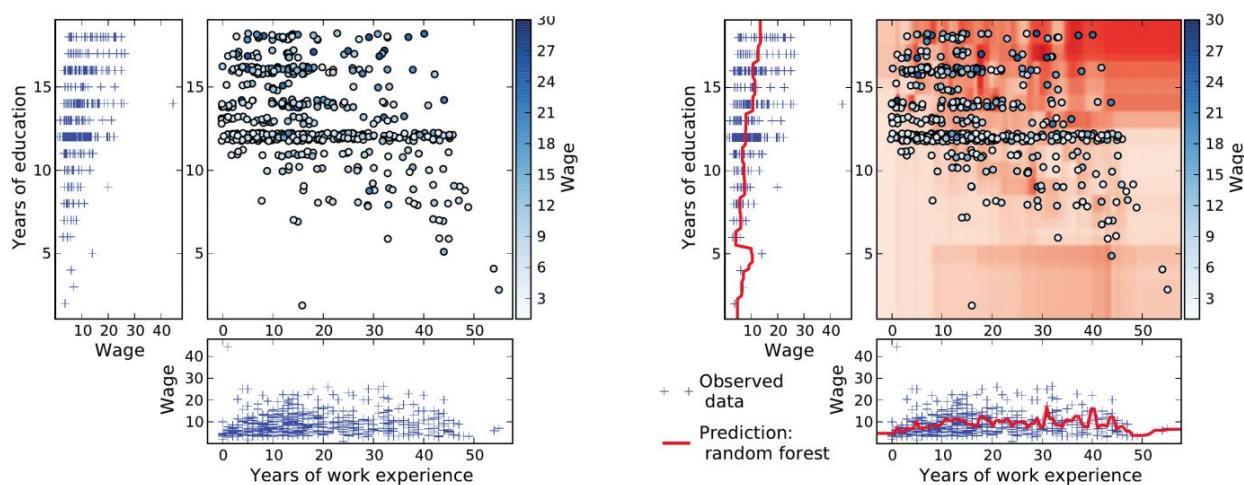


Figure 37: Visualization of a topic model with Gensim (Rehurek & Sojka, 2010).

4.2.4 scikit-learn

- **Provider:** Scikit community⁸
- **Link:** <https://scikit-learn.org>
- **Description:** Python module for machine learning built on top of SciPy and is distributed under the 3-Clause BSD license



⁸ Core contributors are listed at <https://scikit-learn.org/stable/about.html#authors>

Figure 38: Examples of machine learning models built with Scikit-learn (Varoquaux et al., 2015).

4.2.5 NetworkX

- **Provider:** NetworkX developers
- **Link:** <https://networkx.org>
- **Description:** Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks

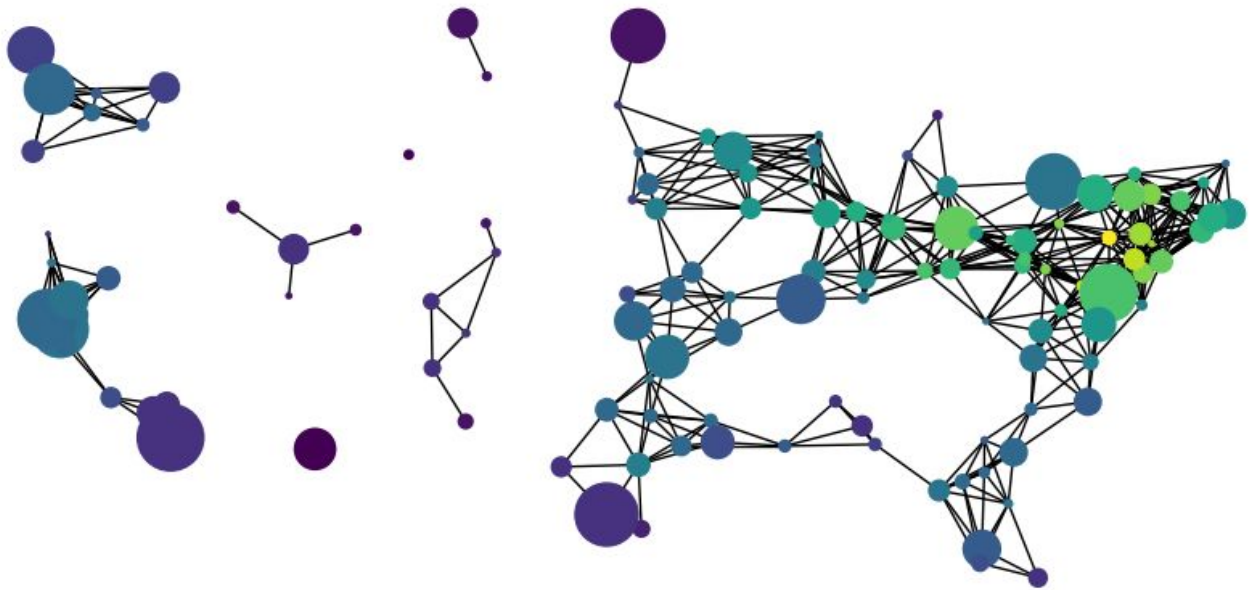


Figure 39: Chart generated with NetworkX (Hagberg, Schult & Swart, 2012).

4.2.6 Plotly

- **Provider:** Plotly
- **Link:** <https://plotly.com/python/>
- **Description:** Interactive graphing library for Python

4.2.7 Matplotlib

- **Provider:** John D. Hunter and the Matplotlib community
- **Link:** <https://matplotlib.org>
- **Description:** Comprehensive library for creating static, animated, and interactive visualizations in Python

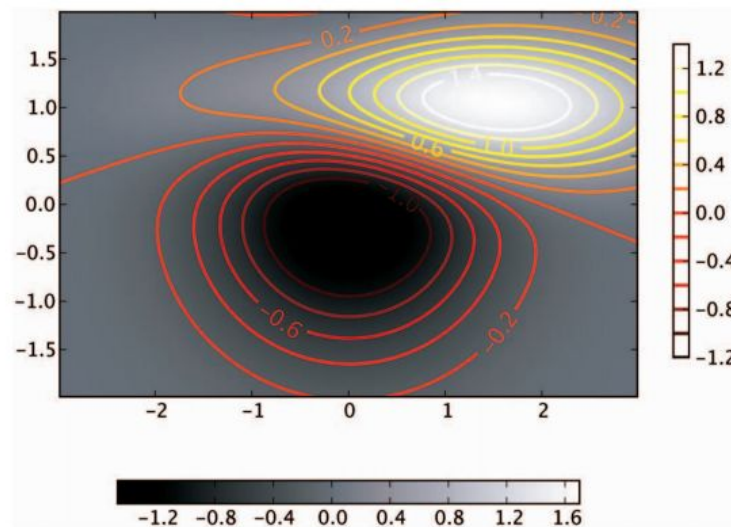


Figure 40: Chart generated with Matplotlib (Hunter, 2007).

4.3 Digital Art

The team of mentors from WP4 has very limited experience in digital art. However, 2 of the 3 tracks of the MediaFutures call involve artists (i.e., Artists for Media and Startup meets Artist). As a consequence, the catalogue includes a list of technologies that have proven to be effective for creative and artistic purposes. The goal of this list is therefore to inspire participants in the artistic tracks with illustrative open source tools for digital art. We list them below with a short description, indicating the provider and link to each tool too.

4.3.1 Processing

- **Provider:** Processing Foundation
- **Link:** <https://processing.org>
- **Description:** Flexible software sketchbook and a language for learning how to code within the context of the visual arts



Figure 41: Chart generated with Processing (Reas & Fry, 2015).

4.3.2 openFrameworks

- **Provider:** openFrameworks community
- **Link:** <https://openframeworks.cc>
- **Description:** Community-developed cross platform toolkit for creative coding in C++

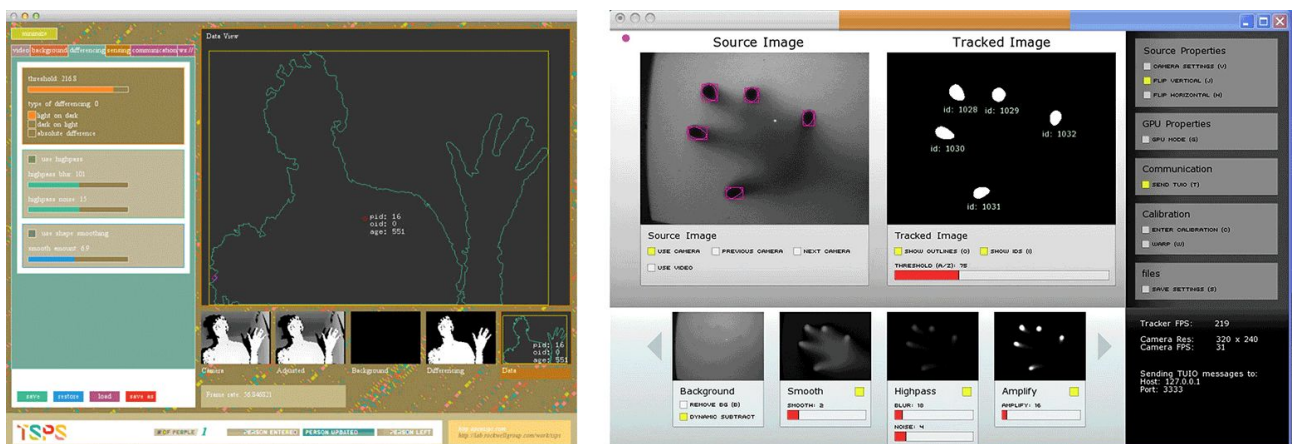


Figure 42: Chart generated with openFrameworks (Levin & Dorsey).

4.3.3 PureData

- **Provider:** Pd-community
- **Link:** <https://puredata.info>
- **Description:** Free real-time computer music system

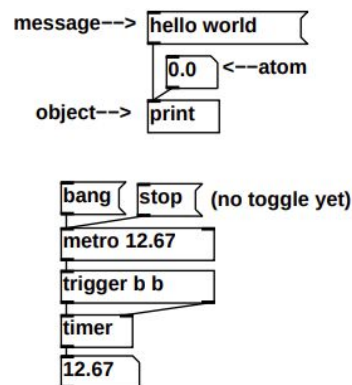


Figure 43: Patchable object generated with PureData (Puckette, 1997).

4.3.4 vvvv

- **Provider:** vvvv community
- **Link:** <https://vvvv.org>
- **Description:** Hybrid visual/textual live-programming environment for easy prototyping and development.

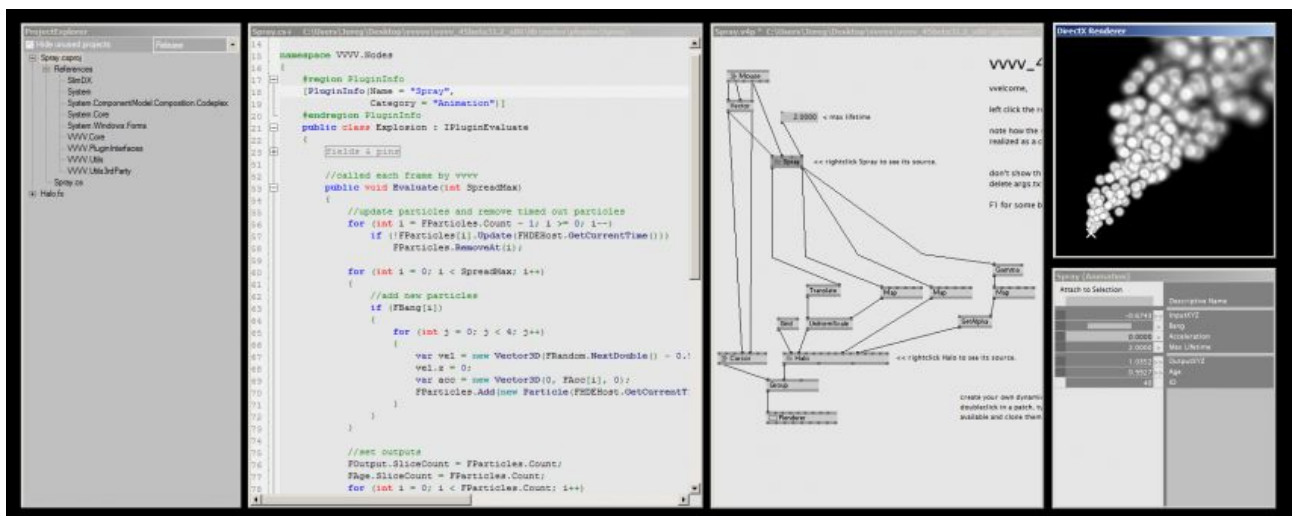


Figure 44: Hybrid development environment with vvvv (Bohnacker, Gross, Laub & Lazzeroni, 2012).

4.3.5 Cinder

- **Provider:** Cinder community
- **Link:** <https://libcinder.org>
- **Description:** Community-developed, free and open source library for professional-quality creative coding in C++

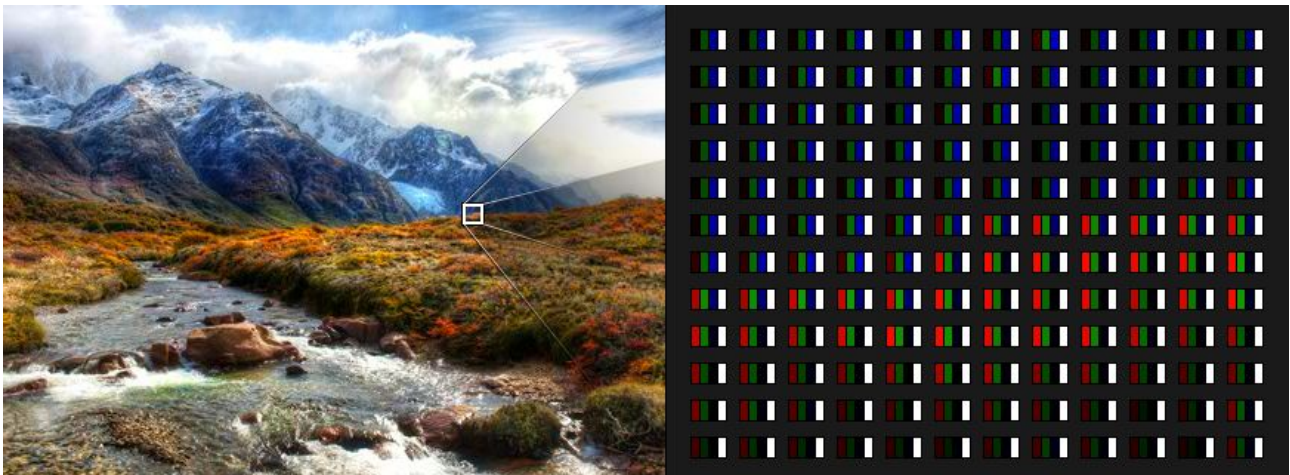


Figure 45: Images in Cinder <https://libcinder.org/docs/guides/cinder-images/index.html>.

4.3.6 Magenta

- **Provider:** Google AI
- **Link:** <https://magenta.tensorflow.org>
- **Description:** Open source research project exploring the role of machine learning as a tool in the creative process

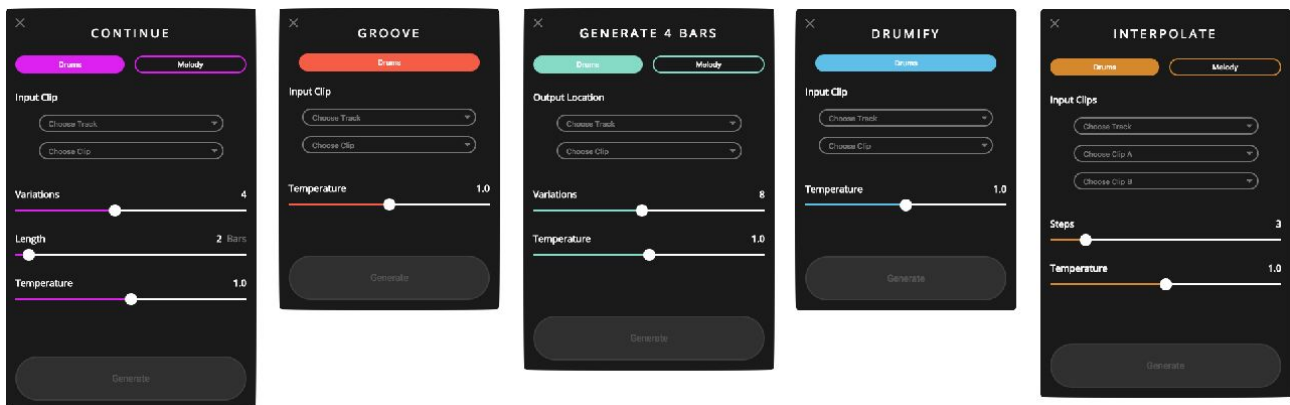


Figure 46: Plug-ins in Magenta Studio (Roberts et al., 2019).

4.3.7 D3.js

- **Provider:** John D. Hunter and the Matplotlib community
- **Link:** <https://d3js.org>
- **Description:** JavaScript library for visualizing data using web standards

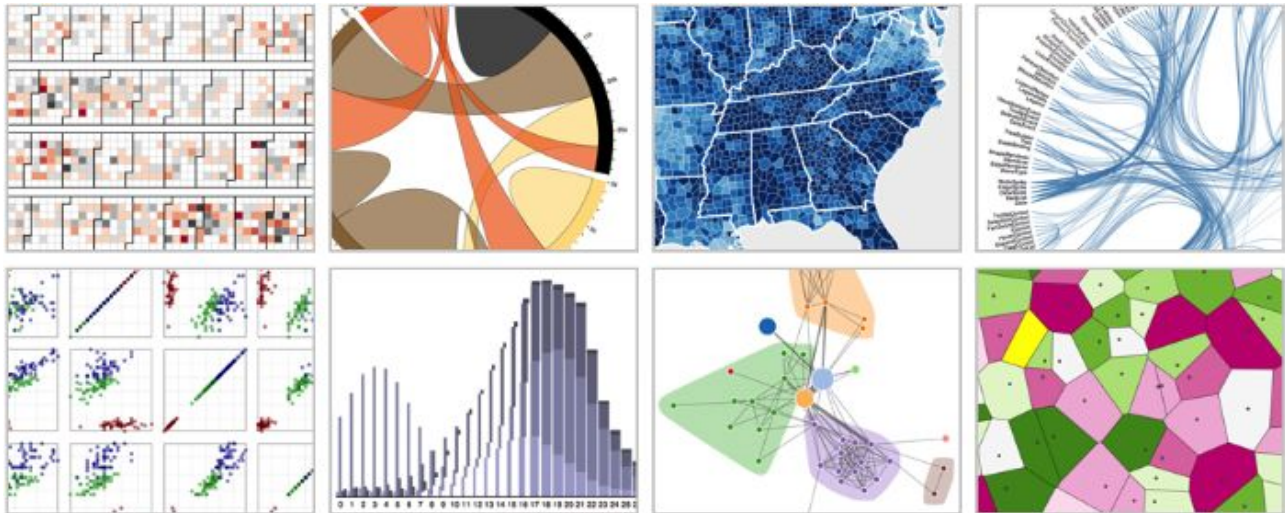


Figure 47: Interactive visualizations built with D3.js (Bostock, Ogievetsky & Heer, 2011).

5 Training

As part of the data innovation and experiments support work package, relevant training and learning opportunities will be provided by the ODI, KCL and other partners to successful startups and artists. The training will be delivered through a range of formats (webinars, elearning, workshops, courses) and will likely be fully remote for the first cohort. These will cover a range of topics from the basics of open data and how to use it to more advanced machine learning and AI techniques. There will also be more informal opportunities for the successful applicants to learn from the wider networks of partners e.g. at the ODI's annual Summit.

Existing material from partners will be drawn upon to provide to successful applicants, but tailored material will be developed to suit the specific needs of the startups and artists. The current full list is shown below^{9 10}.

5.1 Anonymisation is for everyone (ODI)

This course will introduce the latest thinking on anonymisation and work through a practical exercise in anonymising data while maintaining utility and minimising risk. Through the practical exercise, we will explore the needs of different stakeholders, both internal and external and explore how tools like the ODI Data Ethics Canvas and Consequence Scanning toolkit can help.

5.2 Applying Machine Learning and AI Techniques to Data (ODI)

The course takes a practical approach to understand the key machine learning techniques to help you understand what these black boxes might be doing, how they can be applied and what implications each has. During the course you will be challenged to build your own machine learning algorithm for a set of real world data.

⁹ A selection of existing courses have been promoted to interested participants at <https://mediafutureseu.github.io/training>

¹⁰ We will also publish a training calendar by the end of January 2021.

5.3 Introduction to data ethics and the data ethics canvas (ODI)

In this two-hour online course, run by our experts, you will be given an introduction to the concept of data ethics, how ethics is addressed in other domains, and how data ethics can be applied to help organisations avoid harmful impacts and increase trust in new products and services.

5.4 Open data in a day (ODI)

An introductory course to open data. Participants will learn how to discover, use and describe the benefits of open data, and how it can impact an organisation.

5.5 Strategic Data Skills (ODI)

This comprehensive online course is for people who want to work strategically with data, but don't need or want to learn complex programming. Start this course when you want, work at your own pace, and get access to the content for 12-months. Complete and submit your assignments within 3-months of starting and you will be eligible for your completion certificate.

5.6 Datapolis (ODI)

Workshop based on Datapolis, the open data board game exploring building things with open and closed data.

5.7 Open data essentials (ODI)

Learn all the essentials of open data with this easy to follow, online course.

5.8 Finding stories in data (ODI)

Learn how to tell compelling stories from data.

5.9 Annual event: ODI Summit (ODI)

ODI's flagship event, bringing people from a broad range of sectors, backgrounds and countries together to discuss critical issues around data.

5.10 Weekly lectures series: ODI Fridays (ODI)

Weekly lectures with a range of ODI and guest speakers exploring a broad range of data related topics.

5.11 Art-tech collaborations: best practices (IRCAM)

A webinar taking a practical approach to understanding the success factors for art-tech collaborations. It gives advice about how to efficiently start a co-creation process involving artists and startups. Some elements that will be covered are; collaboration rules, co-creation environment, common language, available resources, expectations.

5.12 Public funding opportunities and KAILA tool (Zabala)

A webinar presenting the different opportunities for public funding at the European level and their benefits. It will also include a tool recently launched by ZABALA - KAILA allows you to identify opportunities for public funding, relevant projects, actors with whom to collaborate, as well as the most important actors in each field, monitor technological trends, individualized advice according to your interests, etc. Each team that reaches the BUILD phase will be provided with a free license for the entire program.

5.13 Social Innovation (Zabala)

Training that will cover aligning the business models with the SDGs (Sustainable Development Goals) established by the United Nations and thus establishing a methodology with social impact. Specifically, it will explain how to define social risks based on the business model and implement actions to avoid them. This training will be composed of a webinar where the theoretical part of the SDGs will be explained and a case study of a technology company that has implemented this methodology. Afterwards, the participants will be encouraged to reflect and apply this to their business models. Later on, a workshop will be organised (if possible physically) to evaluate each practical case and help the teams.

5.14 Data journalism (LUISS)

Training on data journalism will be provided by LUISS, in the form of guidance, guidelines and learning materials aimed at improving digital and media literacy skills.

6 Conclusions

This deliverable has presented the infrastructure for data and tools in MediaFutures. Main resources have been listed including (1) datasets for the first Open Call with challenges about misinformation and coronavirus, (2) technical tools of digital methods, data science and digital art to exploit data, and (3) training courses available for participants. New datasets will be required in future calls about new challenges and new tools and training courses are expected to emerge. For that reason, the lists presented in the deliverable will be updated over time in the MediaFutures technical infrastructure for experimental support that has been deployed on GitHub.

7 References

- Bailon, C., Goicoechea, C., Banos, O., Damas, M., Pomares, H., Correa, A., ... & Perakakis, P. (2020). CoVidAffect, real-time monitoring of mood variations following the COVID-19 outbreak in Spain. *Scientific Data*, 7(1), 1-10.
- Barkay, N., Cobb, C., Eilat, R., Galili, T., Haimovich, D., LaRocca, S., ... & Sarig, T. (2020). Weights and Methodology Brief for the COVID-19 Symptom Survey by University of Maryland and Carnegie Mellon University, in Partnership with Facebook. *arXiv preprint arXiv:2009.14675*.
- Bastian, M., Heymann, S., & Jacomy, M. (2009, March). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 3, No. 1).
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."
- Bohnacker, H., Gross, B., Laub, J., & Lazzeroni, C. (2012). *Generative design: visualize, program, and create with processing*. Princeton Architectural Press.
- Borra, E., & Rieder, B. (2014). Programmed method: developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*.
- Bostock, M., Ogievetsky, V., & Heer, J. (2011). D³ data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12), 2301-2309.
- Ceserani, G. (2015). *Interactive Visualizations for British Architects on the Grand Tour in eighteenth-century Italy*. Palladio Components, HTML, CSS, Javascript, JSON, and Markdown files]. Stanford Digital Repository.
- Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance*, 6(2), e19273.
- Cui, L., & Lee, D. (2020). CoAID: COVID-19 Healthcare Misinformation Dataset. *arXiv preprint arXiv:2006.00885*.
- Desvars-Larrive, A., Dervic, E., Haug, N., Niederkrotenthaler, T., Chen, J., Di Natale, A., ... & Ten, A. (2020). A structured open dataset of government interventions in response to COVID-19. *medRxiv*.
- Documenting the Now. (2020). Hydrator [Computer Software]. Retrieved from <https://github.com/docnow/hydrator>
- Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, 20(5), 533-534.

- Gallotti, R., Castaldo, N., Valle, F., Sacco, P., & De Domenico, M. (2020). Covid19 infodemics observatory. DOI, 10, 17605.
- Hagberg, A., Schult, D., & Swart, P. (2012). NetworkX Reference. Python Package.
- Hasell, J., Mathieu, E., Beltekian, D., Macdonald, B., Giattino, C., Ortiz-Ospina, E., ... & Ritchie, H. (2020). A cross-country database of COVID-19 testing. *Scientific data*, 7(1), 1-7.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(3), 90-95.
- Huynh, D. (2011). Google Refine Tutorial. In *Computer Assisted Reporting Conference*, Raleigh, NC, USA.
- Isoni, A. (2016). *Machine learning for the web*. Packt Publishing Ltd.
- Jacomy, M., Girard, P., Ooghe-Tabanou, B., & Venturini, T. (2016, March). Hyphe, a curation-oriented approach to web crawling for the social sciences. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 10, No. 1).
- Lamsal, R. (2020). Coronavirus (COVID-19) Tweets Dataset. IEEE Dataport.
- Levin, G., & Dorsey, B. *Image Processing and Computer Vision*.
- Mauri, M., Elli, T., Caviglia, G., Ubaldi, G., & Azzi, M. (2017, September). RAWGraphs: a visualisation platform to create open outputs. In *Proceedings of the 12th biannual conference on Italian SIGCHI chapter* (pp. 1-5).
- Memon, S. A., & Carley, K. M. (2020). CMU-MisCov19: A Novel Twitter Dataset for Characterizing COVID-19 Misinformation [Data set]. Presented at the 5th International Workshop on Mining Actionable Insights from Social Networks (MAISoN) at CIKM 2020, Online: Zenodo. <http://doi.org/10.5281/zenodo.4024154>
- Memon, S. A., & Carley, K. M. (2020b). Characterizing covid-19 misinformation communities using a novel twitter dataset. arXiv preprint arXiv:2008.00791 <https://arxiv.org/abs/2008.00791>
- Open COVID-19 Data Working Group. (2020). Detailed Epidemiological Data from the COVID-19 Outbreak. Accessed on, 06-16.
- Orduña-Malea, E., Font-Julián, C. I., & Ontalba-Ruipérez, J. A. (2020). Covid-19: análisis métrico de vídeos y canales de comunicación en YouTube. *El profesional de la información (EPI)*, 29(3).
- Pepe, E., Bajardi, P., Gauvin, L., Privitera, F., Lake, B., Cattuto, C., & Tizzoni, M. (2020). COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown. *Scientific data*, 7(1), 1-7.
- Puckette, M. S. (1997, September). Pure data. In *ICMC*.
- Reas, C., & Fry, B. (2015). *Getting Started with Processing: A Hands-on introduction to making interactive Graphics*. Maker Media, Inc..

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

Rieder, B., Matamoros-Fernández, A., & Coromina, Ò. (2018). From ranking algorithms to 'ranking cultures' Investigating the modulation of visibility in YouTube search results. *Convergence*, 24(1), 50-68.

Roberts, Adam, et al. "Magenta studio: Augmenting creativity with deep learning in ableton live." (2019).

Rogers, R. (2013). *Digital methods*. MIT press.

Ryerson University Social Media Lab; The International Federation of Medical Students' Associations (2020). COVID-19 Fact-checkers Dataset. Scholars Portal Dataverse [online] Available at: <<https://doi.org/10.5683/SP2/IMISPE>> [Accessed 4 January 2021].

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn: Machine learning without learning the machinery. *GetMobile: Mobile Computing and Communications*, 19(1), 29-33.

Yu, J. (2020). Open access institutional and news media tweet dataset for COVID-19 social science research. arXiv preprint arXiv:2004.01791.

8 Abbreviation List

AI	Artificial Intelligence
API	Application Programming Interface
CDC	Centers for Disease Control and prevention
CSV	Comma-Separated Values
EC	European Commission
ECDC	European Centre for Disease prevention and Control
EUT	Eurecat
IRCAM	Institut de Recherche et Coordination Acoustique/Musique
JSON	JavaScript Object Notation
KCL	King's College London
ODI	Open Data Institute
URL	Uniform Resource Locator
WHO	World Health Organization

9 More information about this document

Project acronym	MediaFutures
Project full title	MediaFutures, Data-driven innovation hub for the media value chain
Grant Agreement no	951962
Deliverable number	D4.3

Deliverable title	Data, tools and infrastructure for experimental support
Deliverable nature	Report
Dissemination level	Public
Work package and Task	WP4, T4.2
Contractual delivery date	28 February 2021
Actual delivery date	28 February 2021
Authors	Pablo Aragón, EUT Julian Vicens, EUT Elena Simperl, KCL Vicky Hallam, ODI
Reviewers	Emanuele Camarda, LUISS Hugues Vinet, IRCAM

Revision History

Version	Date	Name
0.8	18.01.2021	EUT, KCL, ODI
0.9	01.02.2021	EUT, KCL, ODI
1.0	15.02.2021	EUT



MediaFutures