News articles and front pages from 19 Swedish news sites during the covid-19/corona pandemic 2020–2021

Peter M. Dahlgren

2021-10-12

Abstract

This dataset contains news articles from Swedish news sites during the covid-19 corona pandemic 2020–2021. The purpose was to develop and test new methods for collection and analyses of large news corpora by computational means. In total, there are 677,151 articles collected from 19 news sites during 2020-01-01 to 2021-04-26. The articles were collected by scraping all links on the homepages and main sections of each site every two hours, day and night.

The dataset also includes about 45 million timestamps at which the articles were present on the front pages (homepages and main sections of each news site, such as domestic news, sports, editorials, etc.). This allows for detailed analysis of what articles any reader likely was exposed to when visiting a news site. The time resolution is (as stated previously) two hours, meaning that you can detect changes in which articles were on the front pages every two hours.

The 19 news sites are aftonbladet.se, arbetet.se, da.se, di.se, dn.se, etc.se, expressen.se, feministisktperspektiv.se, friatider.se, gp.se, nyatider.se, nyheteridag.se, samnytt.se, samtiden.nu, svd.se, sverigesradio.se, svt.se, sydsvenskan.se and vlt.se.

Due to copyright, the full text is not available but instead transformed into a *document-term matrix* (in long format) which contains the frequency of all words for each article (in total, 80 million words). Each article also includes extensive metadata that was extracted from the articles themselves (URL, document title, article heading, author, publish date, edit date, language, section, tags, category) and metadata that was inferred by simple heuristic algorithms (page type, article genre, paywall).

Overview

Title:	News articles and front pages from 19 Swedish news sites during the covid-19/corona pandemic 2020–2021
Swedish title:	Nyhetsartiklar och förstasidor från 19 svenska nyhetssajter under coronapandemin 2020-2021
Resarcher:	Peter M. Dahlgren, peter.dahlgren@jmg.gu.se (https://peterdahlgren.com/)
Department:	Department of Journalism, Media and Communication (JMG), University of Gothenburg (https://www.gu.se/jmg)
Research project:	Kriskommunikation och samhällsförtroende i det multipublika samhället (KRISAMS, https://www.gu.se/forskning/krisams)
Funder:	The Swedish Civil Contingencies Agency (MSB) (https://www.msb.se/)
Purpose:	Develop and test new methods for collection and analyses of large news corpora by computational means
Subject:	social science, journalism, media, communication
Keywords:	covid-19, corona, pandemic, news articles, news values, front pages, journalism
Geography:	Sweden
Design:	Longitudinal
Time period:	2020-01-01 to 2021-04-26
Observations:	677,151 articles + 45,337,740 front page timestamps + 80,090,784 words
Data format:	CSV
Data license:	Creative Commons Attribution 4.0 International (CC BY 4.0)
DOI:	https://doi.org/10.5878/d18f-q220

Sharing and data availability

Data license

Publications that use this data

- Dahlgren, P. M. (2021). *Medieinnehåll och mediekonsumtion under coronapandemin: Datoriserade metoder för insamling och analys av stora mängder text- och mediedata*. Göteborg: Institutionen för journalistik, medier och kommunikation (JMG), Göteborgs universitet.
- Dahlgren, P. M. (2021). Svenskar eller utrikesfödda i medierna? – att identifiera födelseland från namn. I *L. Truedson & J. Lundqvist (Red.), Vitt eller brett? – vilka får ta plats i medier och på redaktioner*. Stockholm: Institutet för mediestudier.

Recommended citation

APA6:

> Dahlgren, P. M. (2021). *News articles and front pages from 19 Swedish news sites during the covid-19/corona pandemic 2020–2021*. Swedish National Data Service. doi:10.5878/d18f-q220

See next page for BibTeX.

BibTeX:

```
@misc{dahlgren_news_2021,
  title = {News articles and front pages from 19 Swedish news sites during
           the covid-19/corona pandemic 2020-2021},
  url = {https://doi.org/10.5878/d18f-q220},
  abstract = {This dataset contains news articles from Swedish news sites during the
              covid-19 corona pandemic 2020--2021. The purpose was to develop and
              test new methods for collection and analyses of large news corpora by
              computational means. In total, there are 677,151 articles collected
              from 19 news sites during 2020-01-01 to 2021-04-26. The articles were
              collected by scraping all links on the homepages and main sections of
              each site every two hours, day and night.  The dataset also includes
              about 45 million timestamps at which the articles were present on the
              front pages (homepages and main sections of each news site, such as
              domestic news, sports, editorials, etc.). This allows for detailed
              analysis of what articles any reader likely was exposed to when
              visiting a news site. The time resolution is (as stated previously)
              two hours, meaning that you can detect changes in which articles were
              on the front pages every two hours.  The 19 news sites are
              aftonbladet.se, arbetet.se, da.se, di.se, dn.se, etc.se, expressen.se,
              feministisktperspektiv.se, friatider.se, gp.se, nyatider.se,
              nyheteridag.se, samnytt.se, samtiden.nu, svd.se, sverigesradio.se,
              svt.se, sydsvenskan.se and vlt.se.  Due to copyright, the full text is
              not available but instead transformed into a document-term matrix (in
              long format) which contains the frequency of all words for each
              article (in total, 80 million words). Each article also includes
              extensive metadata that was extracted from the articles themselves
              (URL, document title, article heading, author, publish date, edit
              date, language, section, tags, category) and metadata that was
              inferred by simple heuristic algorithms (page type, article genre,
              paywall). The data is licensed with Creative Commons Attribution 4.0
              International (CC BY 4.0).},
  language = {Swedish},
  publisher = {Swedish National Data Service},
  author = {Dahlgren, Peter M.},
  year = {2021}
}
```

Method of data collection

Web scraping news sites

An open source web scraper called *Mechanical News* was developed to scrape the news articles from 19 news sites.[1] A list of all news sites is available in Table 2.

How and how often

The web scraper scanned the front pages (homepage and all the main section pages such domestic news, sports, editorials etc.) of each news site.

The web scraper followed all links on the front pages (i.e., `<a href="">`). If the link was determined to be an article, information about the article was extracted and saved to a database (including so called timestamps, described below). This process was repeated every two hours, day and night, from 2020-01-01 to 2021-04-26.

Articles collected

In total, 677,151 articles were collected. Note, however, that this figure includes *all* pages found: the homepages, section pages, pages such as "about" and "contact" etc.—all of which are demonstrably *not* articles. These pages can partly be identified via the `page_type` variable found in the dataset. These pages comprise a very small proportion of the total, and they also lack the extensive metadata that articles have. For simplicity's sake, however, all observations are called articles.

Paywalls

All articles were collected, both those locked behind a paywall and those unlocked. But the scraper could not get behind paywalls and therefore only collected the full text from articles that were unlocked.

However, many news sites published their news articles unlocked at first, and only later put them behind a paywall. It is therefore likely that many articles managed to be collected by the web scraper just before they were locked.

Front page timestamps collected

In addition to each article, the date and time when each article was found on each front page (homepage and main section) was also collected.

For example, if a journalist put an article on the front page of their news site for one week, this means that a timestamp will be present every two hours for a week, including the origin URL of the news site and the destination ID of the article.

This is the largest data with 45,337,740 observations.

Missing data

Any gaps in the data collection are primarily due to server or internet connection issues. These have not been tracked systematically, but can be identified through gaps in the front page timestamps.

Any missing data on article level is primarily due to changes in the HTML structure of each news site that happens without prior notice. Similarly, these have not been tracked systematically.

[1] Code in Python for the web scraper is available at https://github.com/peterdalle/mechanicalnews.

News sites scraped

Table 2: Number of articles collected from each news source.

#	Source name	Start collection	End collection	Articles	Homepage
1	aftonbladet	2020-01-01	2021-04-26	72,137	https://www.aftonbladet.se/
2	arbetet	2020-04-06	2021-04-23	1,934	https://arbetet.se/
3	dagensarbete	2020-04-06	2021-04-26	380	https://da.se/
4	dagensindustri	2020-04-06	2021-04-26	31,987	https://www.di.se/
5	dagensnyheter	2020-01-01	2021-04-25	46,456	https://www.dn.se/
6	etc	2020-04-06	2021-04-24	5,184	https://www.etc.se/
7	expressen	2020-04-06	2021-04-26	70,945	https://www.expressen.se/
8	feministisktperspektiv	2020-04-06	2021-04-22	606	https://feministisktperspektiv.se
9	friatider	2020-04-06	2021-04-24	5,429	https://www.friatider.se/
10	goteborgsposten	2020-04-06	2021-04-25	59,454	https://www.gp.se/
11	nyatider	2020-04-06	2021-04-24	1,846	https://nyatider.se/
12	nyheteridag	2020-04-06	2021-04-24	2,703	https://nyheteridag.se/
13	samhallsnytt	2020-04-06	2021-04-24	3,795	https://samnytt.se/
14	samtiden	2020-04-06	2021-04-24	1,538	https://samtiden.nu/
15	svenskadagbladet	2020-04-06	2021-04-25	37,224	https://www.svd.se/
16	sverigesradio	2020-04-06	2021-04-26	137,359	https://sverigesradio.se/
17	sverigestelevision	2020-01-01	2021-04-25	106,579	https://www.svt.se/
18	sydsvenskan	2020-04-06	2021-04-25	66,176	https://www.sydsvenskan.se/
19	vlt	2020-04-06	2021-04-25	25,419	https://www.vlt.se/

Data

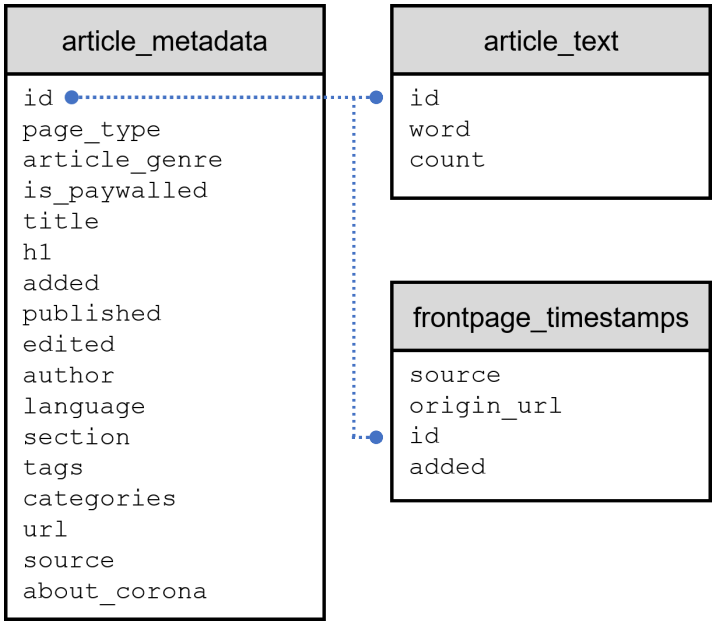

Figure 1: Diagram of the relationships between data structures, where `id` is the identifier (and foreign key) of each article. The data are stored in separate CSV files.

Article metadata

The article metadata is stored in `article_metadata.csv` (UTF-8 character encoding). The file contains information about each news article, one article per row. In total, there are 677,151 observations and 17 variables.

Table 3: Data structure of `article_metadata.csv`.

Variable	Data type	Description
`id`	numeric	Arbitrary ID identifying the news article.
`page_type`	numeric	Type of page (e.g. page, article, video) as indicated by a number, see details on next page.
`article_genre`	numeric	Article genre (e.g., news, sports, editorial) as indicated by a number, see details on next page.
`is_paywalled`	numeric	Whether the article is behind a paywall (`1`) or not (`0`).
`title`	character	Title of the web page extracted from <title>.
`h1`	character	Heading of the article extracted from <h1> or similar.
`added`	datetime	When the article was retrieved by the web scraper (in format `yyyy-MM-dd hh:mm:ss`).
`published`	datetime	When the article was published according to the news site (`yyyy-MM-dd hh:mm:ss`).
`edited`	datetime	When the article was edited according to the news site (`yyyy-MM-dd hh:mm:ss`).
`author`	character	Author(s) of the news article, with a new line separating each new author.
`language`	character	Language of the news article (e.g. `sv`, `sv-SE`).
`section`	character	Section of the article, according to the news site. Each new section is separated by a new line.
`tags`	character	Tags attached to the article, as tagged by the news site. Each new tag is separated by a new line.
`categories`	character	Categories of the news article, as categorized by the news site. Each new category is separated by a new line.
`url`	character	URL to the article.
`source`	character	One of 19 labels identifying the source of the article (see source names in Table 2).
`about_corona`	numeric	Whether the articles mentions "corona" or "covid-19" in `title`, `h1` or article text (`1`) or not (`0`). About 25% of the articles are about corona.

Note: Missing values are denoted by the value `NA`, which primarily affects `edited` and `published`.

Three variables were inferred by the web scraper software Mechanical News[2] that used simple heuristic algorithms for identification. More precisely:

1. `page_type` was primarily identified by keywords in the URL and text present within the page structure of the news article.
2. `article_genre` was primarily identified by keywords in the URL and text present within the page structure of the news article.
3. `language` was firstly identified by the language code within the article HTML metadata, and secondly through a language classifier in Python.

Corresponding labels of `page_type` values

`0` = generic web page
`1` = specific article
`2` = list or collection of articles (as found on frontpages, subsections, categories, themes, tags etc.)
`3` = sound
`4` = video

Corresponding labels of `article_genre` values

`0` = generic web page
`1` = news article
`2` = editorial, news columnist, chronical
`3` = opinion piece, debate article, letter to the editor
`4` = sports
`5` = advertisement, native ad
`6` = entertainment: culture, movies, books, events, life style
`7` = technology: cars, boats, machines, digital, science, tech
`8` = personal: human interest, family, relationships, portraits, health
`9` = economy: world/national economy, business, industry, personal finance

Preview of the first ten rows of `article_metadata.csv`:

```
"id","page_type","article_genre","is_paywalled","title","h1","added","publishe...
3818643,1,1,0,"I migranternas spår","I migranternas spår","2021-04-26 10:26:34...
3818513,0,4,1,"Över 100 profiler tippar finalen mellan Timrå IK och Björklöven...
3818457,4,4,0,"Häckens tredje raka förlust - föll hemma mot Sirius | Highlight...
3818440,1,1,0,"Knölvalen på Öland var nästan nio meter lång och några år gamma...
3818434,1,6,0,"Oscarsgalan 2021: Här kan du streama de nominerade filmerna","S...
3818250,1,1,0,"Rumbutis med bästa svenska placeringen på 27 år - Radiosporten"...
3817575,1,1,0,"Hade explosiva ämnen i studentlägenheten - P4 Uppland","Hade ex...
p4uppland@sverigesradio.se","sv","","","Brott","https://sverigesradio.se/artik...
3817245,4,0,0,"Agendan: Rapporter och räntebesked i fokus - Di TV","","2021-04...
```

Note: The width of the preview is cut by `...` to fit the page.

[2] https://github.com/peterdalle/mechanicalnews.

Article text

The article text is stored in `article_text.csv` (UTF-8). The file contains the `id` of each news article and how many times (`count`) a specific `word` occurs in the news article. The file contains 80,090,784 observations and 3 variables in long format.[3] That is, the `id` variable is repeated for each new `word`.

Table 4: Data structure of `article_text.csv`.

Variable	Data type	Description
`id`	numeric	Arbitrary ID identifying the news article
`word`	character	Word/token found in the news article (converted to lower case)
`count`	numeric	Number of times the word was found in the news article

Preview of the first ten rows of `article_text.csv`:

```
"id","word","count"
3818643,"aldaw",1
3818643,"adem",1
3585168,"adem",1
3576334,"adem",2
3513977,"adem",4
3511142,"adem",2
3096944,"adem",1
2757817,"adem",1
2727200,"adem",1
```

The article texts were tokenized into words/tokens using the following rules:[4]

- keep symbols
- keep numbers
- keep URLs
- keep hyphens
- remove punctuation
- remove separators

[3] In wide format, the data is more than 99.99% sparse, meaning that most words are unique to each article.
[4] R package *quanteda* method `tokens()` was used for tokenization, see https://quanteda.io/reference/tokens.html.

Front page timestamps

The timestampss are stored in `frontpage_timestamps.csv` (UTF-8). The file contains when each news article was found on the front page (homepage and main sections) of the news sites. The file contains 45,337,740 observations and 4 variables in long format.

Table 5: Data structure of `frontpage_timestamps.csv`.

Variable	Data type	Description
`source`	numeric	Label that identifies the news site (see source names in Table 2)
`origin_url`	character	URL from which the news article was found
`id`	numeric	ID identifying the news article that was found on `origin_url`
`added`	datetime	Datetime (`yyyy-MM-dd hh:mm:ss`) when the article was found on `origin_url`

Preview of the first ten rows of `frontpage_timestamps.csv`:

```
"source","origin_url","id","added"
"aftonbladet","https://www.aftonbladet.se/nojesbladet",3815121,2021-04-25 22:00:50
"aftonbladet","https://www.aftonbladet.se/nojesbladet",3814476,2021-04-25 20:01:15
"aftonbladet","https://www.aftonbladet.se/nojesbladet",3814476,2021-04-25 22:00:50
"sverigesradio","https://sverigesradio.se/ostergotland/",3813411,2021-04-25 17:27:19
"sverigesradio","https://sverigesradio.se/ostergotland/",3813411,2021-04-25 19:27:38
"sverigesradio","https://sverigesradio.se/ostergotland/",3813411,2021-04-25 21:26:08
"dagensindustri","https://www.di.se/",3813016,2021-04-25 16:01:37
"dagensindustri","https://www.di.se/bil/",3813016,2021-04-25 16:01:47
"dagensindustri","https://www.di.se/",3813016,2021-04-25 18:00:16
```

Example

If we take the rows with `sverigesradio` as an example above, the article with ID `3813411` was found on `https://sverigesradio.se/ostergotland/` three times: At `2021-04-25 17:27:19`, `2021-04-25 19:27:38` and `2021-04-25 21:26:08`, respectively. We can thus infer that it was present at least five hours on that particular page.

We can then lookup what article `3813411` refers to by consulting the `article_metadata.csv` file and see the title etc. of the news article. And if we want to know which words were used in the article we consult the `article_text.csv` file.

Code to get going

Here's some code in R to make it easier for you to get the data into your environment.

The main thing you need to know is that articles are identified by the `id` variable in all datasets.

```
library(tidyverse)
library(lubridate)

# Load data
words <- read.csv("article_text.csv", fileEncoding = "UTF-8")
metadata <- read.csv("article_metadata.csv", fileEncoding = "UTF-8")
timestamps <- read.csv("frontpage_timestamps.csv", fileEncoding = "UTF-8")

# Labels for corresponding values
page_type <- c("0" = "Generic web page",
               "1" = "Specific article",
               "2" = "List/collection of articles",
               "3" = "Sound",
               "4" = "Video")

article_genre <- c("0" = "Generic web page",
                   "1" = "News article",
                   "2" = "Editorial",
                   "3" = "Opinion",
                   "4" = "Sports",
                   "5" = "Advert",
                   "6" = "Entertainment",
                   "7" = "Technology",
                   "8" = "Personal",
                   "9" = "Economy")

# Convert to appropriate data types
timestamps$added <- as_datetime(timestamps$added)
metadata$is_paywalled <- metadata$is_paywalled == 1
metadata$added <- as_datetime(metadata$added)
metadata$edited <- as_datetime(metadata$edited)
metadata$published <- as_datetime(metadata$published)
metadata$source <- as.factor(metadata$source)
metadata$section <- as.factor(metadata$section)
metadata$language <- as.factor(metadata$language)
metadata$article_genre <- factor(metadata$article_genre, levels=0:9,
                                 labels = article_genre, ordered = FALSE)
metadata$page_type <- factor(metadata$page_type, levels=0:4,
                             labels = page_type, ordered = FALSE)

# Combine metadata + text/words into long format (slow)
articles <- metadata %>%
  left_join(words, by="id")
```