Vluchtige tweets en honderden likes

Sociale media-archivering en het gebruik ervan door onderzoekers

Anne Helmond (Universiteit van Amsterdam).

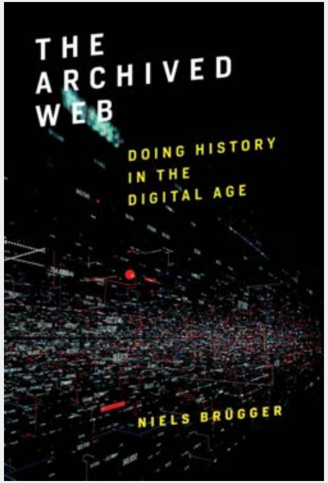
‘Webinar Sociale media-archivering?,’ Netwerk Digitaal Erfgoed, 25 juni 2020.

De onderzoeksmogelijkheden van archieven en platformen

Introductie

- Archieven en archiveringspraktijken ‘vormen’ het corpus en beïnvloeden de soorten vragen die gesteld kunnen worden (zie Brügger, 2005; 2019).
- Sociale media platformen vormen ook het corpus door hun sterk gecontroleerde archiveringsmogelijkheden (zie Helmond en van der Vlist, 2019).
    - Technische controle: beperkte toegang tot een selectie van data
    - Contractuele controle: gebruik en hergebruik van data

Vormen van webarchivering



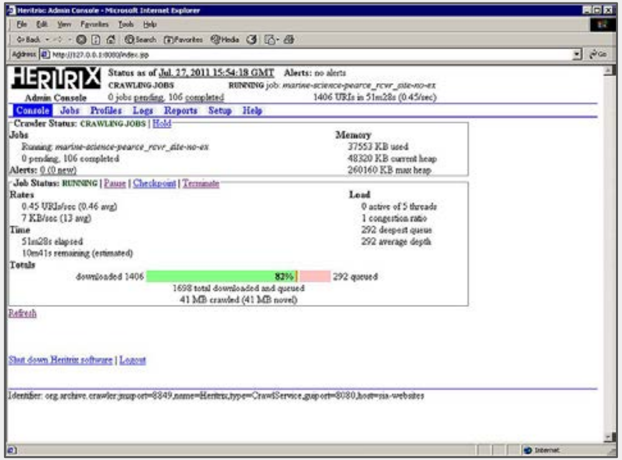
1. Making an image
2. Making a screen movie
3. Downloading individual files
4. Web crawling
5. Collecting web material from a database via an API
6. Collecting the web that has been taken off-line and preserved unchanged
7. Collecting the web as presented in other media types

(Brügger, 2018: 80).

Web crawling

- **Crawl**: A web archiving (or "capture") operation that is conducted by an automated agent, called a crawler, a robot, or a spider. Crawls identify materials on the live web that belong in your collections, based upon your choice of seeds and scope. Crawl can also reference the archived content associated with the action.
- **Crawler**: Explores the web and collects data about its contents. A crawler can also be configured to capture web-based resources. It starts a capture process from a seed list of entry-point URLs (EPUs).

Glossary of Archive-It and Web Archiving Terms


Heritrix: Admin Console - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://127.0.0.1:8080/index.jsp
Go
HERITRIX
Admin Console
Console Jobs Profiles Logs Reports Setup Help
Internet

Status as of Jul. 27, 2011 15:54:18 GMT Alerts: no alerts
CRAWLING JOBS RUNNING job: *marine-science-pearce_rcvr_site-no-ex*
0 jobs pending, 106 completed 1406 URIs in 51m28s (0.45/sec)

Crawler Status: CRAWLING JOBS | Hold

Jobs
Running: *marine-science-pearce_rcvr_site-no-ex*
0 pending, 106 completed
Alerts: 0 (0 new)

Memory
37553 KB used
48320 KB current heap
260160 KB max heap

Job Status: RUNNING | Pause | Checkpoint | Terminate

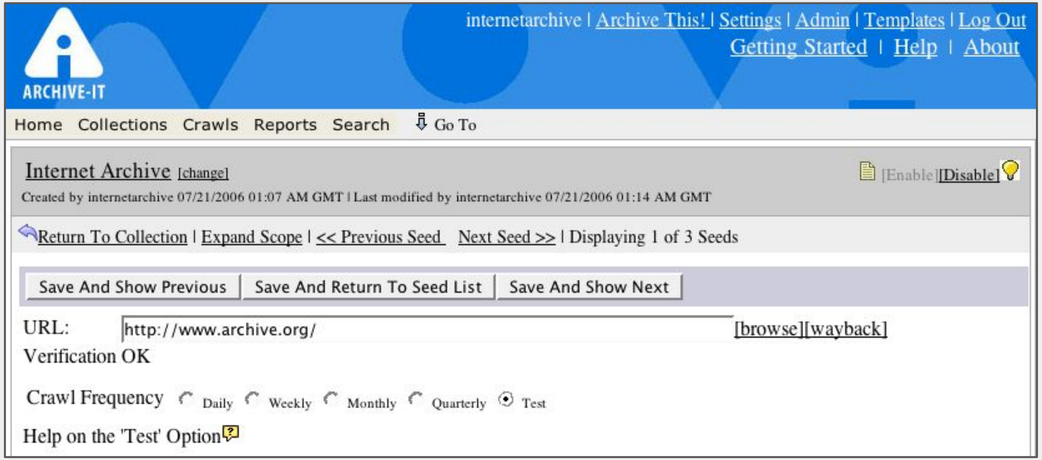
Rates
0.45 URIs/sec (0.46 avg)
7 KB/sec (13 avg)
Time
51m28s elapsed
10m41s remaining (estimated)
Totals
downloaded 1406 82% 292 queued
1698 total downloaded and queued
41 MB crawled (41 MB novel)

Load
0 active of 5 threads
1 congestion ratio
292 deepest queue
292 average depth

Refresh

Shut down Heritrix software | Logout

Identifier: org.archive.crawler:jmxport=8849,name=Heritrix,type=CrawlService,guiport=8080,host=sia-websites


ARCHIVE-IT
internetarchive | Archive This! | Settings | Admin | Templates | Log Out
Getting Started | Help | About
Home Collections Crawls Reports Search Go To
[Enable][Disable]
Save And Show Previous
Save And Return To Seed List
Save And Show Next
[browse][wayback]

Internet Archive [change]
Created by internetarchive 07/21/2006 01:07 AM GMT | Last modified by internetarchive 07/21/2006 01:14 AM GMT

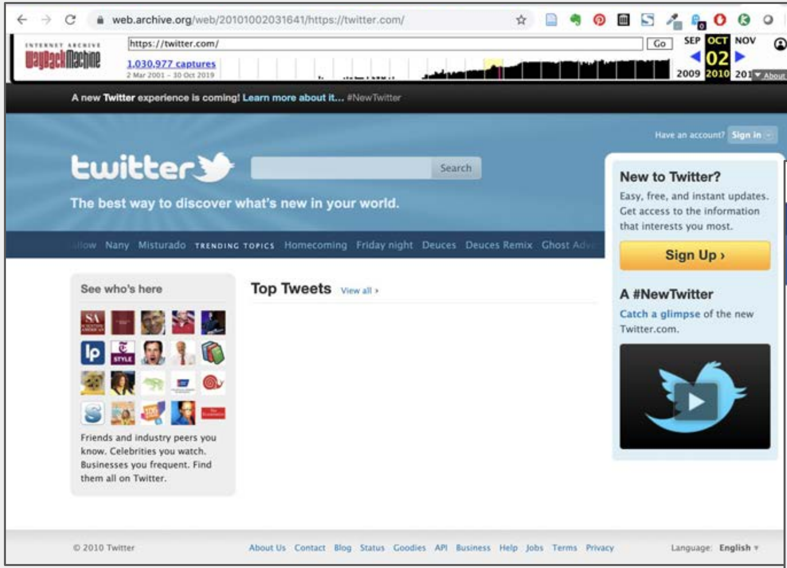
Return To Collection | Expand Scope | << Previous Seed Next Seed >> | Displaying 1 of 3 Seeds

URL: http://www.archive.org/
Verification OK

Crawl Frequency ○ Daily ○ Weekly ○ Monthly ○ Quarterly ● Test

Help on the 'Test' Option

Het crawlen van sociale media


web.archive.org/web/20101002031641/https://twitter.com/
INTERNET ARCHIVE WaybackMachine
https://twitter.com/
Go
1,030,977 captures
2 Mar 2001 - 30 Oct 2019
SEP OCT NOV
02
2009 2010 201
About this capture
Have an account? Sign in
twitter
Search
Sign Up ›
Top Tweets View all ›
See who's here
Nany Misturado TRENDING TOPICS Homecoming Friday night Deuces Deuces Remix Ghost Adv
About Us Contact Blog Status Goodies API Business Help Jobs Terms Privacy
Language: English

A new Twitter experience is coming! Learn more about it... #NewTwitter

The best way to discover what's new in your world.

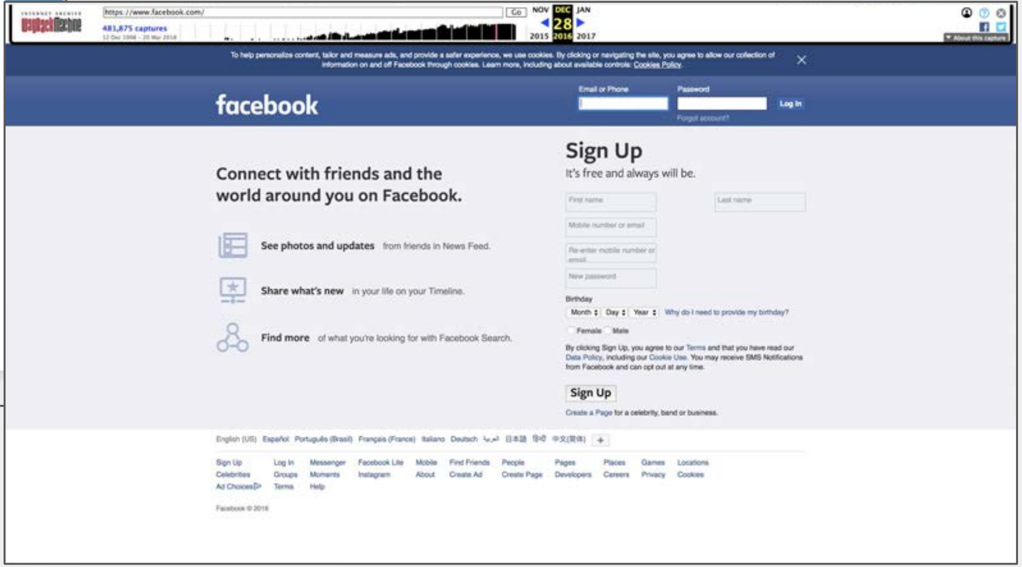
New to Twitter?
Easy, free, and instant updates. Get access to the information that interests you most.

A #NewTwitter
Catch a glimpse of the new Twitter.com.

Friends and industry peers you know. Celebrities you watch. Businesses you frequent. Find them all on Twitter.

© 2010 Twitter

https://twitter.com/


INTERNET ARCHIVE WaybackMachine
https://www.facebook.com/
Go
481,875 captures
12 Dec 1998 - 20 Mar 2018
NOV DEC JAN
28
2015 2016 2017
About this capture
facebook
Email or Phone
Password
Log In
Forgot account?
First name
Last name
Mobile number or email
Re-enter mobile number or email
New password
Birthday
Month Day Year
○ Female ○ Male
Sign Up
English (US) Español Português (Brasil) Français (France) Italiano Deutsch العربية 日本語 हिंदी 中文(简体) +
Sign Up Log In Messenger Facebook Lite Mobile Find Friends People Pages Places Games Locations
Celebrities Groups Moments Instagram About Create Ad Create Page Developers Careers Privacy Cookies
Ad Choices Terms Help

To help personalize content, tailor and measure ads, and provide a safer experience, we use cookies. By clicking or navigating the site, you agree to allow our collection of information on and off Facebook through cookies. Learn more, including about available controls: Cookies Policy.

Connect with friends and the world around you on Facebook.

See photos and updates from friends in News Feed.

Share what's new in your life on your Timeline.

Find more of what you're looking for with Facebook Search.

Sign Up
It's free and always will be.

Why do I need to provide my birthday?

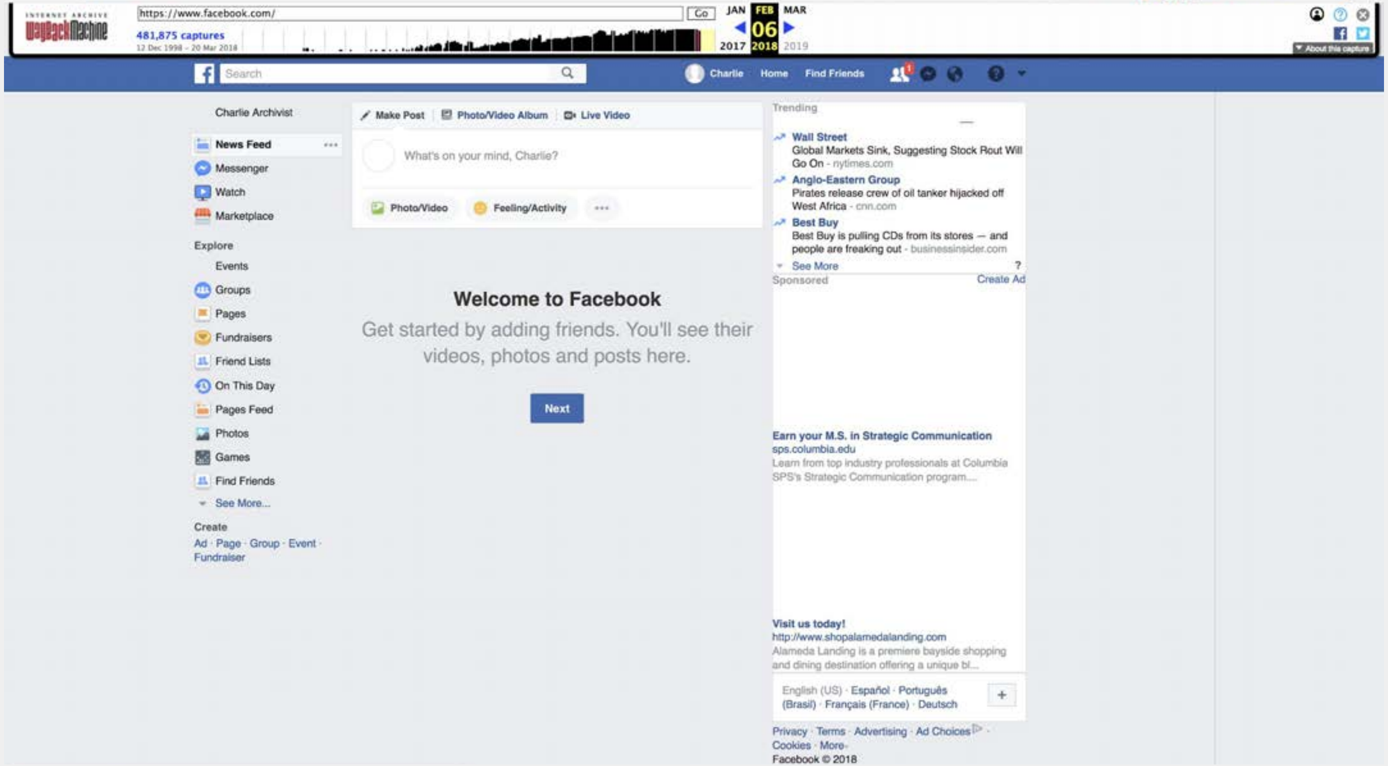
By clicking Sign Up, you agree to our Terms and that you have read our Data Policy, including our Cookie Use. You may receive SMS Notifications from Facebook and can opt out at any time.

Create a Page for a celebrity, band or business.

Facebook © 2016

https://facebook.com/

Crawler profielen (bijv. Charlie Archivist)



What's on your mind, Charlie?

Welcome to Facebook

Get started by adding friends. You'll see their videos, photos and posts here.

Wall Street
Global Markets Sink, Suggesting Stock Rout Will Go On - nytimes.com

Anglo-Eastern Group
Pirates release crew of oil tanker hijacked off West Africa - cnn.com

Best Buy
Best Buy is pulling CDs from its stores — and people are freaking out - businessinsider.com

Earn your M.S. in Strategic Communication
sps.columbia.edu
Learn from top industry professionals at Columbia SPS's Strategic Communication program....

Visit us today!
http://www.shopalamedalanding.com
Alameda Landing is a premiere bayside shopping and dining destination offering a unique bl...

Tools voor sociale media archivering

Voor **kleine** individuele projecten:

Conifer to ‘create high-fidelity, interactive captures’

- 👍Gratis account 5GB opslagruimte. Online interface. Opnemen en afspelen van websites. Standaard .warc file.
- 👎Files nemen veel ruimte in. 1 website tegelijk.

Conifer

Collect and revisit web pages.

Conifer is a web archiving service that creates an interactive copy of any web page that you browse, including content revealed by your interactions such as playing video and audio, scrolling, clicking buttons, and so forth.

Free accounts with 5GB of storage. Get more and support this project by becoming a supporter.

Conifer is an online service based on Webrecorder software.

Webrecorder ‘to capture and replay interactive websites’

- 👍Lokale versie van Conifer. Opnemen en afspelen van websites. Standaard .warc file.
- 👎Files nemen veel ruimte in. 1 website tegelijk.

Webrecorder

Web archiving for all!

Webrecorder provides a **suite of open source projects and tools** to capture interactive websites and replay them at a later time as **accurately as possible**.

Latest News

Read all about the next phase for Webrecorder, Rhizome's Conifer, and ReplayWeb.page!

Read about the collaboration with IIPC to support adoption of pywb

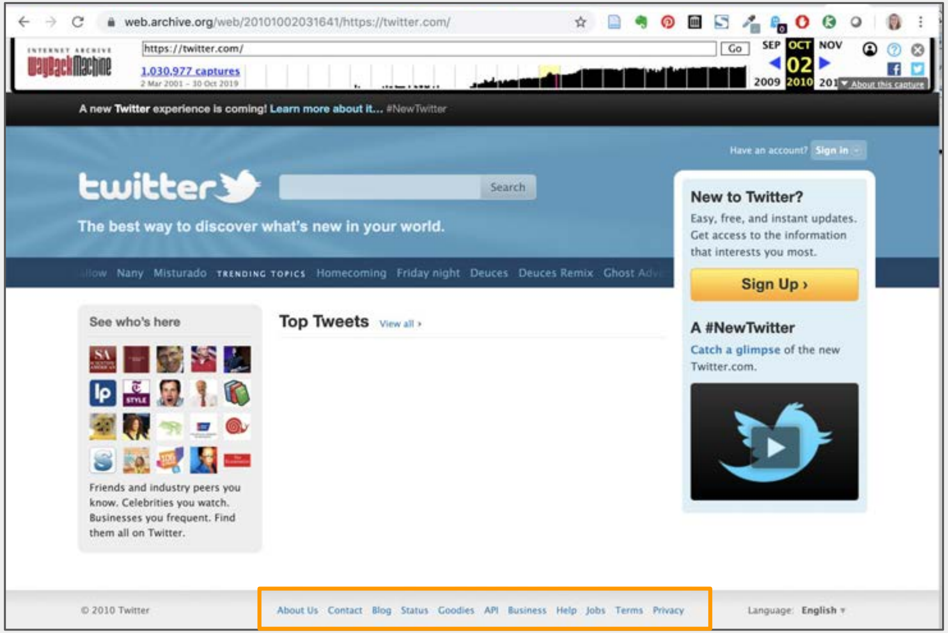
Featured Tool

REPLAYWEB.PAGE is a new Webrecorder tool to allow users to view web archives directly in the browser from anywhere, even when offline.

Sociale media archivering via crawlers

- De archivering van sociale media *via crawlers* levert een (leeg) **gebruikers blik** op het platform die verschillende vormen van **interface analyse** en **platform analyse** mogelijk maakt.
- De archivering van sociale media *via Conifer/Webrecorder* levert een **individuele gebruikers blik** die een **interface analyse** en **autobiografische analyse** mogelijk maakt.

De vele gebruikers van sociale media



web.archive.org/web/20101002031641/https://twitter.com/

INTERNET ARCHIVE WaybackMachine

https://twitter.com/ Go

1,030,977 captures
2 Mar 2001 - 30 Oct 2019

SEP OCT NOV
02
2009 2010 201

About this capture

A new Twitter experience is coming! Learn more about it... #NewTwitter

Have an account? Sign in

Search

The best way to discover what's new in your world.

llow Nany Misturado TRENDING TOPICS Homecoming Friday night Deuces Deuces Remix Ghost Adv

New to Twitter?

Easy, free, and instant updates. Get access to the information that interests you most.

Sign Up ›

A #NewTwitter

Catch a glimpse of the new Twitter.com.

See who's here

Friends and industry peers you know. Celebrities you watch. Businesses you frequent. Find them all on Twitter.

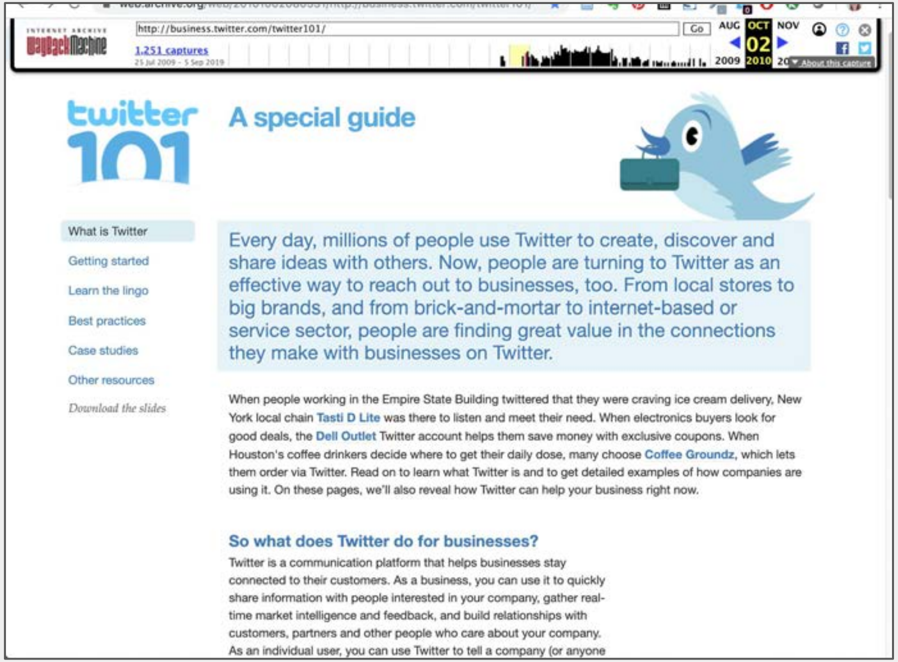
Top Tweets View all ›

© 2010 Twitter

About Us Contact Blog Status Goodies API Business Help Jobs Terms Privacy

Language: English

https://twitter.com/



INTERNET ARCHIVE WaybackMachine

http://business.twitter.com/twitter101/ Go

1,251 captures
25 Jul 2009 - 5 Sep 2019

AUG OCT NOV
02
2009 2010 20

About this capture

twitter 101

A special guide

- What is Twitter
- Getting started
- Learn the lingo
- Best practices
- Case studies
- Other resources
- *Download the slides*

Every day, millions of people use Twitter to create, discover and share ideas with others. Now, people are turning to Twitter as an effective way to reach out to businesses, too. From local stores to big brands, and from brick-and-mortar to internet-based or service sector, people are finding great value in the connections they make with businesses on Twitter.

When people working in the Empire State Building twittered that they were craving ice cream delivery, New York local chain Tasti D Lite was there to listen and meet their need. When electronics buyers look for good deals, the Dell Outlet Twitter account helps them save money with exclusive coupons. When Houston's coffee drinkers decide where to get their daily dose, many choose Coffee Groundz, which lets them order via Twitter. Read on to learn what Twitter is and to get detailed examples of how companies are using it. On these pages, we'll also reveal how Twitter can help your business right now.

So what does Twitter do for businesses?

Twitter is a communication platform that helps businesses stay connected to their customers. As a business, you can use it to quickly share information with people interested in your company, gather real-time market intelligence and feedback, and build relationships with customers, partners and other people who care about your company. As an individual user, you can use Twitter to tell a company (or anyone

http://business.twitter.com/

Typen sociale media gebruikers en hun bronnen

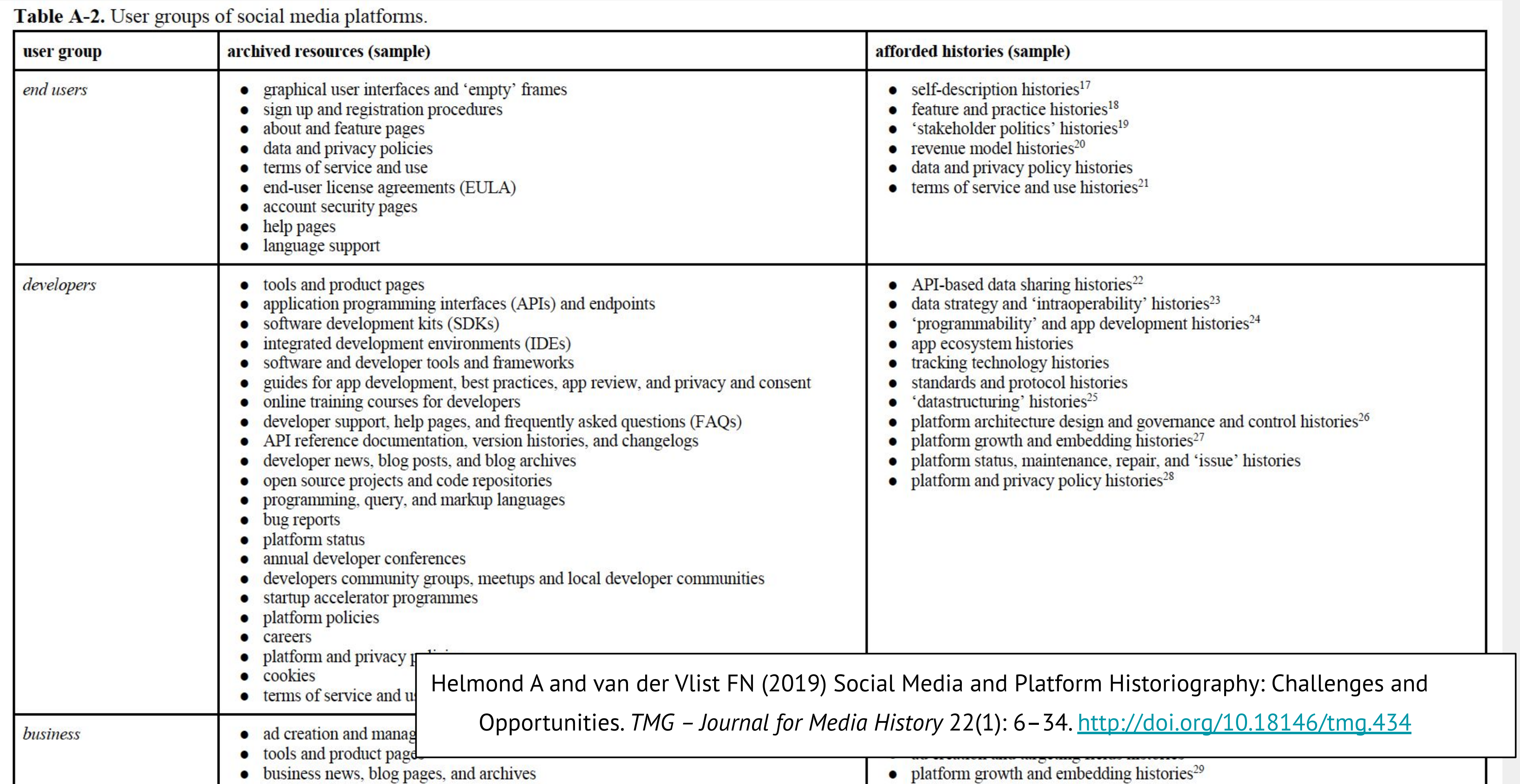


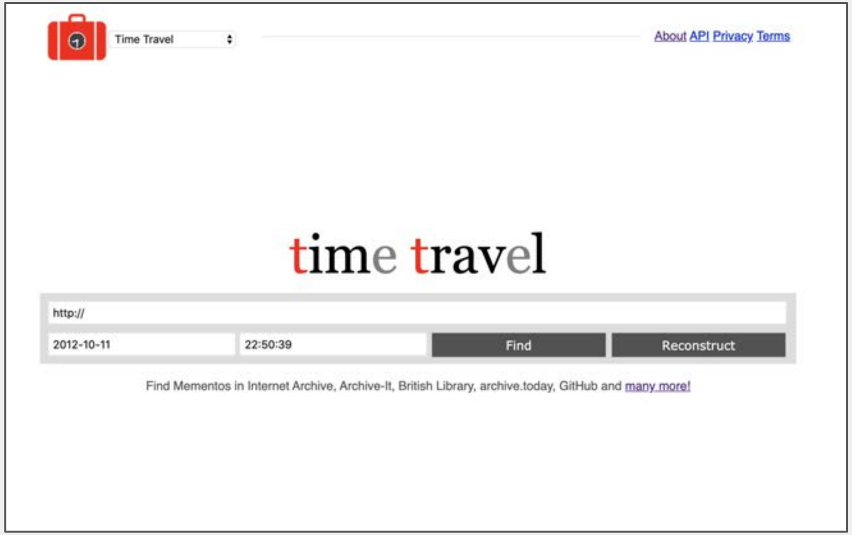
Table A-2. User groups of social media platforms.

user group	**archived resources (sample)**	**afforded histories (sample)**
end users	● graphical user interfaces and 'empty' frames ● sign up and registration procedures ● about and feature pages ● data and privacy policies ● terms of service and use ● end-user license agreements (EULA) ● account security pages ● help pages ● language support	● self-description histories[17] ● feature and practice histories[18] ● 'stakeholder politics' histories[19] ● revenue model histories[20] ● data and privacy policy histories ● terms of service and use histories[21]
developers	● tools and product pages ● application programming interfaces (APIs) and endpoints ● software development kits (SDKs) ● integrated development environments (IDEs) ● software and developer tools and frameworks ● guides for app development, best practices, app review, and privacy and consent ● online training courses for developers ● developer support, help pages, and frequently asked questions (FAQs) ● API reference documentation, version histories, and changelogs ● developer news, blog posts, and blog archives ● open source projects and code repositories ● programming, query, and markup languages ● bug reports ● platform status ● annual developer conferences ● developers community groups, meetups and local developer communities ● startup accelerator programmes ● platform policies ● careers ● platform and privacy p ● cookies ● terms of service and u	● API-based data sharing histories[22] ● data strategy and 'intraoperability' histories[23] ● 'programmability' and app development histories[24] ● app ecosystem histories ● tracking technology histories ● standards and protocol histories ● 'datastructuring' histories[25] ● platform architecture design and governance and control histories[26] ● platform growth and embedding histories[27] ● platform status, maintenance, repair, and 'issue' histories ● platform and privacy policy histories[28]
business	● ad creation and manag ● tools and product page ● business news, blog pages, and archives	● ad creation and targeting fields histories ● platform growth and embedding histories[29]

Helmond A and van der Vlist FN (2019) Social Media and Platform Historiography: Challenges and Opportunities. *TMG – Journal for Media History* 22(1): 6–34. http://doi.org/10.18146/tmg.434

Beschikbaarheid materialen: Memento & Memgator

- De beschikbaarheid van sociale media platform URLs voor alle gebruikersgroepen opvragen in 20+ web archieven
  - facebook.com
  - developers.facebook.com
  - facebook.com/business
  - facebook.com/politics
  - facebook.com/creators
  - etc
- Tools: Memento Time Travel API via Memgator `(Alam & Nelson, 2016).`



http://

2012-10-11

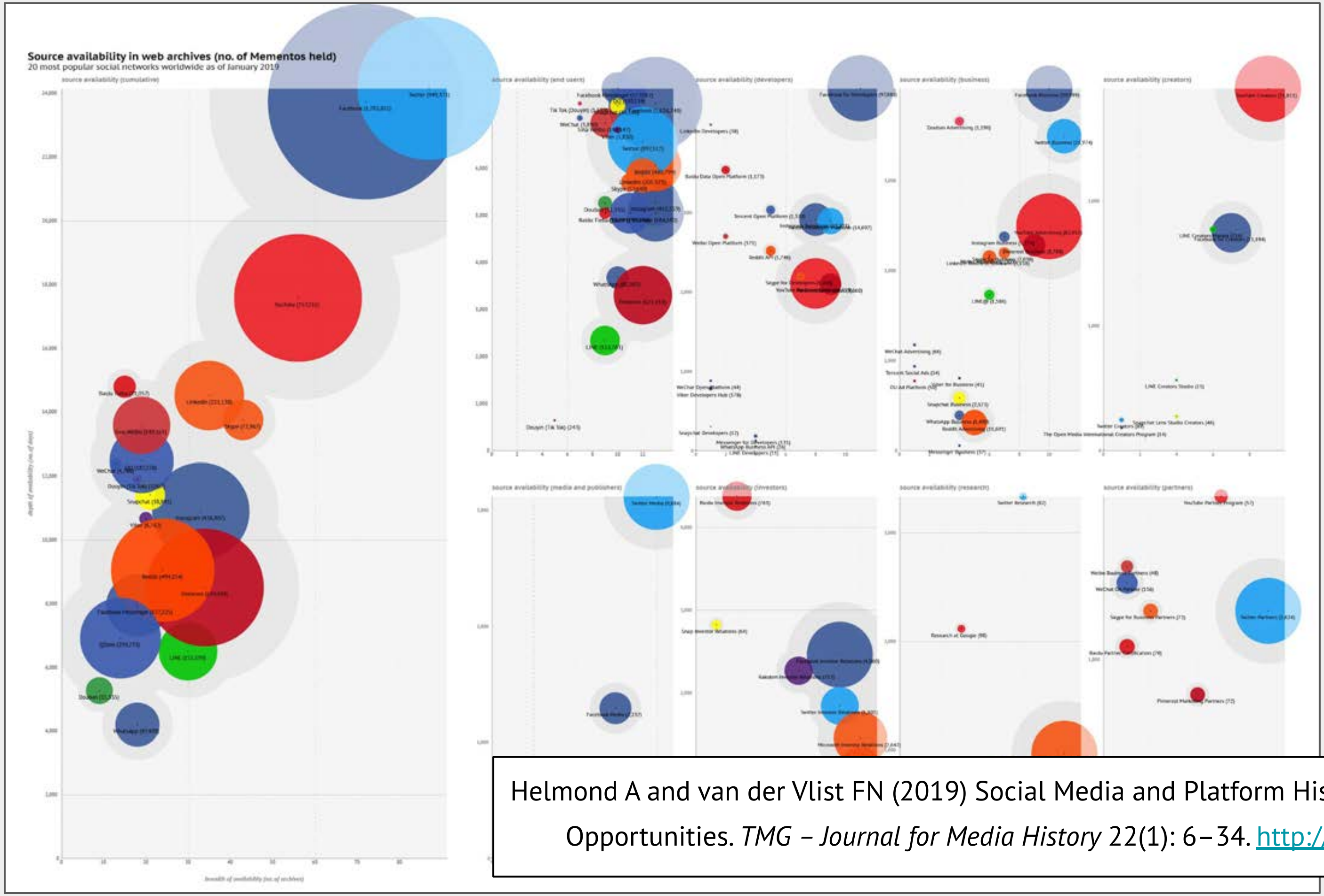
22:50:39

Find Mementos in Internet Archive, Archive-It, British Library, archive.today, GitHub and many more!

https://timetravel.mementoweb.org/

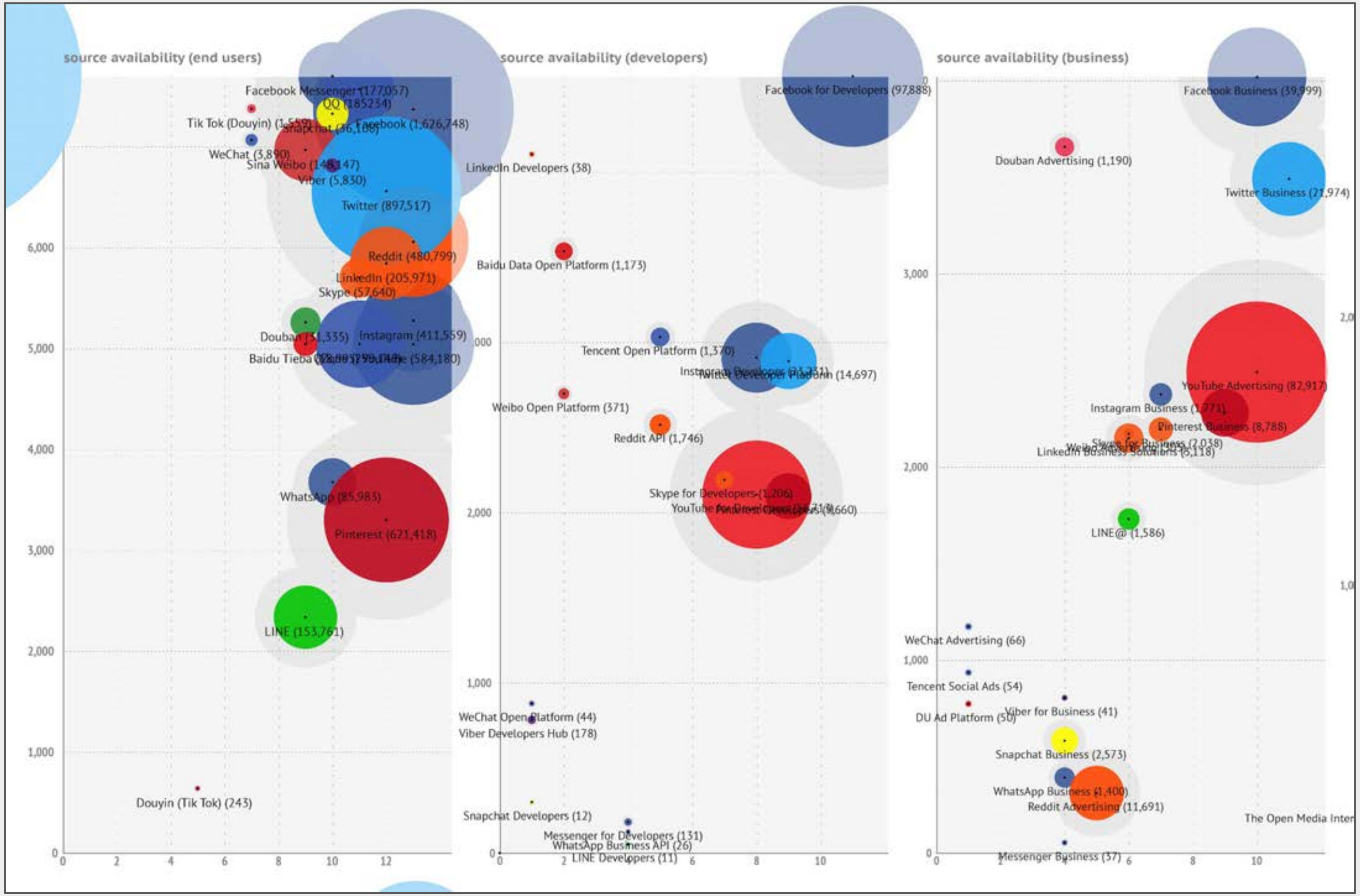
Helmond A and van der Vlist FN (2019) Social Media and Platform Historiography: Challenges and Opportunities. TMG – Journal for Media History 22(1): 6–34. http://doi.org/10.18146/tmg.434

Sociale media bronnen in web archieven

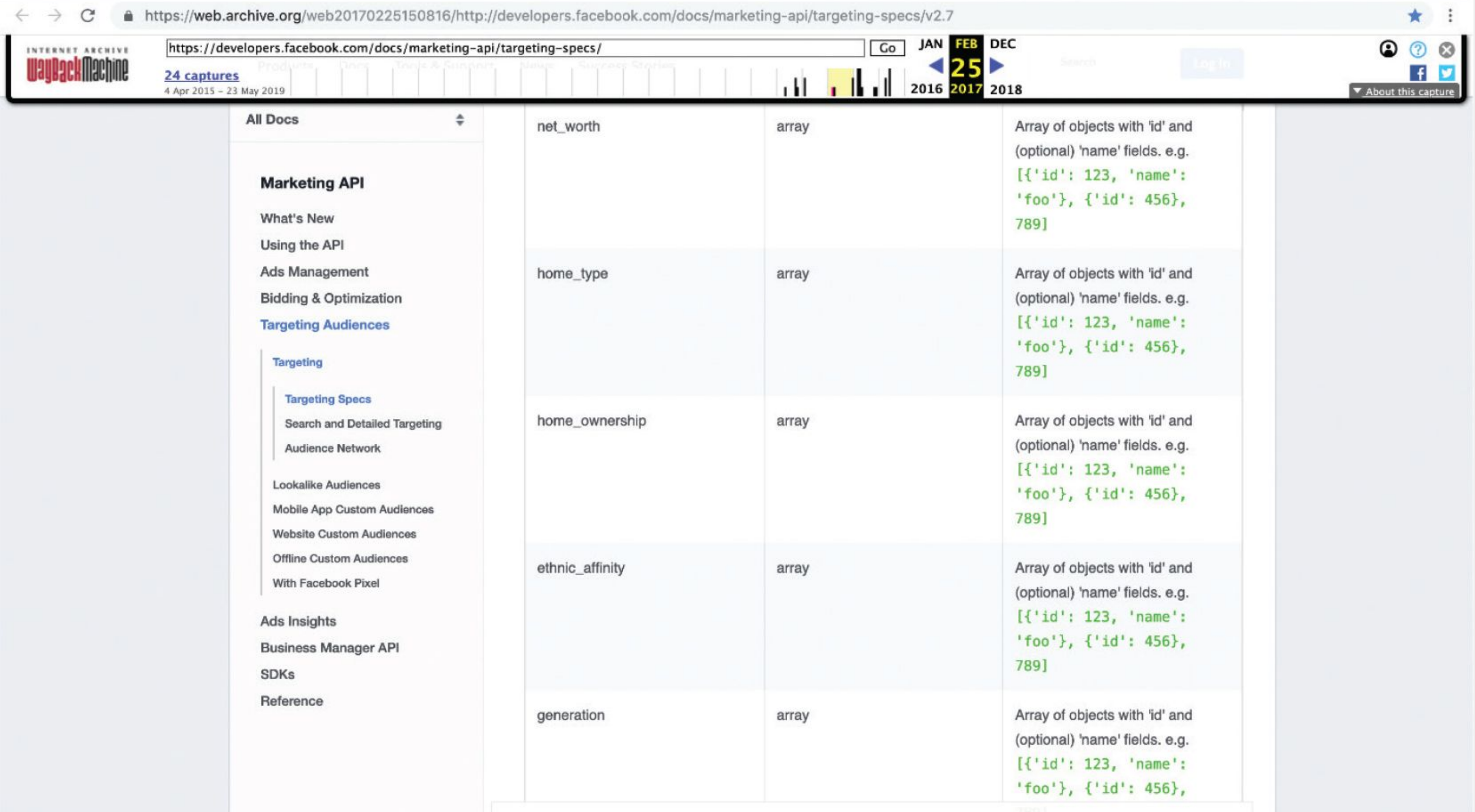


Helmond A and van der Vlist FN (2019) Social Media and Platform Historiography: Challenges and Opportunities. *TMG – Journal for Media History* 22(1): 6–34. http://doi.org/10.18146/tmg.434

Sociale media bronnen in web archieven



Advertentie doelgroepen van Facebook



https://web.archive.org/web20170225150816/http://developers.facebook.com/docs/marketing-api/targeting-specs/v2.7

INTERNET ARCHIVE WaybackMachine

https://developers.facebook.com/docs/marketing-api/targeting-specs/ Go

24 captures
4 Apr 2015 – 23 May 2019

JAN FEB DEC
25
2016 2017 2018

About this capture

All Docs

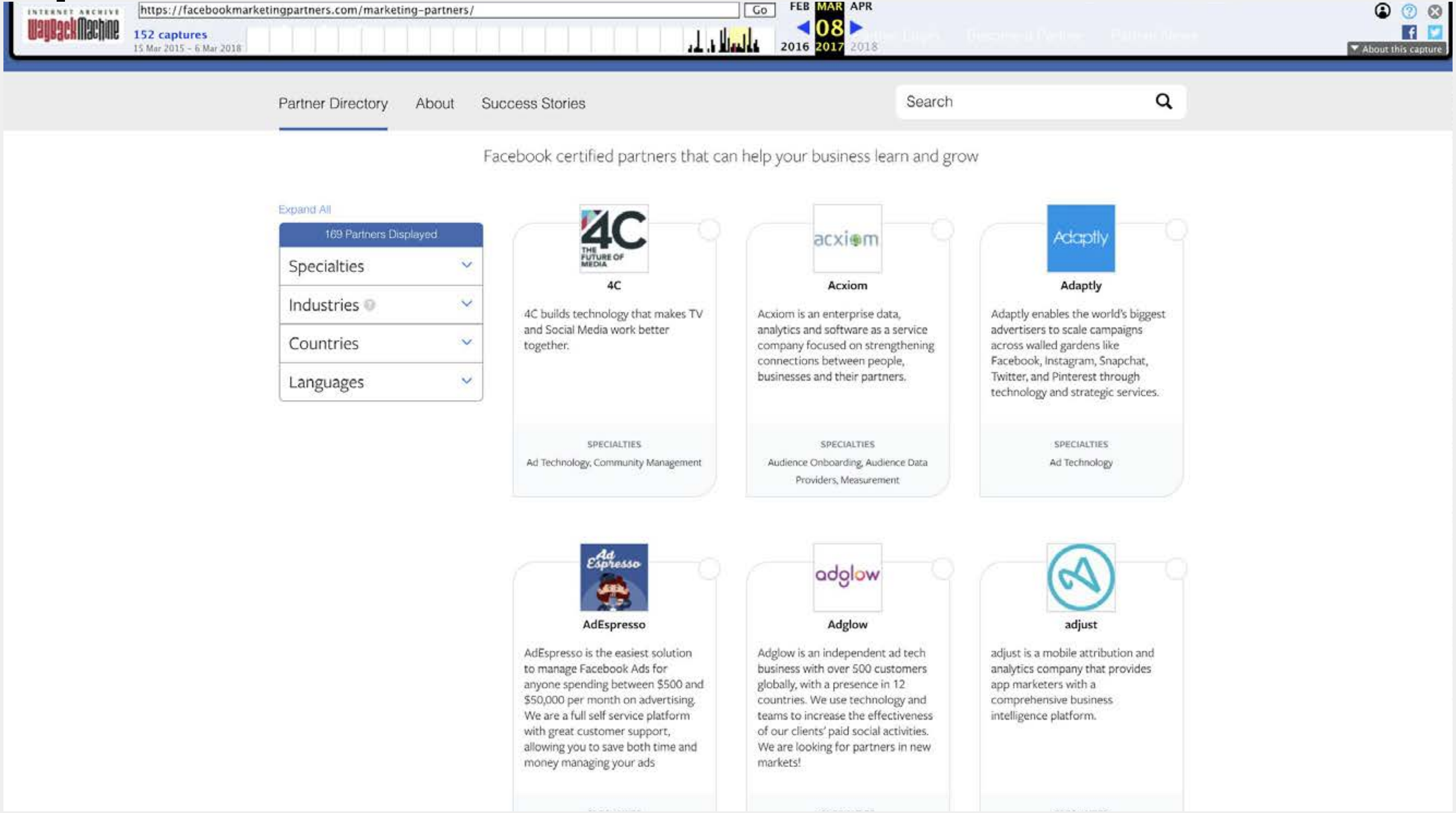
Marketing API

- What's New
- Using the API
- Ads Management
- Bidding & Optimization
- Targeting Audiences
  - Targeting
    - Targeting Specs
    - Search and Detailed Targeting
    - Audience Network
  - Lookalike Audiences
  - Mobile App Custom Audiences
  - Website Custom Audiences
  - Offline Custom Audiences
  - With Facebook Pixel
- Ads Insights
- Business Manager API
- SDKs
- Reference

net_worth	array	Array of objects with 'id' and (optional) 'name' fields. e.g. `[{'id': 123, 'name': 'foo'}, {'id': 456}, 789]`
home_type	array	Array of objects with 'id' and (optional) 'name' fields. e.g. `[{'id': 123, 'name': 'foo'}, {'id': 456}, 789]`
home_ownership	array	Array of objects with 'id' and (optional) 'name' fields. e.g. `[{'id': 123, 'name': 'foo'}, {'id': 456}, 789]`
ethnic_affinity	array	Array of objects with 'id' and (optional) 'name' fields. e.g. `[{'id': 123, 'name': 'foo'}, {'id': 456}, 789]`
generation	array	Array of objects with 'id' and (optional) 'name' fields. e.g. `[{'id': 123, 'name': 'foo'}, {'id': 456},`

Screen capture of Facebook Marketing API audience targeting specifications at developers.facebook.com/docs/marketing-api/targeting-specs/, including the 'ethnic_affinity' field (now disabled). *Internet Archive Wayback Machine, February 25, 2017* (Helmond & van der Vlist, 2019).

De partners van Facebook



Facebook certified partners that can help your business learn and grow

4C

4C builds technology that makes TV and Social Media work better together.

SPECIALTIES

Ad Technology, Community Management

Acxiom

Acxiom is an enterprise data, analytics and software as a service company focused on strengthening connections between people, businesses and their partners.

SPECIALTIES

Audience Onboarding, Audience Data Providers, Measurement

Adaptly

Adaptly enables the world's biggest advertisers to scale campaigns across walled gardens like Facebook, Instagram, Snapchat, Twitter, and Pinterest through technology and strategic services.

SPECIALTIES

Ad Technology

AdEspresso

AdEspresso is the easiest solution to manage Facebook Ads for anyone spending between $500 and $50,000 per month on advertising. We are a full self service platform with great customer support, allowing you to save both time and money managing your ads

Adglow

Adglow is an independent ad tech business with over 500 customers globally, with a presence in 12 countries. We use technology and teams to increase the effectiveness of our clients' paid social activities. We are looking for partners in new markets!

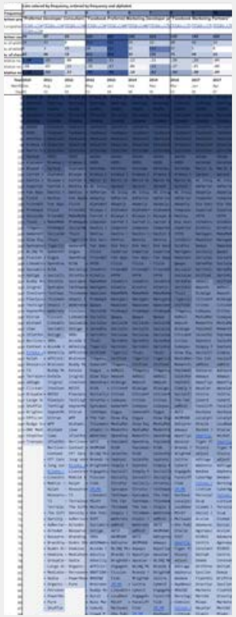
adjust

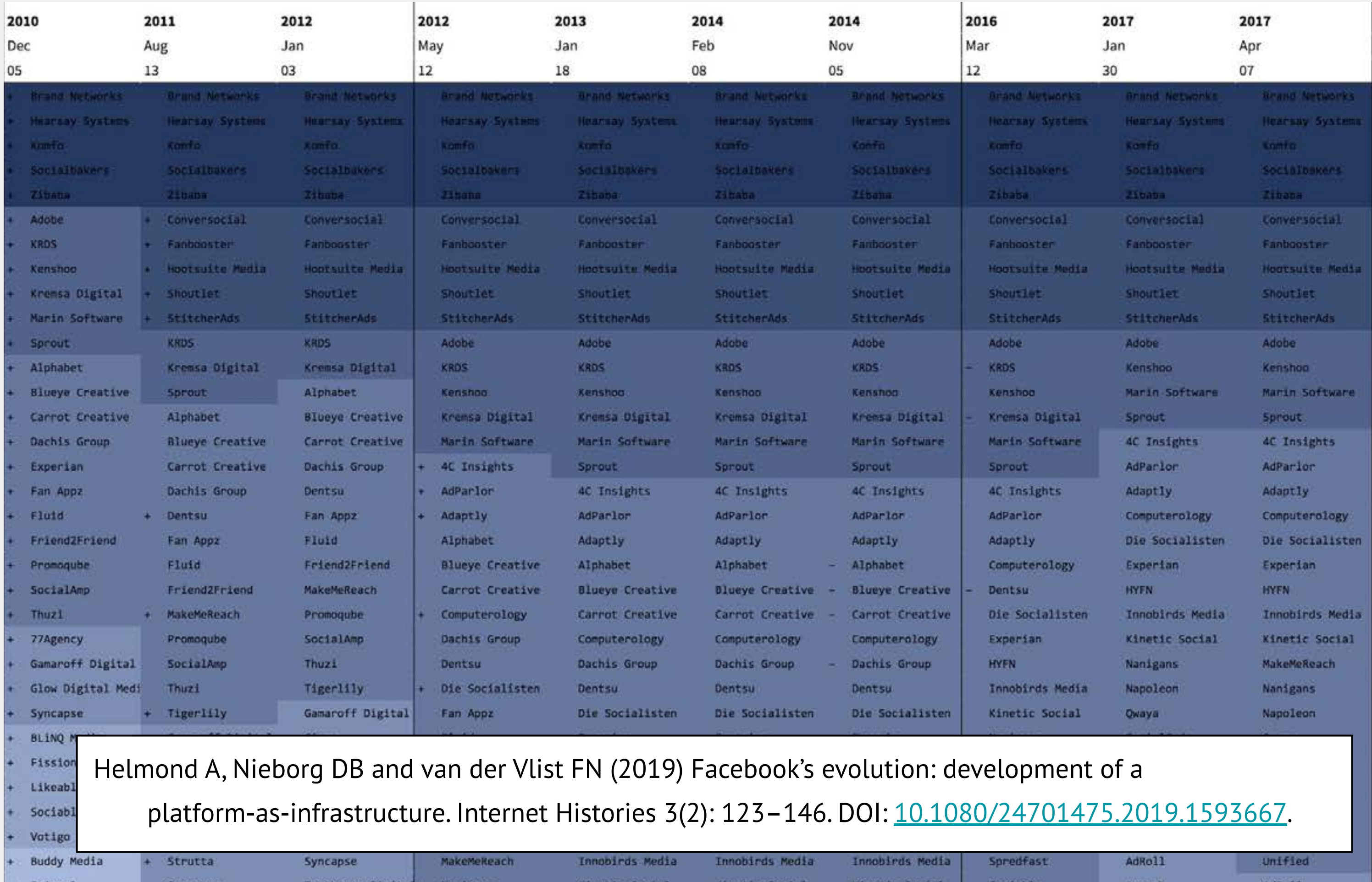
adjust is a mobile attribution and analytics company that provides app marketers with a comprehensive business intelligence platform.

Officiële Facebook partners, 2010–2017

bit.ly/2oYvpI5

Zoom: 25%





2010	2011	2012	2012	2013	2014	2014	2016	2017	2017
Dec	Aug	Jan	May	Jan	Feb	Nov	Mar	Jan	Apr
05	13	03	12	18	08	05	12	30	07
+ Brand Networks	Brand Networks	Brand Networks	Brand Networks	Brand Networks	Brand Networks	Brand Networks	Brand Networks	Brand Networks	Brand Networks
+ Hearsay Systems	Hearsay Systems	Hearsay Systems	Hearsay Systems	Hearsay Systems	Hearsay Systems	Hearsay Systems	Hearsay Systems	Hearsay Systems	Hearsay Systems
+ Komfo	Komfo	Komfo	Komfo	Komfo	Komfo	Komfo	Komfo	Komfo	Komfo
+ Socialbakers	Socialbakers	Socialbakers	Socialbakers	Socialbakers	Socialbakers	Socialbakers	Socialbakers	Socialbakers	Socialbakers
+ Zibaba	Zibaba	Zibaba	Zibaba	Zibaba	Zibaba	Zibaba	Zibaba	Zibaba	Zibaba
+ Adobe	+ Conversocial	Conversocial	Conversocial	Conversocial	Conversocial	Conversocial	Conversocial	Conversocial	Conversocial
+ KRDS	+ Fanbooster	Fanbooster	Fanbooster	Fanbooster	Fanbooster	Fanbooster	Fanbooster	Fanbooster	Fanbooster
+ Kenshoo	+ Hootsuite Media	Hootsuite Media	Hootsuite Media	Hootsuite Media	Hootsuite Media	Hootsuite Media	Hootsuite Media	Hootsuite Media	Hootsuite Media
+ Kremsa Digital	+ Shoutlet	Shoutlet	Shoutlet	Shoutlet	Shoutlet	Shoutlet	Shoutlet	Shoutlet	Shoutlet
+ Marin Software	+ StitcherAds	StitcherAds	StitcherAds	StitcherAds	StitcherAds	StitcherAds	StitcherAds	StitcherAds	StitcherAds
+ Sprout	KRDS	KRDS	Adobe	Adobe	Adobe	Adobe	Adobe	Adobe	Adobe
+ Alphabet	Kremsa Digital	Kremsa Digital	KRDS	KRDS	KRDS	KRDS	- KRDS	Kenshoo	Kenshoo
+ Blueye Creative	Sprout	Alphabet	Kenshoo	Kenshoo	Kenshoo	Kenshoo	Kenshoo	Marin Software	Marin Software
+ Carrot Creative	Alphabet	Blueye Creative	Kremsa Digital	Kremsa Digital	Kremsa Digital	Kremsa Digital	- Kremsa Digital	Sprout	Sprout
+ Dachis Group	Blueye Creative	Carrot Creative	Marin Software	Marin Software	Marin Software	Marin Software	Marin Software	4C Insights	4C Insights
+ Experian	Carrot Creative	Dachis Group	+ 4C Insights	Sprout	Sprout	Sprout	Sprout	AdParlor	AdParlor
+ Fan Appz	Dachis Group	Dentsu	+ AdParlor	4C Insights	4C Insights	4C Insights	4C Insights	Adaptly	Adaptly
+ Fluid	+ Dentsu	Fan Appz	+ Adaptly	AdParlor	AdParlor	AdParlor	AdParlor	Computerology	Computerology
+ Friend2Friend	Fan Appz	Fluid	Alphabet	Adaptly	Adaptly	Adaptly	Adaptly	Die Socialisten	Die Socialisten
+ Promoqube	Fluid	Friend2Friend	Blueye Creative	Alphabet	Alphabet	- Alphabet	Computerology	Experian	Experian
+ SocialAmp	Friend2Friend	MakeMeReach	Carrot Creative	Blueye Creative	Blueye Creative	- Blueye Creative	- Dentsu	HYFN	HYFN
+ Thuzi	+ MakeMeReach	Promoqube	+ Computerology	Carrot Creative	Carrot Creative	- Carrot Creative	Die Socialisten	Innobirds Media	Innobirds Media
+ 77Agency	Promoqube	SocialAmp	Dachis Group	Computerology	Computerology	Computerology	Experian	Kinetic Social	Kinetic Social
+ Gamaroff Digital	SocialAmp	Thuzi	Dentsu	Dachis Group	Dachis Group	- Dachis Group	HYFN	Nanigans	MakeMeReach
+ Glow Digital Medi	Thuzi	Tigerlily	+ Die Socialisten	Dentsu	Dentsu	Dentsu	Innobirds Media	Napoleon	Nanigans
+ Syncapse	+ Tigerlily	Gamaroff Digital	Fan Appz	Die Socialisten	Die Socialisten	Die Socialisten	Kinetic Social	Qwaya	Napoleon
+ BLiNQ M									
+ Fission									
+ Likeabl									
+ Sociabl									
+ Votigo									
+ Buddy Media	+ Strutta	Syncapse	MakeMeReach	Innobirds Media	Innobirds Media	Innobirds Media	Spredfast	AdRoll	Unified

Helmond A, Nieborg DB and van der Vlist FN (2019) Facebook's evolution: development of a platform-as-infrastructure. Internet Histories 3(2): 123–146. DOI: 10.1080/24701475.2019.1593667.

Sociale media platforms: GUI versus API

Graphical User Interface (GUI):

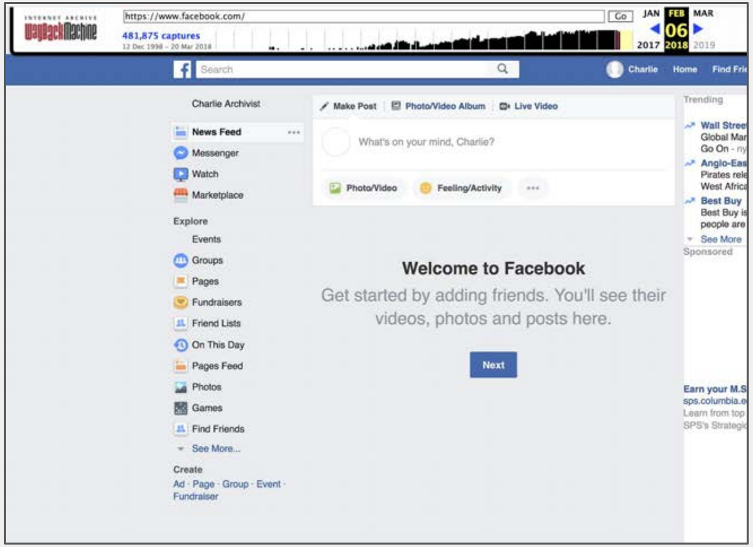
- Archiveringsmethode: crawler
- Toegang: via login
- De “voorkant” van sociale media, toegang tot de gebruikersinterface.
- Gebruikersgroepen: eindgebruikers, maar ook politici, adverteerders, etc.

Application Programming Interface (API):

- Archiveringsmethode: API
- Toegang: via developer account
- De “achterkant” van sociale media, toegang tot de data in de database.
- Gebruikersgroepen: (app) ontwikkelaars, adverteerders en marketing bedrijven.

Sociale media archivering: GUI/API

Crawler-methode



What's on your mind, Charlie?

Welcome to Facebook

Get started by adding friends. You'll see their videos, photos and posts here.

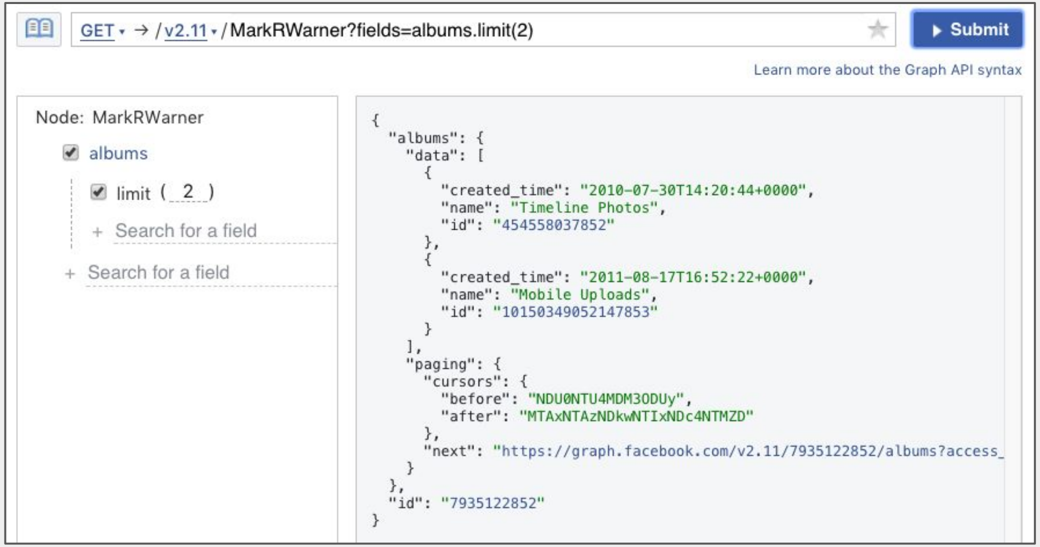
Global Mar Go On - ny

Pirates rele West Afric

Best Buy is people are

Learn from top SPS's Strategi

API-methode



GET → /v2.11 /MarkRWarner?fields=albums.limit(2)

Node: MarkRWarner

☑ albums

☑ limit (2)

```
{
  "albums": {
    "data": [
      {
        "created_time": "2010-07-30T14:20:44+0000",
        "name": "Timeline Photos",
        "id": "454558037852"
      },
      {
        "created_time": "2011-08-17T16:52:22+0000",
        "name": "Mobile Uploads",
        "id": "10150349052147853"
      }
    ],
    "paging": {
      "cursors": {
        "before": "NDU0NTU4MDM3ODUy",
        "after": "MTAxNTAzNDkwNTIxNDc4NTMZD"
      },
      "next": "https://graph.facebook.com/v2.11/7935122852/albums?access_
    }
  },
  "id": "7935122852"
}
```

GUI ≠ API: Niet alle data die je in de GUI ziet kun je via de API krijgen!

Sociale media archivering via APIs

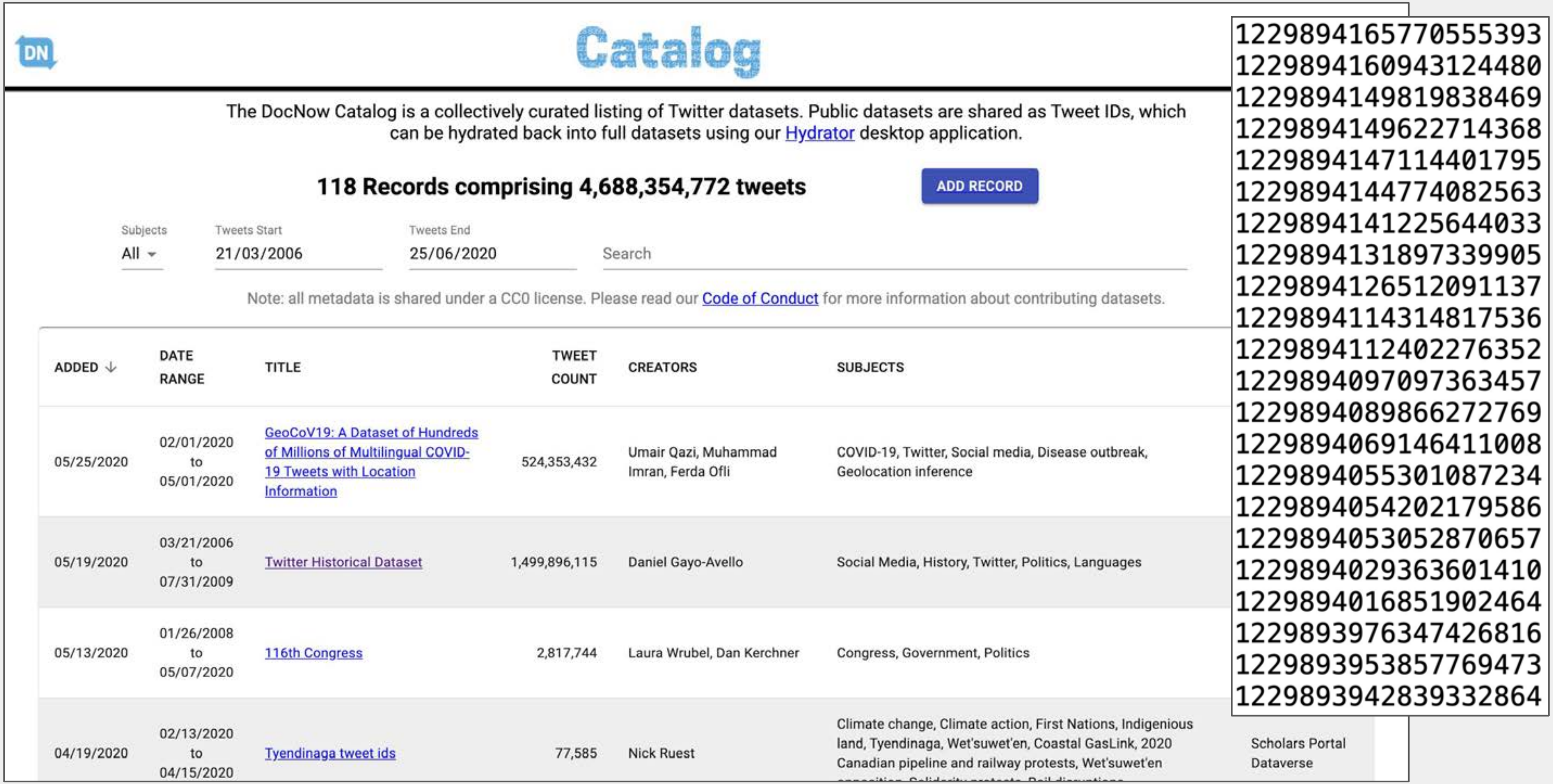
- De archivering van sociale media via APIs levert een **data-blik** op het platform die verschillende vormen van **data-analyse** mogelijk maakt.

- Onderzoekers maken zelf datasets of gebruiken bestaande datasets.
  - Bestaande datasets gebruiken:
    - 👍Iemand anders heeft al waardevolle historische data verzameld
    - 👎Lastig te vinden en niet altijd duidelijk gedocumenteerd
  - Zelf datasets maken:
    - 👍De onderzoeksvraag bepaalt welke data je nodig hebt
    - 👎Tools, skills en (veel) opslagruimte nodig

Bestaande datasets gebruiken

- Opslag en ontsluiting van sociale media data is technisch lastig en kan zeer veel ruimte kosten.
    - ‘The Twitter Archive at the Library of Congress: Challenges for information practice and information policy’ (Zimmer, 2015).
    - ‘The Library of Congress Twitter Archive: A Failure of Historic Proportions’ (Bruns, 2018).
- Zeer sterke restricties voor het maken, delen en gebruiken van datasets.
    - Veranderende terms of service.
    - Platformen stellen steeds minder data beschikbaar ‘APIcalypse’ (Bruns, 2019).
    - Twitter: datasets herpubliceren alleen op basis van tweet-ID.
    - Facebook: Social Science One (partnerovereenkomst)
    - Historische data als verdienmodel (Twitter).

Bestaande Twitter datasets (DocNow Catalog)



Catalog

The DocNow Catalog is a collectively curated listing of Twitter datasets. Public datasets are shared as Tweet IDs, which can be hydrated back into full datasets using our Hydrator desktop application.

118 Records comprising 4,688,354,772 tweets

Subjects: All
Tweets Start: 21/03/2006
Tweets End: 25/06/2020
Search

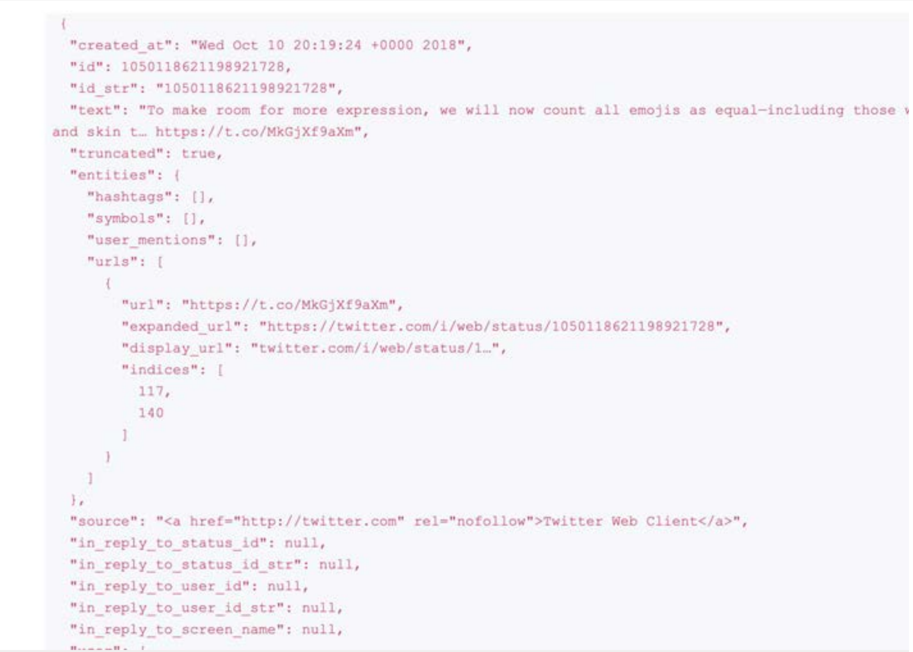
Note: all metadata is shared under a CC0 license. Please read our Code of Conduct for more information about contributing datasets.

ADDED ↓	DATE RANGE	TITLE	TWEET COUNT	CREATORS	SUBJECTS	
05/25/2020	02/01/2020 to 05/01/2020	GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information	524,353,432	Umair Qazi, Muhammad Imran, Ferda Ofli	COVID-19, Twitter, Social media, Disease outbreak, Geolocation inference	
05/19/2020	03/21/2006 to 07/31/2009	Twitter Historical Dataset	1,499,896,115	Daniel Gayo-Avello	Social Media, History, Twitter, Politics, Languages	
05/13/2020	01/26/2008 to 05/07/2020	116th Congress	2,817,744	Laura Wrubel, Dan Kerchner	Congress, Government, Politics	
04/19/2020	02/13/2020 to 04/15/2020	Tyendinaga tweet ids	77,585	Nick Ruest	Climate change, Climate action, First Nations, Indigenious land, Tyendinaga, Wet'suwet'en, Coastal GasLink, 2020 Canadian pipeline and railway protests, Wet'suwet'en	Scholars Portal Dataverse

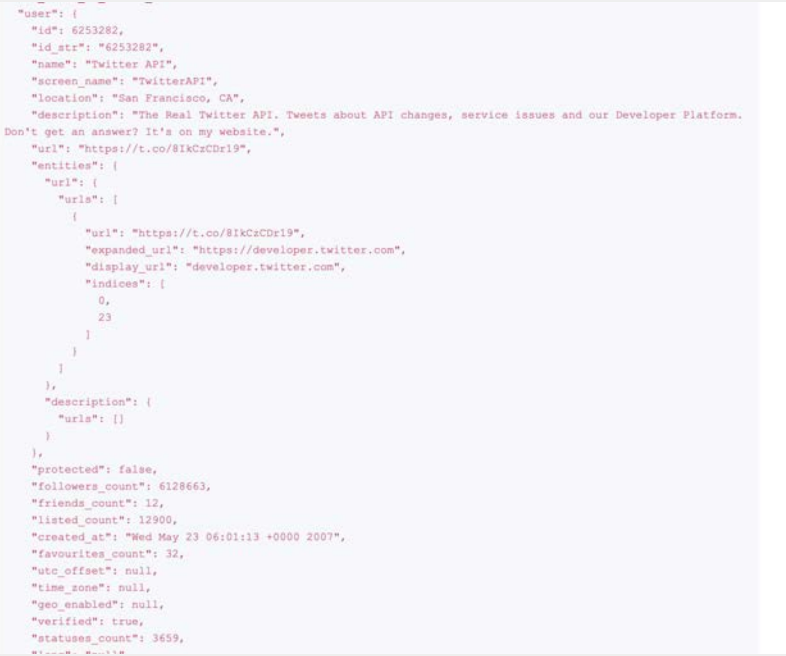
```
1229894165770555393
1229894160943124480
1229894149819838469
1229894149622714368
1229894147114401795
1229894144774082563
1229894141225644033
1229894131897339905
1229894126512091137
1229894114314817536
1229894112402276352
1229894097097363457
1229894089866272769
1229894069146411008
1229894055301087234
1229894054202179586
1229894053052870657
1229894029363601410
1229894016851902464
1229893976347426816
1229893953857769473
1229893942839332864
```

Van tweet id naar volledige tweet

```
"id": 1050118621198921728
```

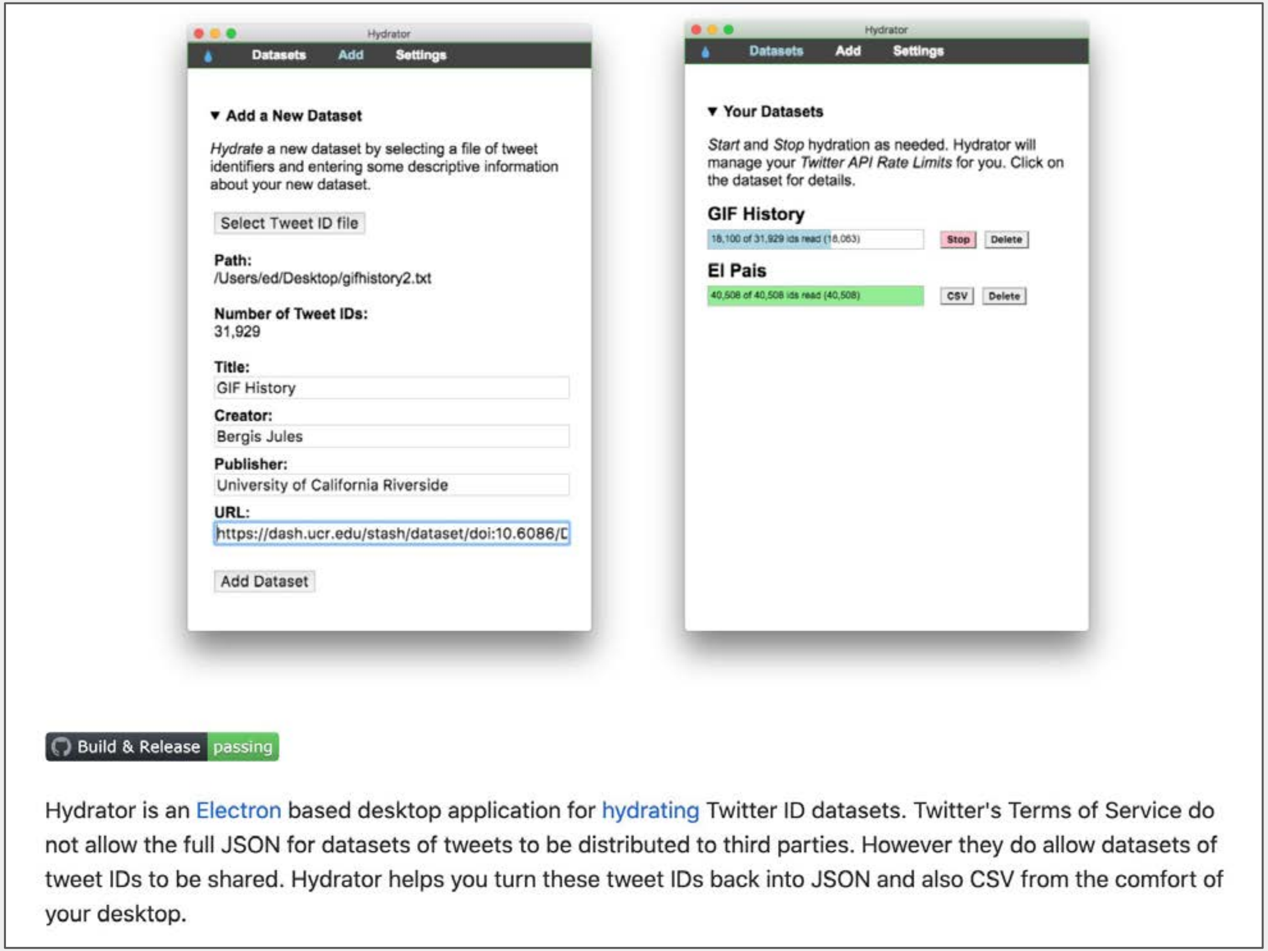


```
{
  "created_at": "Wed Oct 10 20:19:24 +0000 2018",
  "id": 1050118621198921728,
  "id_str": "1050118621198921728",
  "text": "To make room for more expression, we will now count all emojis as equal—including those w
and skin t… https://t.co/MkGjXf9aXm",
  "truncated": true,
  "entities": {
    "hashtags": [],
    "symbols": [],
    "user_mentions": [],
    "urls": [
      {
        "url": "https://t.co/MkGjXf9aXm",
        "expanded_url": "https://twitter.com/i/web/status/1050118621198921728",
        "display_url": "twitter.com/i/web/status/1…",
        "indices": [
          117,
          140
        ]
      }
    ]
  },
  "source": "<a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>",
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
```



```
  "user": {
    "id": 6253282,
    "id_str": "6253282",
    "name": "Twitter API",
    "screen_name": "TwitterAPI",
    "location": "San Francisco, CA",
    "description": "The Real Twitter API. Tweets about API changes, service issues and our Developer Platform.
Don't get an answer? It's on my website.",
    "url": "https://t.co/8IkCzCDr19",
    "entities": {
      "url": {
        "urls": [
          {
            "url": "https://t.co/8IkCzCDr19",
            "expanded_url": "https://developer.twitter.com",
            "display_url": "developer.twitter.com",
            "indices": [
              0,
              23
            ]
          }
        ]
      },
      "description": {
        "urls": []
      }
    },
    "protected": false,
    "followers_count": 6128663,
    "friends_count": 12,
    "listed_count": 12900,
    "created_at": "Wed May 23 06:01:13 +0000 2007",
    "favourites_count": 32,
    "utc_offset": null,
    "time_zone": null,
    "geo_enabled": null,
    "verified": true,
    "statuses_count": 3659,
```

DocNow Twitter Hydrator



▼ Add a New Dataset

Hydrate a new dataset by selecting a file of tweet identifiers and entering some descriptive information about your new dataset.

Path:
/Users/ed/Desktop/gifhistory2.txt

Number of Tweet IDs:
31,929

Title:
GIF History

Creator:
Bergis Jules

Publisher:
University of California Riverside

URL:
https://dash.ucr.edu/stash/dataset/doi:10.6086/D

▼ Your Datasets

Start and *Stop* hydration as needed. Hydrator will manage your *Twitter API Rate Limits* for you. Click on the dataset for details.

GIF History

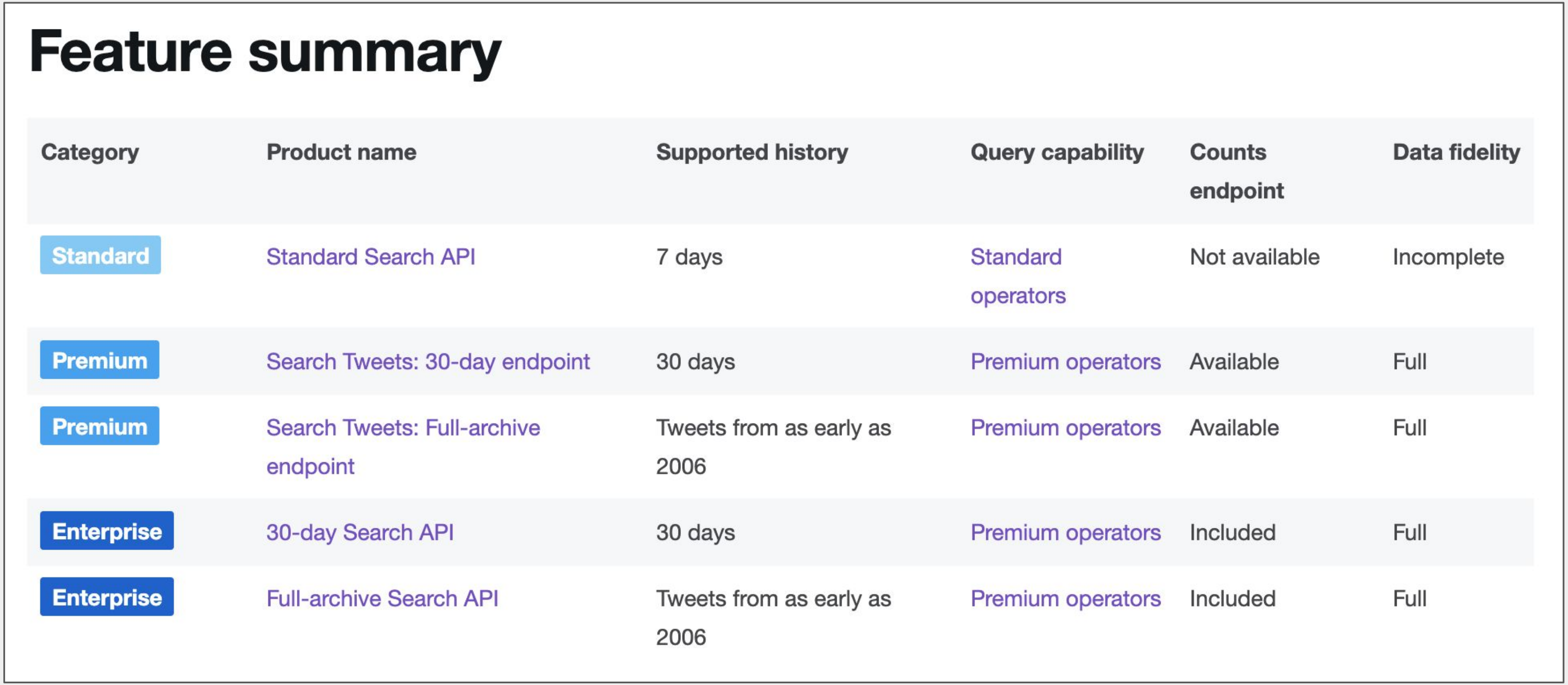
18,100 of 31,929 ids read (18,063)

El Pais

40,508 of 40,508 ids read (40,508)

Hydrator is an Electron based desktop application for hydrating Twitter ID datasets. Twitter's Terms of Service do not allow the full JSON for datasets of tweets to be distributed to third parties. However they do allow datasets of tweet IDs to be shared. Hydrator helps you turn these tweet IDs back into JSON and also CSV from the comfort of your desktop.

Zelf data verzamelen: Twitter Search API



Feature summary

Category	Product name	Supported history	Query capability	Counts endpoint	Data fidelity
Standard	Standard Search API	7 days	Standard operators	Not available	Incomplete
Premium	Search Tweets: 30-day endpoint	30 days	Premium operators	Available	Full
Premium	Search Tweets: Full-archive endpoint	Tweets from as early as 2006	Premium operators	Available	Full
Enterprise	30-day Search API	30 days	Premium operators	Included	Full
Enterprise	Full-archive Search API	Tweets from as early as 2006	Premium operators	Included	Full

https://developer.twitter.com/en/docs/tweets/search/overview

Tools voor sociale media archivering (Twitter 1)

Twitter Archiving Google Sheet (TAGS): https://tags.hawksey.info/

- 👍Gratis, veel mogelijkheden om Twitter data te archiveren via de Twitter API. Gekoppeld aan Google Sheets. Enkele ingebouwde analyse mogelijkheden voor onderzoekers.
- 👎Alleen tweets van de laatste 7 dagen (limitatie Twitter API)

Tools voor sociale media archivering (Twitter 2)

The Digital Methods Initiative Twitter Capture and Analysis Toolset (DMI-TCAT) allows one to retrieve and collect tweets from Twitter and to analyze them in various ways: https://github.com/digitalmethodsinitiative/dmi-tcat

- 👍Gratis, open source. Zeer veel mogelijkheden om Twitter data te archiveren via de Twitter API. Zeer veel ingebouwde analyse mogelijkheden voor onderzoekers (zie instructie video).
- 👎Installeren op eigen server (ook een voordeel).

Analyses

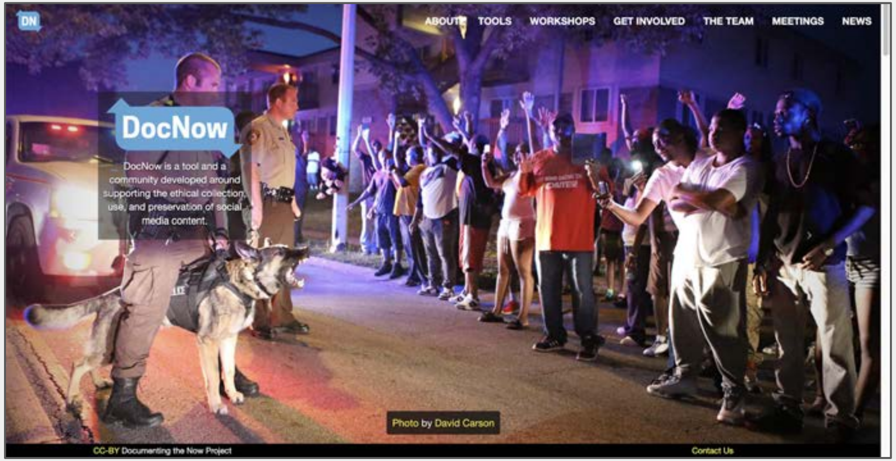
- **#hashtags**: sociale en culturele issues, bijvoorbeeld #BlackLivesMatterNL, #coronamaatregelen
- **mentions**: identificeren van expertise, discussie netwerken
- **following**: volgnetwerken
- **retweets**: resonantie, viraliteit
- **URL**: distributie van content

De rol van web archieven en instanties

Tools voor archivarissen

DocNow is an **appraisal tool** for the web.

“DocNow allows archivists to tap into conversations in Twitter to help them discover what web resources are in need of archiving. Its goal is to help ensure **ethical practices** in web archiving by building conversations between archivists and the communities they are documenting.”



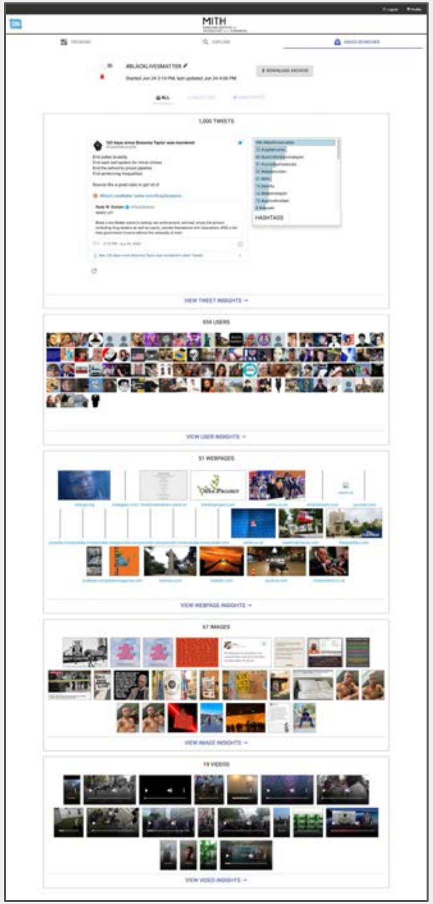
DocNow is a tool and a community developed around supporting the ethical collection, use, and preservation of social media content.

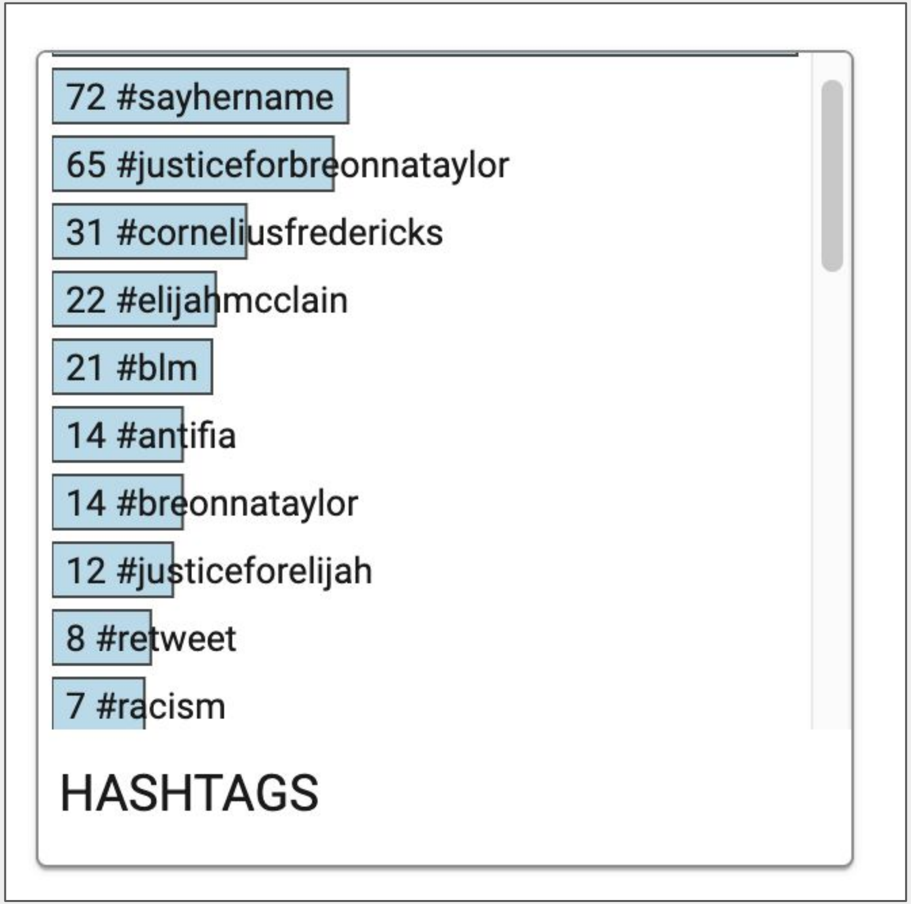
Photo by David Carson

CC-BY Documenting the Now Project

Contact Us

DocNow (demo)





72 #sayhername

65 #justiceforbreonnataylor

31 #corneliusfredericks

22 #elijahmcclain

21 #blm

14 #antifia

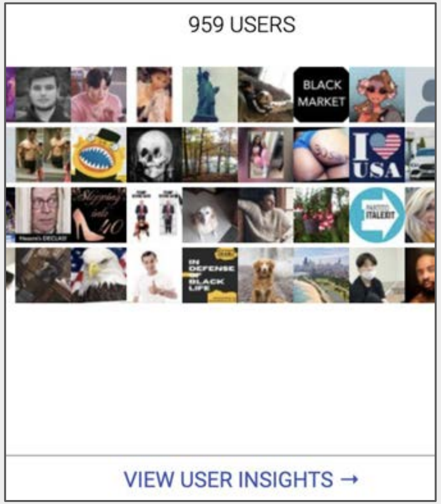
14 #breonnataylor

12 #justiceforelijah

8 #retweet

7 #racism

HASHTAGS



Doel van de collectie

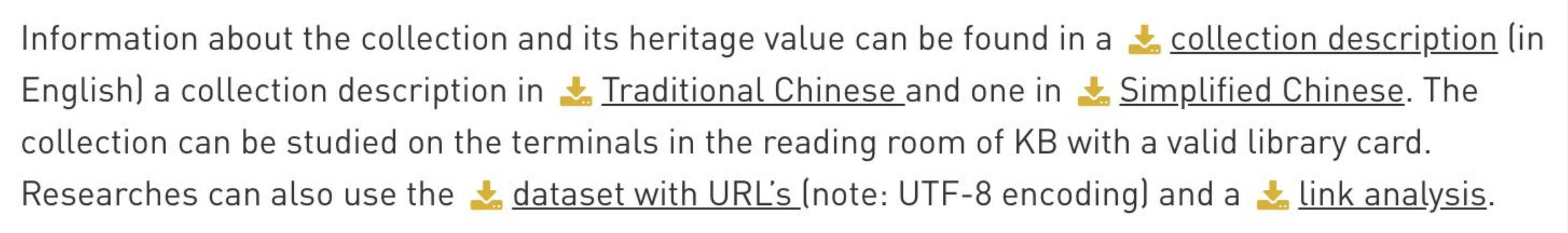
- Keuzes over selectie bepalen later welke vragen onderzoekers kunnen stellen.
- Wie krijgt er toegang tot de collectie en hoe en waar?
    - Onderzoekers werken het liefst met ‘rauwe data’ op hun eigen computer.
    - Maar: privacy en ethiekvraagstukken en platform regels.

Collectiebeschrijving en metadata

- Essentieel voor onderzoekers is informatie over:
  - Hoe is de collectie gemaakt?
  - Welke beslissingen zijn er genomen?
  - Wat zijn de selectiecriteria?
  - Welke data zit er in de set?
  - Hoe is deze data verzameld (en met welke API)?
- “Web archives **provenance**: what users need to know about how a collection was made as they use, analyze, and make inferences from these aggregations” (Maemura et al., 2018).
- Een collectiebeschrijving: Metadata over het gearchiveerde corpus.
- Rauwe data, bijvoorbeeld in .csv of .json (.warc is lastig om mee te werken)

Voorbeeld collectiebeschrijving (1)

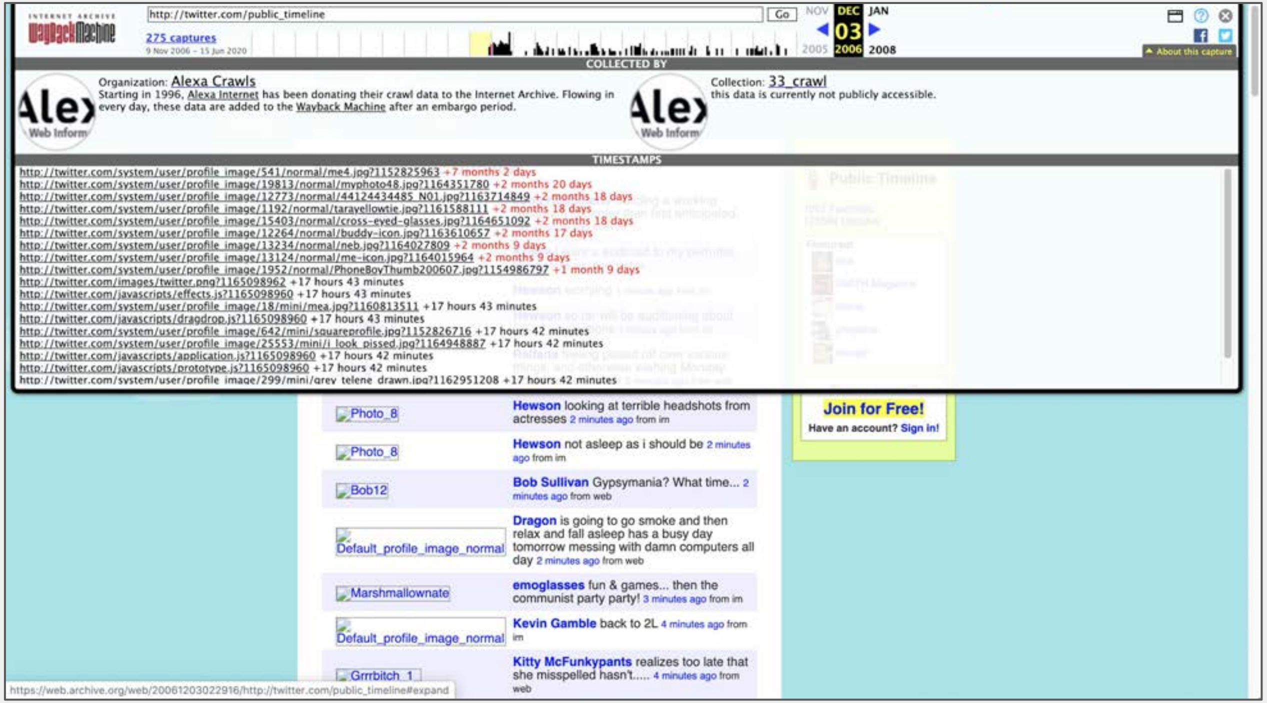
- Bode, P. de, Lin, K., Teszelszky, K. (2019) Chinese Netherlands web collection. KB Lab: The Hague. https://lab.kb.nl/dataset/web-collection-chinese-netherlands



Information about the collection and its heritage value can be found in a collection description (in English) a collection description in Traditional Chinese and one in Simplified Chinese. The collection can be studied on the terminals in the reading room of KB with a valid library card. Researches can also use the dataset with URL's (note: UTF-8 encoding) and a link analysis.

Voorbeeld collectiebeschrijving (2)

- Internet Archive Wayback Machine ‘About this capture’



275 captures
9 Nov 2006 – 15 Jun 2020

COLLECTED BY

Organization: Alexa Crawls
Starting in 1996, Alexa Internet has been donating their crawl data to the Internet Archive. Flowing in every day, these data are added to the Wayback Machine after an embargo period.

Collection: 33_crawl
this data is currently not publicly accessible.

TIMESTAMPS

```
http://twitter.com/system/user/profile_image/541/normal/me4.jpg?1152825963 +7 months 2 days
http://twitter.com/system/user/profile_image/19813/normal/myphoto48.jpg?1164351780 +2 months 20 days
http://twitter.com/system/user/profile_image/12773/normal/44124434485_N01.jpg?1163714849 +2 months 18 days
http://twitter.com/system/user/profile_image/1192/normal/tarayellowtie.jpg?1161588111 +2 months 18 days
http://twitter.com/system/user/profile_image/15403/normal/cross-eyed-glasses.jpg?1164651092 +2 months 18 days
http://twitter.com/system/user/profile_image/12264/normal/buddy-icon.jpg?1163610657 +2 months 17 days
http://twitter.com/system/user/profile_image/13234/normal/neb.jpg?1164027809 +2 months 9 days
http://twitter.com/system/user/profile_image/13124/normal/me-icon.jpg?1164015964 +2 months 9 days
http://twitter.com/system/user/profile_image/1952/normal/PhoneBoyThumb200607.jpg?1154986797 +1 month 9 days
http://twitter.com/images/twitter.png?1165098962 +17 hours 43 minutes
http://twitter.com/javascripts/effects.js?1165098960 +17 hours 43 minutes
http://twitter.com/system/user/profile_image/18/mini/mea.jpg?1160813511 +17 hours 43 minutes
http://twitter.com/javascripts/dragdrop.js?1165098960 +17 hours 43 minutes
http://twitter.com/system/user/profile_image/642/mini/squareprofile.jpg?1152826716 +17 hours 42 minutes
http://twitter.com/system/user/profile_image/25553/mini/i_look_pissed.jpg?1164948887 +17 hours 42 minutes
http://twitter.com/javascripts/application.js?1165098960 +17 hours 42 minutes
http://twitter.com/javascripts/prototype.js?1165098960 +17 hours 42 minutes
http://twitter.com/system/user/profile_image/299/mini/grey_telene_drawn.jpg?1162951208 +17 hours 42 minutes
```

Photo_8 — **Hewson** looking at terrible headshots from actresses 2 minutes ago from im

Photo_8 — **Hewson** not asleep as i should be 2 minutes ago from im

Bob12 — **Bob Sullivan** Gypsymania? What time... 2 minutes ago from web

Default_profile_image_normal — **Dragon** is going to go smoke and then relax and fall asleep has a busy day tomorrow messing with damn computers all day 2 minutes ago from web

Marshmallownate — **emoglasses** fun & games... then the communist party party! 3 minutes ago from im

Default_profile_image_normal — **Kevin Gamble** back to 2L 4 minutes ago from im

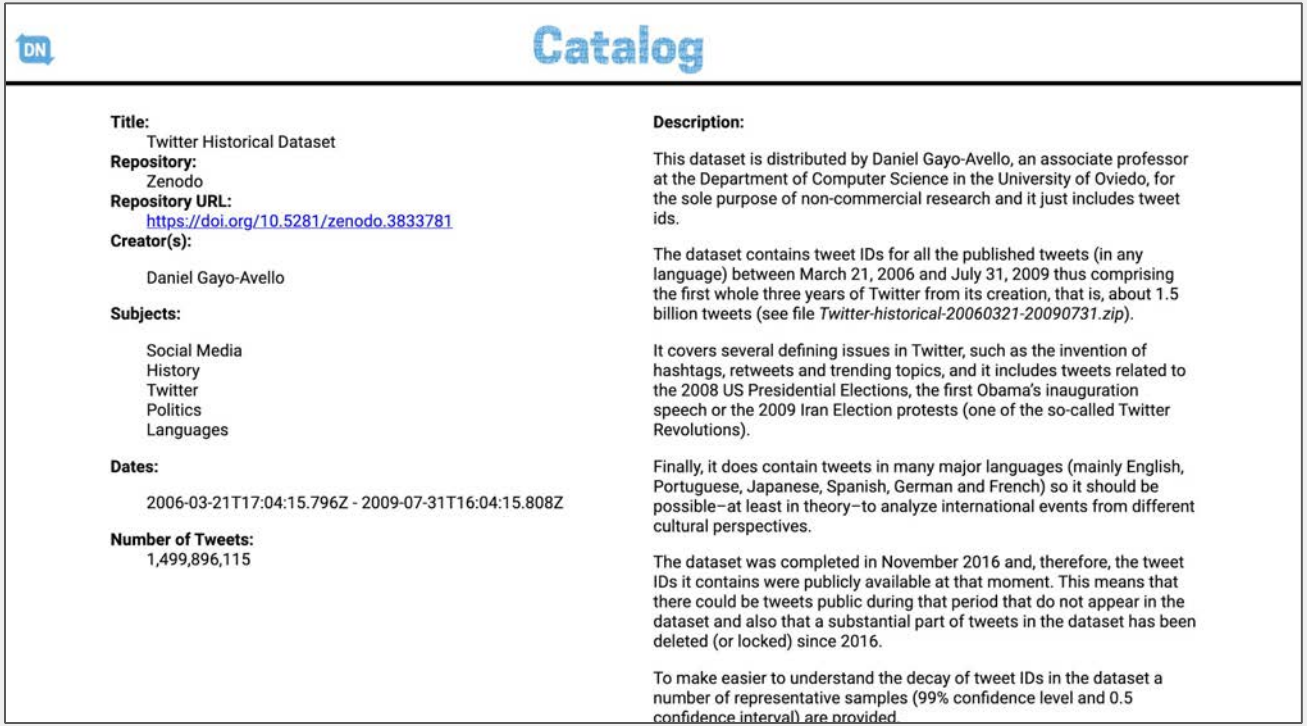
Grrrbitch_1 — **Kitty McFunkypants** realizes too late that she misspelled hasn't..... 4 minutes ago from web

Join for Free!
Have an account? Sign in!

https://web.archive.org/web/20061203022916/http://twitter.com/public_timeline#expand

Voorbeeld collectiebeschrijving (3)

- The DocNow Catalog



Catalog

Title:
Twitter Historical Dataset
Repository:
Zenodo
Repository URL:
https://doi.org/10.5281/zenodo.3833781
Creator(s):

Daniel Gayo-Avello

Subjects:

Social Media
History
Twitter
Politics
Languages

Dates:

2006-03-21T17:04:15.796Z - 2009-07-31T16:04:15.808Z

Number of Tweets:
1,499,896,115

Description:

This dataset is distributed by Daniel Gayo-Avello, an associate professor at the Department of Computer Science in the University of Oviedo, for the sole purpose of non-commercial research and it just includes tweet ids.

The dataset contains tweet IDs for all the published tweets (in any language) between March 21, 2006 and July 31, 2009 thus comprising the first whole three years of Twitter from its creation, that is, about 1.5 billion tweets (see file *Twitter-historical-20060321-20090731.zip*).

It covers several defining issues in Twitter, such as the invention of hashtags, retweets and trending topics, and it includes tweets related to the 2008 US Presidential Elections, the first Obama's inauguration speech or the 2009 Iran Election protests (one of the so-called Twitter Revolutions).

Finally, it does contain tweets in many major languages (mainly English, Portuguese, Japanese, Spanish, German and French) so it should be possible–at least in theory–to analyze international events from different cultural perspectives.

The dataset was completed in November 2016 and, therefore, the tweet IDs it contains were publicly available at that moment. This means that there could be tweets public during that period that do not appear in the dataset and also that a substantial part of tweets in the dataset has been deleted (or locked) since 2016.

To make easier to understand the decay of tweet IDs in the dataset a number of representative samples (99% confidence level and 0.5 confidence interval) are provided.

Conclusie

- Archieven, archiveringspraktijken en sociale media platformen ‘vormen’ het corpus en beïnvloeden de soorten vragen die gesteld kunnen worden.
- De onderzoeksvraag is leidend voor welke archief of welke data je nodig hebt.
- Het nauwkeurig documenteren van een collectie is essentieel voor onderzoekers om gebruik te maken van de gearchiveerde sociale media data.
- Voor web archieven, instellingen en onderzoekers is API kennis belangrijk.

Bedankt! Vragen?

Helmond A and van der Vlist FN (2019) Social Media and Platform Historiography: Challenges and Opportunities. *TMG – Journal for Media History* 22(1): 6–34. doi:10.18146/tmg.434

Helmond A, Nieborg DB and van der Vlist FN (2019) Facebook's evolution: Development of a platform-as-infrastructure. *Internet Histories: Digital Technology, Culture and Society*, 3(2), 123–146. doi: 10.1145/3097286.3097324.

Anne Helmond (Universiteit van Amsterdam)
a.helmond@uva.nl

This work is part of the research programme Innovational Research Incentives Scheme Veni with project number 275-45-009, which is (partly) financed by the Dutch Research Council (NWO).

Bronnen en links

- Brügger N (2018a) *The Archived Web: Doing History in the Digital Age*. Cambridge, MA: The MIT Press.
- Brügger N (2018b) *Web History and Social Media*. In: Burgess J, Poell T, and Marwick A (eds) The SAGE Handbook of Social Media. London: SAGE Publications, pp. 196–212.
- Bruns A (2019) After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society* 22(11). Routledge: 1544–1566. DOI: 10.1080/1369118X.2019.1637447.
- Helmond A and van der Vlist FN (2019) Social Media and Platform Historiography: Challenges and Opportunities. *TMG – Journal for Media History* 22(1): 6–34. http://www.tmgonline.nl/articles/434/
- Helmond A, Nieborg DB and van der Vlist FN (2019) Facebook's evolution: development of a platform-as-infrastructure. *Internet Histories* 3(2): 123–146. DOI: 10.1080/24701475.2019.1593667.
- https://tags.hawksey.info/
- https://github.com/digitalmethodsinitiative/dmi-tcat
- https://www.docnow.io/
- https://catalog.docnow.io/
- https://webrecorder.net/
- https://conifer.rhizome.org/
- https://timetravel.mementoweb.org/
- https://socialscience.one/
- https://github.com/emilymae/web-archives-bib/blob/master/research-by-theme.md#5-twitter-and-social-media