Social Media Archiving Tools

Zefi Kavvadia & Robert Gillesse

23 June 2020

NDE Social Media Archiving Webinar

international institute of social history

network digitaal erfgoed

(Social Media) Archiving at IISH

- The IISH preserves and makes accessible one of the largest collections on socioeconomic history in the world
- Serving the research community and contributing to the preservation of global and local memory and heritage
- Social media archiving is a new undertaking for us but very significant for the IISH mission:
    - Women's Marches, Black Lives Matters movement, Coronavirus crisis in the Netherlands, informal labour networks

Social Media Archiving Tools Project

- NDE-funded research into tools for the harvesting of social media content
  - Free and open-source tools only
  - Focus on tools for capture (thus tools for preservation, processing, and/or access are in principle excluded)
- Series of Social Media Archiving Question Hour sessions during April and May 2020
- Final report expected ca. October 2020

From web to social media archiving and back again

- Since the mid-90s, web archiving has made strides
- Primarily using web crawlers (aka scrapers) to request pages from the host server of a website and then store the requested content
- Heritrix, NetArchive Suite, Web Curator Tool
- Well-suited for static content, i.e. pages that depend primarily on HTML to be loaded and rendered
- Social media can be captured by crawlers only to a limited extent if at all

- Social media websites are examples of dynamic content, i.e. pages that depend on scripts to be run in browsers, servers, or both, in order to be rendered
- The personalized nature of social media presupposes user interaction to trigger these scripts
- Tools have/are being developed that can mimic user interaction through a browser
- Brozzler, Browsertrix, Conifer (formerly Webrecorder)

- *But wait - there's more!*

Two broad approaches to archiving social media

“Look and feel” - saving pages

- Using browsers and crawlers to crawl and capture social media content as browsable pages
- Stored in WARC files
- Replayable by software like Replayweb.page, Webrecorder Player, OpenWayback, pywb, etc.

Structured data - saving information

- Connecting to a social media platform API to pull data
- Requires authenticated account - difficult with some platforms
- CSV, JSON files
- Usable with data analysis, visualization software

	I	J	K	L	
1	hashtags	mentions	in_reply_to_screen_name	twitter_url	text
2				http://twitter.com/HaneefICE/status/1008381298463920130	🙏🙏🔥🔥 "i caught her textin another nigga, how in hell is she mad at me?"
3	Amsterdam, Munich, UnveilingInfinity	jyothirmayah, SriSri		http://twitter.com/KamleshJhalta/status/1008381304696659971	RT @jyothirmayah: After #Amsterdam 🇳🇱 #Munich 🇩🇪 will Meditate w
4				http://twitter.com/Allwork_space/status/1008381307653689345	Dubbed "a Silicon Valley under one roof", the TQ tech hub in Amsterdam u
5		SriSri		http://twitter.com/ankitchitkara/status/1008381316902084608	RT @SriSri: People from across Europe gathered in Amsterdam to experie
6				http://twitter.com/Jooherfection/status/1008381345398181888	SDIFHDSDKASJHDASHD IT'S ME IN AMSTERDAM https://t.co/nvDW8A
7				http://twitter.com/p2kamsterdam/status/1008381419415064577	18:10:38 A1 13109 Rit 67890 Amsterdam Reinwardtstraat 1093GW
8				http://twitter.com/corgisland/status/1008381442580246529	I AM READY FOR AMSTERDAM
9				http://twitter.com/LarsWuijster/status/1008381456354349058	Dit zijn grote innovatieve stappen op gebied van duurzaamheid! https://t.c
10		urbanchestnut		http://twitter.com/STLCityNews1/status/1008381462373109760	Drinking an Amsterdam Pils by @urbanchestnut @ Amsterdam Tavern —
11				http://twitter.com/112zwnl/status/1008381468043825153	A1 13109 Rit 67890 Amsterdam Reinwardtstraat 1093GW https://t.co/GIcI
12				http://twitter.com/112zwnl/status/1008381478097612801	A1 13109 Rit 67890 Amsterdam Reinwardtstraat 1093GW https://t.co/c2c
13	EULAR2018	fibrofella		http://twitter.com/SimonRStones/status/1008381522716626945	RT @fibrofella: Leaving Amsterdam today after my first congress, #EULA
14		HanAltena, JohnVVD	HanAltena	http://twitter.com/MartinKarindeHa/status/1008381554299727873	@HanAltena @JohnVVD verdomme, excuez le mot: https://t.co/SF5PVy2
15	IndieBrew			http://twitter.com/Indie_Brew/status/1008381588252647424	Watch Death Cab Debut New Song "Summer Years" at Amsterdam Tour O
16		IndinNederlands, venurajamony		http://twitter.com/blissfulls/status/1008381594527297536	RT @IndinNederlands: Ambassador @venurajamony with Minister of Infra
17		patel4witham		http://twitter.com/TheSpeakingRog/status/1008381597815595008	RT @patel4witham: These political games have serious consequences.Th
18		37paday		http://twitter.com/gavin_steven/status/1008381605080072193	RT @37paday: I cannot believe I am now planning for a no deal Brexit at h
19				http://twitter.com/AkeebLDN/status/1008381614345342976	Fam this slyly reminds me of that movie Hostel🤔 https://t.co/5MUQJMw
20	meditation, IDY2018	ArtofLivingYoga, SriSri		http://twitter.com/saurabh1031/status/1008381616811593730	RT @ArtofLivingYoga: Gurudev @SriSri address yoga enthusiasts and lea
21	meditation, IDY2018	SriSri		http://twitter.com/anuvenks/status/1008381655785099265	Gurudev @SriSri address yoga enthusiasts and lead a #meditation at Ams
22		cristinar0218		http://twitter.com/lizetteSLS22/status/1008381662915203073	Done with finals. Now time to plan our trip to Korea and Amsterdam for Jir
23		ozy_cagla, BTS_twt, BTS__Europe		http://twitter.com/Soph3279/status/1008381704518578177	RT @ozy_cagla: Kijkkk!!!!! Bij Amsterdam Centraal😍😍 @BTS_twt @BTS
24	5YearsWithBTS	BTS__Europe, BTS_twt		http://twitter.com/duomo00/status/1008381706418708481	RT @BTS__Europe: 🇪🇺 #5YearsWithBTS 🇰🇷©@BTS_twt, thank you fo
25		SriSri		http://twitter.com/shreyagoyal3333/status/1008381728984006657	RT @SriSri: People from across Europe gathered in Amsterdam to experie
26	NONcentralCONF, GoMadrid	GoTechMadrid		http://twitter.com/pablo_lopez/status/1008381730703790086	RT @GoTechMadrid: Do you have your NON TICKETS for the #NONcentr
27				http://twitter.com/p2kamsterdam/status/1008381734801608710	18:11:54 A2 13176 Rit 67891 Amsterdam Meibergdreef 1105AZ AMC H3 Z
28		TheBulldogAMS		http://twitter.com/legendiscover/status/1008381741646663681	RT @TheBulldogAMS: Tomorrow is Cannabis Liberation Day here in Amst
29		patel4witham		http://twitter.com/Noritaeden/status/1008381763708702720	RT @patel4witham: These political games have serious consequences.Th
30				http://twitter.com/112zwnl/status/1008381777558298630	A2 13176 Rit 67891 Amsterdam Meibergdreef 1105AZ AMC H3 ZUID verl
31				http://twitter.com/112zwnl/status/1008381785691099136	A2 13176 Rit 67891 Amsterdam Meibergdreef 1105AZ AMC H3 ZUID verl
32		mxthadxmic		http://twitter.com/wexrelosers/status/1008381785875591170	RT @mxthadxmic: 16.06.18 Amsterdam, Holland. https://t.co/Fg483jYmbg
33		Huobi_Pro	Huobi_Pro	http://twitter.com/MediaDisney/status/1008381795673493504	@Huobi_Pro Posted... https://t.co/KZnvNHYKiF
34		SriSri		http://twitter.com/LatikaTharani/status/1008381805915865088	RT @SriSri: Conducted yoga and meditation in Museumplein, Amsterdam
35				http://twitter.com/NachoSilgoria/status/1008381808478744576	Que belleza de gol https://t.co/c0zuEpb1i1
36	5YearsWithBTS	BTS__Europe, BTS_twt		http://twitter.com/Kookies_teeth/status/1008381812224258048	RT @BTS__Europe: 🇪🇺 #5YearsWithBTS 🇰🇷©@BTS_twt, thank you fo

“Look and feel” tools

Brozzler: Uses the Chromium or Chrome browser to automatically interact with pages and capture them

- Dashboard available to view crawl status, command-line control and configuration
- Actively developed by the Internet Archive

Browsertrix: System for automated capture and display of web content based on browser crawlers and crawling behaviours

- Requires Docker
- Basic GUI available but advanced configuration requires YAML files and command line

Webrecorder Desktop/Conifer: Uses remote browsers to record pages as the user interacts with them

- Manual capture, very user-friendly
- Some automation is available (Autopilot to scroll and load specific social media pages automatically)

“Structured data” tools

Twarc: Command-line tool for capturing Twitter API data

- Requires a registered developer account and Twitter app
- Very supportive community (DocNow)

TAGS: Google Sheet template that captures and stores Twitter API data

- User-friendly
- Fewer filtering options

Social Feed Manager: Application for harvesting of Twitter, Flickr, Tumblr and Sina Weibo API data

- GUI
- Requires some technical skill for installation and maintenance

API harvesting vs. crawling

- Content from API harvesting can be readily used with a variety of data analytics tools
  - It allows for large-scale collecting
  - It lacks the visual experience of navigating the pages
  - Platform terms and conditions significantly limit our ability to make collections accessible and to also guarantee their integrity (e.g. tweet IDs policy)
- Crawling social media and storing it in WARCs provides easily replayable content
  - Browsing experience is (mostly) preserved
  - Source code is captured and available but requires more researcher tinkering
  - Difficult for large-scale collecting

➢ *The approach to archiving social media content we choose affects the uses that the content can be put into in the future*

Preliminary findings

- Facebook is actively resisting archiving
    - Restricted access to API, no broad accommodations for researchers, tight control of data dissemination (e.g. Facebook Ad Library)
    - Constant changes put developers in an arms race to keep up
- Twitter's restrictive re-publishing policy may cause issues with integrity of collections
- Platforms like Instagram, YouTube, Vimeo can also be archived by scraping only images and videos, rather than by capturing entire pages
- In some cases, co-operating directly with creators/users to obtain content might be the only solution
- In many cases, we must jump in before we have time to figure everything out

So many tools, so little time

Choosing the right tool largely depends on the context, needs, and circumstances of each organization

Social media archiving is as much a technical issue as it is a conceptual and ethical one

Some considerations for starting out

- Flexibility
  - There is no single tool to capture all kinds of social media content in all forms - mixing and matching is needed
- Planning
  - Finding out which your designated community/user group is and what their needs are, and planning about ethical and legal aspects of handling sensitive data
- Documentation
  - Ability to record provenance and context of social media content consistently
- Investing in skills building
  - Combination of technical and archival knowledge, awareness of research processes and needs, constant learning

Into the future

- ➔ What is considered social media? How do we archive WhatsApp and Telegram? How about TikTok, Discord, and Twitch?
- ➔ What can we learn from researchers already working with data?
- ➔ What can we learn from the fields of data curation, research data management, digital asset management?
- ➔ How can we use our position as heritage professionals to advocate for ethical social media archiving in the service of research, memory and accountability?

Thank you!

Social media archiving tools FAQ

Social Media Archiving tools preliminary report

Contact us here:

zefi.kavvadia@iisg.nl

robert.gillesse@iisg.nl