Vragenuur over het gebruik van social media archiveringstools
#coronacollectie

Verslag 30 april

Met 14 deelnemers stond het vierde vragenuur op donderdag 30 april in het licht van het gebruik van social media archiveringstools. Zefi Kavvadia en Robert Gillesse (IISG) begeleidden het vragenuur.

Over het archiveren van Instagram
Het onderzoek van IISG kijkt onder meer naar het archiveren van Instagram posts. Er was ooit een tool (Lentil), bedoeld om Instagram met de Instagram API te archiveren, maar de API van Instagram werd gewijzigd naar een graph-API en werkt nu niet meer. Er bestaan aardig watl tools voor 'dataset scraping', alleen zijn deze niet ontwikkeld met archivering als doel. Voorbeelden zijn Instamancer en Instalooter. In het uiteindelijke rapport van IISG zal hier dieper op in worden gegaan.

Webrecorder werkt ook prima om Instagram vast te leggen. De autopilot feature van Webrecorder doet het goed als je een pagina wilt archiveren. Dit is een prima alternatief als je niet per se de API data wilt. Nadeel: als je de comments bij beelden of video's wilt archiveren met Webrecorder, moet je dat per beeld doen. Het is een keuze om Webrecorder alvast een deel van het archiveren automatisch te laten doen en daarna handmatig te controleren wat er mee is gekomen.

Ook via Brozzler kun je Instagram archiveren. Vanuit het onderzoek is dit nog niet geprobeerd, maar dat is wel gepland. Je moet Brozzler telkens configureren voor ieder afzonderlijk social media platform. Instructies daarvoor staan op Github.

Over Archiv-It en hun abonnement
Ook via de Archive-it dienst kun je gebruik maken van Brozzler. Wie zich abonneert op Archive-it, heeft ook volledige controle over de tools. Archive-it heeft itt Brozzler ook een graphical user interface. Je kunt alles opslaan in je eigen repository en cloud storage bij hen kopen. Archive-it also hosts already made data that you can access online on the Wayback Machine.

Over tools die bepaalde content (HTML of video) uit een WARCfile extraheren of deselecteren
Er bestaat een fikse lijst met dit soort tools, zie The Awesome Webarchiving List: https://github.com/iipc/awesome-web-archiving. Om er twee uit te lichten (die zijn gebruikt door het IISG):

- Warcit is een command line tool die het omgekeerde doet: het extraheert niets, maar bundelt HTML files in een Warc file.
- Met de tool Warc-extractor j kun je ook onbedoelde info (URL's) uit je WARC file filteren. Dat doet Webrecorder niet, die pakt de hele pagina en daarin kun je niets veranderen. Warcit is dus een handige tool als je je warc file wilt opschonen.

Over het archiveren van YouTube video's en Twitch video's
Daar zijn aparte tools voor, bijvoorbeeld youtube-dl en Webrecorder. Twitch video's (veel gebruikt door gamers die willen streamen) werken anders, daar kun je Webrecorder niet voor gebruiken. Het

is beter om de makers van deze video's te benaderen om de oorspronkelijke videobestanden op te vragen en toestemming te vragen om deze te archiveren.

Over de technische kwaliteit van geharveste video's
Youtube.dl en Webrecorder proberen altijd de video met de hoogste kwaliteit te harvesten. Hou daarbij wel ingedachte dat juist video's veel opslagruimte in beslag kunnen nemen. Houd daarom in je achterhoofd wat je wilt archiveren en selecteer.

Een discussie over wat je wilt bewaren volgt. Is het erg als je de comments van een Instagrampost mist? Als archivaris wil je alles bewaren en weten of je iets van belang hebt gemist. *An archive is not a data storage, it needs context.*

In dat licht bezien: het team van Webrecorder heeft ook BrowserTrix ontwikkeld. Deze tool heeft een replay-mode. Als je een collectie hebt gearchiveerd, kun met een screenshot een snel overzicht krijgen of de vastgelegde de website er goed uitziet en of er iets mist. Dit is handig voor een eerste review van je content, zonder dat je handmatig hoeft te checken of alle onderdelen van de pagina zijn vastgelegd. Uit onderzoek blijkt dat deze tool niet erg accuraat is. Voor ieder social media platform is er namelijk een ander script, en dit zorgt ervoor dat de screenshot niet de hele pagina toont terwijl de hele pagina wél is vastgelegd. Beste manier is om iedere pagina zelf te openen en te kijken of alles erin staat, inclusief iconen, grafieken en foto's. Tip: open de bestanden en bekijk ze zelf, eventueel met add-ons om de code van de pagina te lezen en te zien waar er zaken missen. Nog een tip: lees de logs (als deze worden gemaakt, Webrecorder doet dit niet) om te kijken of alles mee is genomen met het archiveren.

Over het belang van het documenteren van collectieproces en de gebruikte tools
Uit gesprekken met onderzoekers blijkt dat zij graag willen weten waar de data vandaan komen en hoe de collectie tot stand is gekomen, precies zoals bij een analoge collectie. We zouden zeker moeten proberen om de technische info zoals software, versie en extra componenten in de tools ook te geven. De informatie in een Warc-bestand benoemt dit niet automatisch.

IISG neemt tijdens het onderzoek naar social media archivering ook mee hoe je goede beschrijvingen van website-collecties en social media collecties opstelt. IISG probeert de optimale manier te vinden om hiervoor EAD archiefbeschrijvingen te gebruiken: niet alleen beschrijvende info maar ook technische informatie om dit in catalogi op te kunnen nemen. Dit zou onderdeel moeten zijn van een digital preservation system of een repository.

Rondvraag over beschrijvingen die deelnemers hebben gemaakt van websites of social media.

SSA heeft twitter data sets vastgelegd: het is niet makkelijk om te zien welke informatie erin zit en wat er mist. Als je een hashtag harvest, dan worden de reacties bijvoorbeeld niet meegenomen. Moet je dit doorgeven aan de onderzoekers in zo'n beschrijving? Er zijn dan best veel restricties te noemen van bepaalde software.

'How much context does an researcher need to do his research?' is een al oudere bestaande vraag onder archivarissen. Met deze nieuwe online archiveringswereld zou een bredere discussie hierover goed zijn.

Over Coosto
Wouter Brunner vertelt vervolgens iets over Coosto. Dit is een tool om je eigen social media te onderhouden en om social media analyses te maken. Je kunt er niet mee archiveren, maar Coosto harvest veel media zelf (FB, Insta Twitter en NU.nl , nieuwsblogs). Er is dus heel veel content beschikbaar, waarover het bedrijf ook social media analyses maakt. Dit heeft niets van doen met de doelstellingen die erfgoedinstellingen voor ogen hebben. Toch zijn die analyses interessant om te zien welke social media bronnen zij gebruiken. Coosto maakt worksheets met alle data die je samen met de sources kunt gebruiken om een collectie compleet te maken. Voorbeeld: Sentiment analyse van corona-tweets. In die zin kan Coosto dus een belangrijke hulpmiddel zijn bij de selectie van de te archiveren materiaal. Wouter zal een voorbeeld posten in de #coronacollectie-mailgroep.

Update research IISG
Zefi onderzoekt momenteel een tool die zich echt op Facebook richt. Het verslag over de Brozzler, Webrecorder en BrowserTrix tools is bijna klaar, daar komen nog user stories bij vanuit een webarchivist-invalshoek. Dit verslag is binnenkort beschikbaar via https://confluence.socialhistoryservices.org/display/ESMAT/User+Stories+for+Social+Media+Archiving+Tools+Testing+at+IISH.

Feedback hierop is welkom.

Over het volgende vragenuur
Het volgende vragenuur sociale media archivering vindt plaats op donderdag 7 mei van 11:00 tot 12:00 uur. Samen met auteursrechtjurist Annemarie Beunen van de Koninklijke Bibliotheek, digitaal archivaris Mirjam Schaap van het Stadsarchief Amsterdam en juridisch adviseur informatie- en archiefrecht Noor Schreuder van het Nationaal Archief verkennen we de juridische aspecten van social media archivering. Het vragenuur is zoals altijd toegankelijk via deze link: https://zoom.us/j/525710213.

Vragen?
Heb je nog vragen naar aanleiding van het vragenuur, dit verslag of anderszins, maak dan gebruik van de #coronacollectie-mailgroep via NDE-coronacollectie@googlegroups.com. Of lees alvast de veelgestelde vragen uit eerdere vragenuren.