Best practices voor de archivering van sociale media in Vlaanderen en Brussel

Fase 3. D3
Verslag van de pilootprojecten betreffende toegang tot en hergebruik van gearchiveerde sociale media

Thomas Peeters, Fien Messens, Katrien Weyns

30 augustus 2023

Dit is de derde deliverable (D3) van de derde fase binnen het project *Best Practices voor de Archivering van Sociale Media in Vlaanderen en Brussel* getrokken door KADOC-KU Leuven en meemoo.

In de derde projectfase werd raadpleging en gebruik van socialemedia-archieven en socialemediaberichten onder de loep genomen. Hoe kunnen erfgoedorganisaties raadpleging door verschillende gebruikers met verschillende noden faciliteren? Kunnen we zelf met de gegevens aan de slag? Tijdens pilootprojecten namen we samen met verschillende projectpartners de proef op de som.

Contents

1. Pilootproject “Opmaak oeuvreoverzicht van kunstenaars via Instagramposts”

Volgt nog.

2. Pilootproject “Ontsluiting naar onderzoekers vanuit een erfgoedorganisatie”

2.1. Opzet

Deze piloot focuste op het gebruik van verzamelde socialemedia-archieven in academisch onderzoek. Met name wilden we een beter zicht krijgen op het gebruik voor text mining. Gedurende de derde fase vonden er verschuivingen plaats bij Twitter (X), wat een impact had op de uitwerking van de geplande piloot. Verder wilde Letterenhuis mee deze piloot vormgeven. We kozen ervoor om het opzet bij te sturen en onze blik te verruimen. In plaats van het faciliteren van één onderzoekscase, brachten we ervaren onderzoekers samen uit verschillende disciplines en universiteiten die input gaven op de datasets die we in fase 1 en 2 verzamelden en we in te toekomst nog kunnen aanleveren.

2.2. Gebruikte datasets

De aanwezigen werden geïntroduceerd in het project en het verzamelde materiaal. Datasets in verschillende formaten afkomstig van KADOC en Letterenhuis werden tijdens een samenkomst getoond en van uitleg voorzien. Volgende data werden voorgelegd:

- WARC-bestanden van het Facebook account van Gaea Schoeters
- Download van een Facebookaccount (txt, html, jpg, …) gemaakt en gedeponeerd door een archiefvormer, met name Tom Naegels
- Ruwe data (json, txt, mp4, mkv, jpg) verzameld via Tartube afkomstig van het Youtube-kanaal van CD&V
- Ruwe data (json, txt, jpg) verzameld via Instaloader afkomstig van het Instagramaccount van CD&V

2.3. Interactie

Vervolgens werd aan de hand van stellingen feedback gevraagd op het gepresenteerde materiaal en de archiverings-en ontsluitingspraktijk van (private) archieven. Aan de tafel zaten vertegenwoordigers van onderzoeksgroepen aan UA (met name een onderzoeker Computational Linguistics, Psycholinguistics and Sociolinguistics), KU Leuven (met name een onderzoeksondersteuner Digital Humanities) en UGent (met name een onderzoeker imec-mict-UGent). Deze beperkte groep was zeer onderlegd en voldoende representatief om ook het werk van collega’s mee te nemen in hun feedback. Onderzoekers van IMS (KU Leuven), van het Antwerp Centre for Digital humanities and literary Criticim en het Brussels Platform for Digital Humanities en andere genodigden verbonden aan bovenstaande initiatieven waren verhinderd.

2.4. Conclusie

Alle onderzoekers gaven aan dat ze de (door archieven) verzamelde datasets zouden evalueren en gebruiken binnen onderzoek, ongeacht of het nu om een kleine of een grote dataset zou gaan en ongeacht in welk formaat deze wordt aangeboden. Twee van de drie aanwezigen kenden dat aanbod niet. Voor onderzoekers is het interessant dat archieven een selectie maken van socialemediaberichten die zij binnen hun collectieprofiel relevant vinden en waarbij ze ook kunnen aangeven waarom die relevant is. Onderzoekers staan open voor suggesties vanuit archieven. Er is ook interesse in meer uitwisseling tussen archieven en onderzoekers: wie is waar mee bezig? En wie kan wat aanbieden? Zo kunnen vraag en aanbod beter op elkaar inspelen. Het verduurzamen van dergelijke uitwisseling is echter niet evident, gezien onderzoeksmandaten vaak beperkt zijn in tijd. (*Misschien kan dit een plaats krijgen binnen het praktijknetwerk?)*

Private archieven verzamelen veelal kleine, kwalitatieve sets (volledig, met behoud van look-and-feel,

gestructureerd en gecontextualiseerd). Dit is waardevol voor kwalitatief onderzoek. De opgebouwde datasets zijn vaak onvoldoende groot om grootschalige textminingprojecten uit te werken bv. covidonderzoek, maar de kleine sets kunnen andere datasets aanvullen wanneer ze voldoende metadata (over volledigheid, inhoud, …) en structuur bevatten. Het verzamelen van grote datacorpora gebeurt nu door onderzoekers zelf, maar is zeer tijdsintensief. Het aanvullen van verzamelde corpora met gearchiveerde datasets zou dus een tijdswinst opleveren. Het is echter wel van belang dat gegevens in bulk kunnen opgehaald worden.

Academici zijn bereid om zelf extracties te maken uit WARC-bestanden en andere bestanden als dat nodig is voor het onderzoek, maar tegelijk wordt er opgemerkt dat niet alle onderzoekers de technische vaardigheden hebben of er voldoende in onderlegd zijn. De aanwezigen zien een (gedeelde) taak weggelegd voor archieven om samen met onderwijsinstellingen beginnende onderzoekers te introduceren in socialemedia-archieven, en bij uitbreiding webarchieven, als bron en in de methodieken om deze bronnen te capteren. Door de snelle technologische veranderingen is het immers ook voor onderzoeksondersteuners en onderzoekers moeilijk om op de hoogte te blijven van alle mogelijke tools. Archieven zouden in een seminarie deze introductie kunnen geven, maar (academische) onderzoekers zouden ook meer betrokken kunnen worden in het praktijknetwerk sociale media archiveren.

3. Pilootproject “Hergebruik in het kader van gezichtsherkenning door een erfgoedorganisatie”

Volgt nog.

4. Pilootproject “Ontsluiting naar en hergebruik van data in het onderwijs (vanuit een erfgoedorganisatie)”

4.1. Opzet

In het kader van het FARO-initiatief “erfgoedklasbakken” grepen we de kans om samen met enkele content partners na te gaan of een dataset met gecapteerd materiaal uit sociale media accounts bruikbaar kan zijn als lesmateriaal. In tweede instantie werden enkele manieren van ontsluiting, de presentatie van zo’n datasets in lescontext, getest.

Dit pilootproject werd uitgevoerd als een samenwerking tussen KADOC en content partners Amsab-ISG en CAVA-VUB. Elke content partner gaf een aantal lessen die als bron- en illustratiemateriaal gecapteerde datasets gebruikten. Hoewel de uitwerking van elke les verschilde van content partner tot content partner werd er wel steeds gewerkt rond dezelfde centrale vragen. Hoe evolueert maatschappelijke communicatie? Zijn nieuwe, online vormen van communicatie duurzaam en voor altijd online? Wat willen we uit nieuwe bronnen wel of net niet bewaren? Met die issues als raamwerk werkten we respectievelijk omtrent de recente Boerenprotesten, het Abortusdebat en Digitaal Erfgoed in het algemeen. Telkens werd ouder, fysiek materiaal afgezet tegen datasets al nieuwe bron of nieuw erfgoed. De bedoeling was dan ook simpel, nagaan of zo’n digitale bron ontsloten en gebruikt kan worden als lesmateriaal.

Concreet vonden er een tiental lessen plaats. KADOC nam twee daarvan voor z’n rekening. Er gingen twee lessen door, telkens in een 5dejaarsklas ASO. De les werd opgebouwd rond de, toen zeer recente, stikstofprotesten van boeren in België. Oudere, fysieke stukken uit de collectie werden vergeleken met nieuwe vormen van communicatie zoals sociale media. De leerlingen dachten aan de hand van diverse datasets na over de standvastigheid van sociale netwerken, de elementen die schuilgaan achter een post en de mogelijkheden van dat materiaal als bron voor de toekomst.

De collega’s van CAVA-VUB gaven op hun beurt zeven lessen verspreid over vier verschillende scholen. De lessen werden gegeven aan een brede waaier klassen, van OKAN-klas met leerlingen van 9 tot 14 jaar,

over een brede immersieklas tot enkele klassen uit een 5de en 6de middelbaar. Het centraal thema van elke les was de vraag of het internet een duurzaam iets was, een informatiebron die voor altijd online staat. Vanuit dat vertrekpunt dachten de leerlingen na over wat ze zelf posten op sociale media, wat daarmee kan gebeuren en hoe ze zelf kijken naar de vergankelijkheid van online informatie.

Amsab-ISG zat tot slot in diezelfde lijn. De nadruk van de les lag eveneens op de evolutie van communicatiemiddelen en de gevolgen daarvan. Rond het thema van abortus, en de manier waarop de maatschappij hierover in debat gaat, werden fysieke, papieren bronnen zoals affiches en flyers afgewogen tegen nieuwe vormen van communicatie. Tijdens de piloot gingen zo'n tien lessen door, telkens in een derde graad middelbaar.

4.2. Gebruikte datasets

Het eigenlijk lesmateriaal bestond telkens enerzijds uit ouder, fysiek materiaal uit de collecties en anderzijds uit nieuwe digitale bronnen. KADOC en AMSAB gingen aan de slag met flyers en affiches uit de omvangrijke affichecollectie mee, CAVA-VUB werkte met dagboeken en schoolagenda's. Tegenover dat materiaal stonden dan zeer recente datasets gecapteerd uit sociale media accounts. Elk van de content partners bouwde een eigen dataset op en presenteerde die telkens op een andere manier. Bij de les van KADOC stonden een aantal tweets van CD&V en Boerenbond centraal, telkens berichten over het verloop of de nasleep van de protesten. Die datasets werden weergegeven als screenshot, als JSON-document en als WARC-document, een breed gamma aan captaties dus. Zowel het screenshot als het JSON-document konden simpel en statisch gepresenteerd zonder gespecialiseerde tools. Het WARC-bestand werd dynamisch gepresenteerd via replayweb.page. CAVA-VUB werkte met één enkele post om die dan in groep in depth uit te spitten. De post in kwestie haakte niet specifiek in op een maatschappelijk thema maar was bewust een generieke persoonlijke post. Op die manier werd de link gelegd met ander archiefmateriaal zoals dagboeken of agenda's, allemaal materiaal dat op het eerste zicht als een persoonlijk iets zonder veel archiefwaarde wordt beschouwd. De post werd zowel als screenshot en als JSON-document gepresenteerd tijdens de les, zonder verdere tools. Ze werden dus statisch geprojecteerd of als print verdeeld, dit omdat het dynamisch presenteren van een WARC net teveel voeten in de aarde had om efficiënt te doen in amper 50 lesminuten. De lessen van Amsab gebruikten een dataset bestaande uit 10 individuele posts. Om die gecapteerde posts te presenteren werd er gebruik gemaakt van een tijdelijke en publiek verborgen webpagina aangelegd. Daarop vonden leerlingen dan de posts als embedded sociale media.

In elk van de drie cases werd er gewerkt met posts die op het moment van de les online nog steeds raadpleegbaar waren op de gecapteerde accounts. Vooral met het oog op de juridisch grijze zone rond de ontsluiting van de capteerde posts buiten een leeszaal werd er niet gewerkt met archiefkopieën van (mogelijk) verwijderde posts.

4.3. Ontvangst

Over het algemeen werd het gebruik van deze relatief nieuwe bron goed ontvangen door het doelpubliek. Het thema van evoluerende communicatie werd kracht bijgezet door het concrete van de datasets, zonder zou alles te vaag gebleven zijn. Leerlingen reageerden met andere positief op de actuele inhoud van het materiaal en op het feit dat dit een bron uit hun onmiddellijke leefwereld is.

De leerkrachten zagen eveneens veel toekomst in het materiaal. De nabijheid bij de leefwereld van de leerlingen en de actualiteit van het materiaal werden aangehaald als positieve kanten. Enkel de technische kant schrikte nog af. Het omslachtige van een WARC inladen en presenteren vergde volgens hen net iets te veel tijd en voorkennis om vlot te gebruiken in een lescontext of als huistaak. Een eigen webpagina opzetten om posts te presenteren is vervolgens ook organisatorisch een moeilijk haalbare kaart voor de meeste leerkrachten.

4.4. Evaluatie

Het gebruik van sociale media in een lescontext heeft zijn voor- en nadelen. Inhoudelijk is het materiaal zeer dankbaar in een educatieve context. De posts bevatten een actuele inhoud die maatschappelijk leeft, omvat vaak diverse perspectieven die in korte en bondige boodschappen naar voren komen. Dit is zeker

zo met het oog op breed gedragen onderwerpen zoals verkiezingen. De keerzijde daarvan is dan weer dat datasets snel gedateerd kunnen aanvoelen. Om écht up-to-date te zijn mogen datasets niet té ver op voorhand gecreëerd worden of net eerder algemeen zijn. Wat vandaag een hot topic is, is morgen mogelijk weer oud nieuws. In het verlengde daarvan is het ook niet inzetbaar voor élk thema. Een onderwerp moet natuurlijk leven in de maatschappij alvorens het veel tractie krijgt op sociale media. Zonder posts kan er natuurlijk geen dataset gecapteerd worden.

Technisch zijn er ook nog een aantal issues. Elk van de in de piloten gebruikte formaten had eigen voor- en nadelen. Waar het screenshot en het JSON-document de bovenhand hadden qua gebruiksgemak, stond het WARC-bestand pal aan de leiding als het aankomt op volledigheid, duurzaamheid en leesbaarheid. Vanuit een educatief opzicht is werken met formaten die look-and-feel bewaren interessanter. WARCs gebruiken is dan weer niet evident door de noodzaak van een specifieke player, waardoor het moeilijk is om het op te nemen in een powerpoint of huistaak. JSON is op zijn beurt te omvangrijk qua tekst en te minimalistisch van uitzicht om vlot gelezen te worden door lagere of middelbare scholieren. Kortom, de ideale tool om sociale media snel en eenvoudig te presenteren in een lescontext is voorlopig nog niet voor handen.

Een dataset is in deze context bij voorkeur zo gebald en gericht mogelijk. Hele hashtagstreams zijn te omvangrijk om in één of twee lesuren efficiënt te gebruiken. Dat toont zich ook in de gegeven lessen tijdens deze piloot. Telkens stond er één, enkele of een tiental posts in een spotlight, ver verwijderd van de aantallen die onderzoekers wensen te gebruiken. Inhoudelijk heeft het materiaal ook nood aan een kritische blik.

4.5. Conclusie

Hoewel het materiaal inhoudelijk zeer dankbaar en erg interessant is voor gebruik in een educatieve context blijkt de technische kant nog te wensen over te laten. De geteste manieren van presenteren zijn op dit moment nog niet optimaal voor efficiënt gebruik in een klassituatie. De meest laagdrempelige optie, het statisch als screenshot of print presenteren van een post, laat te wensen over wat betreft volledigheid. Ingebedde dynamische media vallen bijvoorbeeld al uit de boot. De meer dynamische optie zoals het tonen van een WARC via *replayweb.page* is tijdsintensief. Er moet bovendien zeer gericht gecapteerd worden bij deze methode, in een WARC die het hele account bevat is het bijna onmogelijk een specifieke post efficiënt terug te vinden zonder kostbare minuten van een lesuur te verliezen. Hoewel een eigen webpagina met daarop de gebruikte posts ingebed handig in gebruik tijdens de eigenlijk les is, is de voorbereiding daarop logischerwijs niet evident.

5. Pilootproject “Ontsluiting naar een breed publiek vanuit een archiefinstelling”

5.1. Opzet

Het doel was om een workflow voor ontsluiting te ontwikkelen voor cultureel-erfgoedinstellingen die sociale media archiveerden. Deze bevat specifieke stappen en richtlijnen voor het ontsluiten van gearchiveerde sociale media data naar het brede publiek, waaronder onderzoek vanuit persoonlijke interesse of meer professionele context als erfgoedwerker, journalist of academicus.

5.2. Methodologie

Tijdens het pilootproject is de volgende aanpak gebruikt:

- **Stap 1: Survey noden breed publiek:** Er werd een overzicht gevormd van de noden en vereisten die het brede publiek geïmplementeerd zou willen zien als zij sociale media wensen te bekijken. Het Belgische publiek werd met beroep van een survey hierover bevraagd (zie verslag survey breed publiek).
- **Stap 2: Overzicht playback-landschap**: Het coördinerend team heeft contact opgenomen met hun netwerk van nationale en internationale experts, waaronder de International Internet Preservation Consortium (IIPC), om inzicht te krijgen in de tools die goed presteren op de markt voor het ontsluiten van gearchiveerde data (en die makkelijk in de omgang zijn). Er werd ook onderzoek gedaan naar de tools die het meest worden gebruikt in België voor het ontsluiten van gearchiveerde sociale media data.
- **Stap 3: Testen van tools**: Uiteindelijk werden de meest gangbare tools die mogelijk gemakkelijk te gebruiken zijn voor een breed publiek getest. Dit stelde het projectteam in staat om de functionaliteit, effectiviteit en gebruikerservaring van elke tool te beoordelen.
  De tooltestings van meemoo, werden eveneens benut als bron van informatie en ervaring met betrekking tot verschillende tools voor het ontsluiten van gearchiveerde data.
- **Stap 4: Haalbaarheidsbeoordeling**: Er is een grondige beoordeling uitgevoerd om de haalbaarheid van verschillende tools voor het beoogde brede publiek te bepalen. Hierbij werden criteria zoals gebruiksvriendelijkheid, technische vereisten en compatibiliteit overwogen. De beperkingen en belemmeringen van elke geteste tool zijn hierbij ook zorgvuldig vastgelegd

5.3. Resultaten

Stap 1: Noden breed publiek (zie apart verslag)

Stap 2: Overzicht playback-landschap

Naast het vastleggen van de behoeften voor de toegankelijkheid van gearchiveerde sociale media data voor een breed publiek, hebben we als tweede stap een overzicht gemaakt van de verschillende beschikbare tools voor het herbekijken van sociale media. Dit omvatte een uitgebreide evaluatie van voornamelijk open-source, waarbij we ons richtten op functionaliteiten, gebruiksgemak en de mate van actieve ontwikkeling en ondersteuning. Er werd zowel binnen België (bv. BESOCIAL project) als internationaal (bv. IIPC gemeenschap) gekeken welke tools vandaag het meest gebruikt worden voor een breed publiek.

Op basis van dit uitgebreide overzicht hebben we een zorgvuldige selectie gemaakt van de tools die nog steeds actief werden bijgewerkt en die geschikt zijn voor gebruik door een breed publiek. Deze tools, voldeden vanop het zicht aan onze criteria op het gebied van betrouwbaarheid, stabiliteit en gebruikersvriendelijkheid. De “goedgekeurde” tools worden in een volgend stadium grondig uitgetest om te zien wat de beste aanpak voor het best practices project is.

Tool	Passend voor breed publiek?	Notities
Sol R wayback	Waar	
openwayback	Onwaar	Niet langer actief in ontwikkeling. Voor een nauwkeurige weergave van webarchieven raadt de IIPC aan om Webrecorder's pywb te gebruiken. Voor degenen die momenteel OpenWayback-instanties hosten, biedt de documentatie van pywb een overgangsgids.
replayweb.page	Waar	
Waybackmachine	Onwaar	Niet in lijn met de scope van het Best practices project.
Memento tracer	Onwaar	Niet frequent gebruikt binnen de IIPC gemeenschap. Eveneens ook weinig opvolging bij *bugs*.
Archive-it	Onwaar	Commerciële tool. Getest in een eerdere fase van het project, maar financieel te duur bevonden.
WAIL	Onwaar	Niet frequent gebruikt binnen de IIPC gemeenschap. Eveneens ook weinig opvolging bij *bugs*.
Webrecorder Player	Onwaar	Webrecorder Player is vervangen door ReplayWeb.page. Gebruikers van Webrecorder Player worden aangemoedigd om over te stappen naar de nieuwste ReplayWeb.page App.
Conifer	Waar	
pywb	Waar	
ARCH	Onwaar	Te uitgebreid en meer de focus op analyses dan op het voorzien van een toegangsplatform.
Warclight	Onwaar	Installatie is te ingewikkeld.
WacZ	Waar	Geen playback-tool. Container-formaat voor WARC bestanden.
Archives Unleashed Toolkit	Onwaar	Meer de focus op analyses dan op het voorzien van een toegangsplatform.

Tabel 2: overzicht meest courante playback-tools met oordeel of het geschikt is voor ontlsuiting voor een breed publiek. Het is belangrijk op te merken dat deze lijst niet exhaustief is en dat er andere tools beschikbaar kunnen zijn die niet in de tabel worden vermeld.

Tijdens onze analyse konden we naast het overzicht hierboven ook **drie types scenario's** onderscheiden om gecapteerde sociale media data beschikbaar stellen. Elk scenario heeft zijn eigen voordelen en nadelen in termen van implementatie, gebruiksgemak en functionaliteiten.

Scenario A De gearchiveerde data zal downloadbaar zijn in WARC formaat (en Json-formaat). Er worden handleidingen voorzien om gebruikers te gidsen naar online replay tools	**Scenario B** De gearchiveerde data zal via link raadpleegbaar zijn. Er wordt direct doorverwezen naar de replaypagina waar de data interactief te bekijken is.	**Scenario C** De gearchiveerde data zal in een platform worden geplaatst waarbij de bezoeker zowel kleinschalige als uitgebreide zoekfuncties kan uitvoeren.
Voordelen: Vrij gemakkelijke implementatie	- Weinig drempels voor het brede publiek om de gearchiveerde data te bekijken (gering aantal "kliks") - Mogelijkheid om verschillende captaties over tijd te bekijken	- Veel DH tools ter visualisatie: wordcloud, Linkgraph, Domain stats, Link graph Gephi export, Ngram Netarchive - Zoek en filterfunctionaliteiten
Nadelen: Het is aan het publiek om zelf actie te ondernemen om een replaytool te gebruiken. Er zullen hierbij veel kliks van de gebruiker nodig zijn om de data te kunnen raadplegen.	- Het publiek kan de data enkel bekijken als het publiek toegankelijk wordt gemaakt. - Geen mogelijkheid tot zoeken in website	- Veel implementatiewerk (opzetten server) - Onderhoud?
Tools: WARC bestanden of meer gestructureerde WACZ containers kunnen herbekeken worden met ReplayWeb.page of Conifer	Een hele reeks (*zie overzicht*). Hieronder de voornaamste: - Pywb: voorbeeld - SolrWayback: voorbeeld - OpenWayback: voorbeeld - Conifer: voorbeeld	- Een hele reeks (*zie overzicht*). SolrWayback: demo-platform, netwerkvisualisatie, etc
Middelen: - Tijd: creatie/bijschaven handleidingen Replaytools + evt. creatie WACZ container	- Tijd: implementatie/opzetten webserver + uploaden collectie + monitorring + actie ondernemen bij foutmeldingen + … - Kosten: opzet en onderhoud webserver	- Tijd: implementatie/opzetten webserver + uploaden collectie + monitorring + actie ondernemen bij foutmeldingen + … - Kosten: opzet en onderhoud webserver

Tabel 2: overzicht types scenario bij playback-tools.

Deze onderverdeling gaf ons een beter idee van welke tools voorhanden zijn per niveau en wat haalbaar is binnen het kader van het project.

Stap 3: testen van tools

Voor de derde fase werden de meest gangbare tools die gemakkelijk in gebruik zijn voor een breed publiek getest. Dit stelde het projectteam in staat om de functionaliteit, effectiviteit en gebruikerservaring van elke tool te beoordelen.

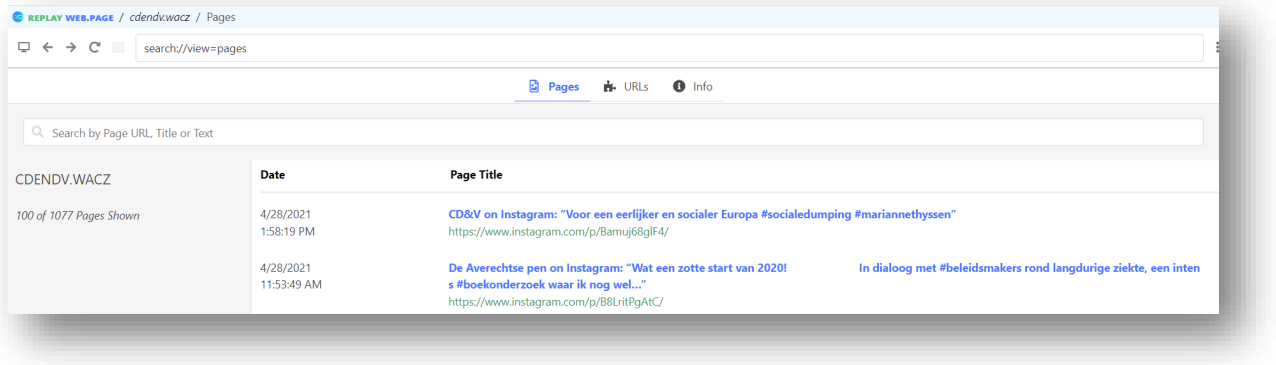
Playback tool	**Beschrijving**
Replayweb.page	ReplayWeb.page is een op de browser gebaseerde viewer die webarchiefbestanden laadt die door de gebruiker worden verstrekt en deze weergeeft voor herbeleving in de browser.
Py-Wacz	WACZ is niet een playback tool, maar eerder een bestandsformaat dat het verpakken en delen van webarchiefcollecties op het web mogelijk maakt. Deze "container" bevat alle gegevens die nodig zijn voor het weergeven van gearchiveerde inhoud, evenals contextuele informatie die nodig is voor gebruikers om het te interpreteren. Het kan gebruikers meer structuur bieden bij het bekijken van

	gecapteerde sociale media. Met Py-Wacz kan je gearchiveerde sociale media als WARC omzetten naar WACZ.
Archiveweb.page	Je kan WACZ rechtstreeks aanmaken als je crawlt met Archiveweb.page.
SolrWayback	SolrWayback is een open-source softwaretool waarmee gebruikers toegang kunnen krijgen tot en zoeken in gearchiveerde webpagina's die zijn opgeslagen in het Solr-indexformaat, waardoor effectieve en flexibele exploratie van het webarchief mogelijk is.
Conifer *(voorheen Webrecorder.io)*	Conifer is een browser-platform met een gebruiksvriendelijke interface dat zich voornamelijk focust op het capteren van content. Het biedt echter ook toegang tot gearchiveerde inhoud en de mogelijkheid wordt geboden om gecapteerde content te bekijken en te delen.
Pywb	Pywb is een python softwaretoolkit die een replay-omgeving biedt om gecapteerde WARC-bestanden te bekijken en te doorzoeken zoals ze waren op het moment van archivering.

Tabel 1: overzicht van de geteste playback tools door KADOC-KU Leuven en meemoo.

WACZ is niet een playback tool, maar eerder een bestandsformaat dat het verpakken en delen van webarchiefcollecties op het web mogelijk maakt. Het kan gebruikers meer structuur bieden bij het bekijken van gecapteerde sociale media (hiervoor dienen ze nog steeds een aparte playback tool te gebruiken). Het gebruik van WACZ heeft echter enkele beperkingen. Om WARC-bestanden om te zetten naar WACZ kan je gebruik maken van Py-Wacz, maar dat kan enkel een Linux- en MacOS-besturingssystemen. Bovendien vereist het werken met Py-Wacz enige kennis van de command-line en het hebben van Python (laatste versie) geïnstalleerd. Daarnaast kan je WACZ-bestanden aanmaken bij het crawlen via Archiveweb.page. Dit is wel een gebruiksvriendelijke methode. Om gebruikers te ondersteunen wanneer ze zelf WARC-bestanden willen samenvoegen tot WACZ-bestand, kunnen er handleidingen worden aangeboden die specifiek zijn afgestemd op het gebruik van WACZ. Het is belangrijk dat deze handleidingen up-to-date zijn. WACZ-bestanden kunnen echter wel eenvoudig gelezen worden d.m.v. Replayweb.page. De drempel voor het lezen en gebruik door leeszaalbezoekers is dus eerder laag.

Als testcase genereerde KADOC-KU Leuven een WACZ-formaat van de sociale media platforms (Facebook, Twitter en Instagram) van de CD&V. Het is mogelijk dat het publiek zelf meer gestructureerde collecties aanmaakt (in WACZ) mits beschikbaarheid van duidelijke handleidingen en beginnerskennis van de commandline. Echter is het eerder aangewezen aan de contentpartners om zelf collecties (in WACZ-formaat) aan te maken die gebruikers kunnen gebruiken.



Replayweb.page is een browsergebaseerde viewer die webarchiefbestanden laadt die door de gebruiker worden verstrekt en deze weergeeft voor herbeleving in de browser. Er werd getest of de mogelijkheid bestaat om opgeladen WARC/WACZ bestanden extern te delen. Eveneens werd er ook gekeken naar het verschil in het gebruik van WARC- en WARC-bestanden.

Om begrijpelijke redenen kunnen URL's met de bron=file://... niet worden gedeeld, omdat ze verwijzen naar een lokaal bestand op jouw computer. Het laden van die URL door iemand anders zal resulteren in een fout. De organisatie maakt er echter wel werk van om een optie aan te bieden om lokale webarchieven peer-to-peer te delen via het DAT hyper://-protocol. DAT en de onderliggende P2P-architectuur bieden verschillende voordelen, waaronder een betere beschikbaarheid van gegevens, een grotere fouttolerantie en een verminderde afhankelijkheid van gecentraliseerde infrastructuur. Het maakt gedistribueerd delen en samenwerken mogelijk zonder te vertrouwen op traditionele client-servermodellen, wat nuttig is in verschillende toepassingen zoals het delen van bestanden, gedecentraliseerde webhosting en gezamenlijke inhoudscreatie.

Op vlak van het gebruik van WARC of WACZ bestanden zijn WACZ het handigst. Gearchiveerde websites worden veel

overzichtelijker omdat je nu een overzicht hebt met paginatitels en inhoud van de pagina's en niet langer op het niveau van de URL's moet zoeken. De verschillende pagina's zijn full-text search doorzoekbaar. Er wordt gezocht op het niveau van de paginatitel en op het niveau van de tekst van de pagina's.

Als men een WARC gebruikt moet men vaak filteren (op bv. HTML-bestanden) om bij het juiste resultaat te komen. Een andere manier om bij de output te geraken is het correct invullen van het webadres in de zoekbalk.

Gearchiveerde sociale media accounts gecreëerd met tools die scrollen doorheen het sociale media account zijn hiervoor niet geschikt. py-warz en ook Archiveweb.page herkent dit slechts als één pagina, en is zodoende niet doorzoekbaar. Het werkt enkel goed als je eerst alle URL's van de posts scrapet en vervolgens via browsertrix-crawler alle URL's apart crawlt. Laat die laatste methode wel net een minder toegankelijke manier zijn om sociale media te archiveren.

SolrWayback Deze tool biedt veel functionaliteiten, waaronder vrije tekstzoekopdrachten, export van zoekresultaten, diverse visualisaties zoals domeinkoppelingen en word clouds, beeldzoekopdrachten en visualisaties van zoekresultaten. Het vereist Java (JDK 8, 9, 10, 11), Solr 7.x en Tomcat 8.x of een andere J2EE-server voor het deployen van een Java WAR-bestand.

Voordelen van de tool zijn onder andere het bestaan van een compleet pakket met Solr, de warc-indexer en Tomcat, en een overzichtelijke documentatie met statistieken en een goede zoekfunctie. Bovendien wordt de tool regelmatig bijgewerkt en wordt het veel gebruikt binnen het IIPC-netwerk.

Er zijn echter ook nadelen, zoals het verouderde karakter van Solr 7.x en de beperkte zoekmogelijkheden voor subdomeinen en sociale media. Sommige functies werken mogelijk niet goed voor sociale media en er zijn enkele prestatieproblemen, zoals de trage solr-indexer. Ook worden er enkele fouten genoemd in het script.

Tot slot wordt vermeld dat de fulltext-zoekfunctie goed werkt, maar mogelijk alleen effectief is als de URL van elke post apart is gecrawld. Verdere onderzoeken zijn nodig om te bepalen of individuele filters mogelijk zijn via een boolean search.

Op lange termijn lijkt dit platform een veelbelovende optie voor het gewenste doel. Het biedt een scala aan functionaliteiten die waardevol kunnen zijn bij het uitvoeren van zoekopdrachten, exporteren van resultaten en visualiseren van gegevens. De installatie en monitoring van dit platform vragen de nodige expertise, tijd en opslagruimte om een efficiënte werking te garanderen. Deze oplossing loont zeker de moeite om verder te exploreren, maar dat was binnen het bestek van deze piloot niet mogelijk. Het was ook geen oplossing die direct voor alle partners binnen hun mogelijkheden lag.

Conifer (voorheen Webrecorder.io) is een gebruiksvriendelijke playback tool. Het stelt gebruikers in staat om opgeladen WARC-bestanden te delen. Weet wel dat het momenteel niet mogelijk is om WACZ-bestanden te uploaden. Gebruikers kunnen de privacy-instellingen van hun collecties aanpassen, waarbij standaard collecties privé zijn. Het beperkte gebruik van Conifer in de webarchiveringswereld zien we eerder als een beperking.

Pywb is een webarchiveringstoolkit die is ontworpen voor het afspelen van webarchieven. Het is erkend als de beste software om webarchieven af te spelen door de International Internet Preservation Coalition (IIPC) in 2020. Met pywb kunnen gebruikers webarchieven direct in de browser afspelen en creëren. Het is vooral handig voor het archiveren van dynamische websites waarvoor geen inloggegevens vereist zijn, zoals openbare accounts op Twitter.

De vereisten voor het gebruik van pywb zijn Python en een basiskennis van het werken met de command line. Het voordeel is dat pywb draait op verschillende besturingssystemen, waaronder Windows, macOS en Linux. Bovendien archiveert het sociale media in het standaardformaat WARC en maakt het gebruik van dezelfde software voor zowel het maken als het afspelen van webarchieven. De toolkit biedt uitgebreide documentatie om gebruikers te begeleiden.

Er zijn echter ook enkele nadelen verbonden aan pywb. Zowel de installatie en het gebruik van de software vereist het werken met de command line, wat een technische drempel kan vormen voor sommige gebruikers. In het algemeen biedt pywb een nuttige functionaliteit voor het afspelen en creëren van webarchieven, vooral voor dynamische websites zonder inlogvereisten. Het is belangrijk om rekening te houden met de technische vereisten en beperkingen van de toolkit bij het overwegen van het gebruik ervan voor webarchivering. Na een poging voor installatie hebben we besloten dat dit niet de beste tool voorhanden is voor ontsluiting voor een breed publiek.

5.4. Aanbevelingen

Wat is haalbaar om op korte termijn te implementeren?

Replayweb.page is de meest eenvoudige manier om gebruikers gearchiveerde sociale media te laten bekijken met behoud van look-and-feel (op basis van WARC-bestanden).

Daarnaast wordt aanbevolen om het container-formaat WACZ aan te bieden in combinatie met de playbacktool Replayweb.page. WACZ kunnen aangemaakt worden met Archiveweb.player of py-wacz (opgelet, enkel mogelijk met MacOS of Linux). WACZ zorgt voor een gestructureerde manier om je WARC (en extra) bestanden aan te bieden. De contentpartners heeft de keuze of om zelf enkele collecties te genereren of om het publiek vrij te laten om dit te doen. Ter promotie van het webarchief in je instituut lijkt het echter wel een goed idee om al één collectie in dit formaat aan te bieden (mits aan juridische voorwaarden is voldaan). Zo kunnen gebruikers deze al uittesten. Hiernaast kwam er naar boven in het nodenonderzoek dat het breed publiek baat zou hebben bij documentatie over context van i) het concept gearchiveerde sociale media ii) creatie van de collectie en iii) het verschil tussen WARC en WACZ, etc.

In het kader van het project werden enkele laagdrempelige handleidingen voor leeszaalbezoekers opgemaakt en gepubliceerd op CEST: *Handleiding voor het raadplegen van gearchiveerde websites en sociale media (HTML, WARC, WACZ)* (https://www.projectcest.be/)

Wat is haalbaar om op lange termijn te implementeren?

Op lange termijn is het wenselijk om de mogelijke implementatie van SolrWayback verder te onderzoeken voor ontsluiting naar een breed publiek toe. De SolrWayback implementatie kan zowel het brede publiek als de meer onderzoeksgerichte bekijker een betrouwbaar platform aanbieden. De levensvatbaarheid daarvan moet echter nog verder onderzocht worden wat niet mogelijk of zelfs wenselijk was binnen het bestek van deze pilootprojecten.

Voorbeeld: https://webarchivum.oszk.hu/en/webarchive/sub-collections/archive-of-the-websites-of-the-national-szechenyi-library/#webkettes