BESOCIAL

Towards a sustainable social media archiving strategy for Belgium

Contract - B2/191/P2/BESOCIAL

Samenvatting

Context

Archivering van sociale media is niet nieuw. In het afgelopen decennium is de output van sociale media, zijnde big data, een van de belangrijkste bronnen geworden voor kwantitatief (en kwalitatief) onderzoek. Het is van cruciaal belang gebleken om fascinerende inzichten te onthullen in een verscheidenheid van vragen over menselijk gedrag. Culturele instellingen zoals nationale bibliotheken verzamelen sociale-mediagegevens niet alleen voor onderzoeksdoeleinden, maar zoeken ook hun weg naar het archiveren en bewaren van dit snel veranderende datatype, via het opzetten van infrastructuur voor web- en/of sociale-media-archivering.

Van juli 2020 tot en met september 2022 ging de Koninklijke Bibliotheek van België (KBR) samen met de UGent, UNamur en UCLouvain een partnerschap aan met als doel een duurzame strategie te ontwikkelen voor het archiveren en bewaren van sociale media in België. Dit BESOCIAL-onderzoeksproject werd gefinancierd door het Federaal Wetenschapsbeleid (BELSPO) in het kader van zijn BRAIN.be-programma. KBR was de coördinator van dit project dat werd geleid in nauwe samenwerking met CRIDS (Research Centre in Information, Law and Society) aan de Universiteit van Namen, CENTAL (Centre de traitement automatique du langage), aan de UCLouvain, IDLab (Internet Technology & Data Science Lab), GhentCDH (Ghent Centre for Digital Humanities), en MICT (Research Group for Media, Innovation and Communication Technologies), aan de Universiteit Gent. De interdisciplinariteit van het onderzoeksnetwerk zorgde ervoor dat zowel technische, juridische en operationele aspecten in verband met gebruikerseisen aan bod kwamen, en bevorderde tevens kruisbestuiving binnen het project.

Doelstellingen

Het BESOCIAL onderzoeksproject werd opgedeeld in 7 werkpakketten (WP) en Taken (T) die een stapsgewijze aanpak schetsen voor de ontwikkeling van een duurzame sociale media archiveringsstrategie voor België.

De eerste doelstelling binnen het project "het onderzoeken van bestaande sociale media archiveringsprojecten in België en in het buitenland" was gekoppeld aan WP1, waar vier specifieke taken

als doel hadden om een beknopte (inter)nationale state-of-the-art van sociale media archivering (SMA) op te stellen. Twee bijkomende doelstellingen in het BESOCIAL project waren het opzetten van pilots voor sociale media archivering en het verlenen van toegang tot het sociale media archief. WP2 fungeerde als voorbereidingsfase voor het opzetten van een pilot voor social media archivering (WP3), en een pilot voor het verlenen van toegang tot het social media archief (WP4).

De eindresultaten van deze doelstellingen werden als aanbevelingen/strategie neergeschreven (WP5) bestaande uit het wettelijk kader, de technische en functionele vereisten, een business model, een definitie voor SM(A) in België, de institutionele inbedding van SM(A) in KBR, en de definitie van procedures.

Gedurende het project werden (voorlopige) resultaten nationaal en internationaal gedeeld in de vorm van het bijwonen en houden van conferenties, het geven van presentaties, en het publiceren van artikelen (WP6). Een andere cruciale doorlopende taak was het opzetten van een juridische helpdesk waar BESOCIALs juridische partner doorlopend advies gaf over de belangrijkste en relevante vragen over privacy en ICT-recht die men tegenkwam tijdens het project. Dit leidde tot een intern FAQ-document met een samenvatting van deze SMA-vragen en antwoorden.

Methodologie

Doelstelling 1: Overzicht van bestaande sociale media archiveringsprojecten in België en in het buitenland: om te onderzoeken hoe het Belgische en internationale landschap van SMA is vertegenwoordigd, kozen we voor een beschrijvende onderzoeksaanpak. Dit omvat een aantal stappen: het vinden van initiatieven die sociale media data harvesten via secundair onderzoek, het verzamelen van data van deze initiatieven via *desk-research*, enquêtes en semi gestructureerde interviews, en het analyseren en interpreteren van de data, gebruik makend van een kwalitatieve thematische analyse aanpak.

Doelstelling 2 en 3: Het opzetten van pilots voor sociale media archivering, en het verlenen van toegang tot het sociale media archief: voor de voorbereidingsfase (WP2) en uitvoeringsfase (WP3 en WP4) gebruikten we een mixed method aanpak. Een juridisch kader werd gecreëerd met behulp van desk-research. Semi-gestructureerde interviews werden gebruikt om meer inzicht te krijgen in de behoeften van de gebruikers. We experimenteerden ook met bepaalde tools tijdens de haalbaarheidsstudies om de dataverzameling op te starten. Voor de creatie van een piloot toegangsplatform werd ook een experimentele methode toegepast.

Voorstelling van de resultaten en de voornaamste besluiten en aanbevelingen

Doelstelling 1: Een overzicht van bestaande sociale-media-archiveringsprojecten in België en in het buitenland: Het in kaart brengen van het Belgische en internationale SMA-landschap is een eerste stap naar het verwerven van inzicht in en het identificeren van overlappende kenmerken van de verschillende

benaderingen van sociale media-archivering die worden toegepast. Onze bevindingen tonen aan dat veel instellingen bezig zijn met SMA, maar dat de omvang en reikwijdte ervan varieert.

Algemeen kunnen we concluderen dat de keuzes en verschillende stappen in het social-media-archiveringsproces afhankelijk zijn van bepaalde middelen. Er zijn drie struikelblokken: i) tijd om SMA te verkennen naast andere lopende taken binnen culturele instellingen, ii) (technische) *in-house* kennis en iii) beperkt budget om bijvoorbeeld over te stappen op commerciële tools en daarmee tijd te winnen. Daarnaast speelt ook het wettelijk kader een belangrijke rol in SMA-processen. Als het gaat om juridische overwegingen die een belemmering vormen voor de evolutie van de digitale samenleving, kunnen we stellen dat deze evolutie zich in zo'n snel tempo heeft voltrokken dat het voor het recht moeilijk is geweest om de laatste ontwikkelingen bij te benen.

Internationaal werden volgende gemeenschappelijke trends vastgesteld: i) Twitter is het social media platform dat het vaakst wordt gearchiveerd; ii) verzamelingen van accounts en/of hashtags focussen op belangrijke personen, organisaties en gebeurtenissen; iii) er wordt voorrang gegeven aan selectieve harvests; iv) er wordt meestal gebruik gemaakt van open source tools; v) het WARC dataformaat wordt het vaakst gebruikt, en tot slot, vi) er werd een duidelijk gebrek aan een gemeenschappelijk begrip van bewaringsconcepten (bv. 'preservation formats' of 'preservation standards') vastgesteld.

Op Belgisch niveau concludeerden we dat Facebook het vaakst wordt gearchiveerd, gevolgd door Twitter en Instagram. Deze voorkeur wordt gedreven door het thematische aspect; organisaties merken dat de accounts die ze willen vastleggen eerder op Facebook te vinden zijn. Er werd ook vastgesteld dat het aantal accounts en/of hashtags in evenwicht is met het aantal sociale-mediaplatformen. Als men veel accounts en hashtags wenst te archiveren, kiezen organisaties voor eerder voor minder platformen. Uit onze bevindingen blijkt ook dat de meeste harvesting wordt uitgevoerd met behulp van open source tools waarbij de focus ligt op het vastleggen van de look en feel van een account.

In de loop van de tijd zal het concept van sociale media en de manier waarop de samenleving deze platforms gebruikt, ongetwijfeld aanzienlijk veranderen. Of instellingen in staat zullen zijn deze veranderingen bij te benen, moet nog blijken. Op basis van deze conclusies wordt aangeraden om vervolgonderzoek te overwegen om zo het Belgische en internationale overzicht te actualiseren om de implicaties van bepaalde SMA-beslissingen voor de toekomst beter te begrijpen.

Doelstelling 2: Pilots opzetten voor sociale media archivering: Door het uitvoeren van verschillende haalbaarheidsstudies in WP2 kwamen we tot de beslissing dat de focus van het BESOCIAL project op tekst moet liggen, dit om juridische en technische implicaties te vermijden. Twitter en Instagram werden naar voren geschoven als de platformen om te archiveren met behulp van de tools Social Feed Manager en Instaloader. De selectie van het corpus werd bepaald via een duale strategie(T2.1): i) het BESOCIAL team creëerde verschillende seed lists, en ii) er werd ook rekening gehouden met de input van het publiek. Voor dit laatste werd een crowdsourcing campagne opgezet, die leidde tot vele suggesties en publiciteit voor

BESOCIAL, sociale-media-archivering in België, en KBR. Het thema voor de collecties richtte zich op cultureel erfgoed in België.

In de archiveringsfase (WP3) werden richtlijnen opgesteld om de techniciteit van de gebruikte tools op te vangen (T3.1). Ook werd meer inzicht verkregen in de noden en behoeften (T2.2) van de brede waaier aan stakeholders tijdens het gebruik van een social media archief. De volgende conclusies kwamen naar voren: i) er is een gebrek aan bewustzijn van het bestaan van (sociale media) webarchieven, ii) er is een behoefte om de academische wereld te betrekken bij selectiebeslissingen en beleid, iii) we hebben een algemene overeenkomst nodig over hoe gearchiveerde inhoud doorzoekbaar moet zijn, en iv) we moeten verwijzingen naar bepaalde methodologieën of bepaalde software of tools openstellen. Deze inzichten werden meegenomen bij de controle van de kwaliteit van de geoogste data-inhoud (T3.2). Uit deze analyse bleek dat i) voorafgaande ervaring met .csv- of .json-bestanden, of meer in het algemeen, datageletterdheid essentieel is voor het beheren, openen en kritisch analyseren van gegevens en het gegevensverzamelingsproces; ii) vooraf meer contextuele informatie moet worden verstrekt over waar de gegevensverzameling over gaat, zodat de benodigde specifieke domeinkennis kan worden vastgesteld; en iii) dat het criterium "licentie", het criterium "gebruiksvoorwaarden" en het criterium "prototypes en documentatie" verder kunnen worden verbeterd.

Al deze bevindingen werden meegenomen bij het opstellen van een ideaal preserveringsplan bij KBR (T3.3). Dit plan omvatte aanbevelingen voor de KBR, waaronder het verstrekken van betere documentatie rondom digitale preservatie.

Doelstelling 3: Pilot om toegang te verlenen tot het sociale-media-archief: In WP4 werd een analyse uitgevoerd op basis van de behoeften van gebruikers (T4.3) bij het gebruiken van een sociale media interface (gebaseerd op de bestaande interface van CENTAL). We stelden vast dat de gearchiveerde inhoud doorzoekbaar moet zijn, en dat het idee van een klassieke zoekinterface niet volstaat voor (academisch) onderzoek rond/met sociale media. De volgende criteria moeten in aanmerking worden genomen:

- Oriënteren (bv. een nieuwe gebruiker die de website en gebruikersinterface bezoekt zonder enige voorkennis heeft contextuele informatie nodig over wat hij/zij hier kan doen, over welke data beschikbaar is via de gebruikersinterface, enz)
- Controleren (bv. de gebruiker moet op weg worden geholpen bij het opstellen van zijn of haar zoekopdracht)
- Construeren (bv. beschikken over visualisaties van staafdiagrammen en cirkeldiagrammen om snel de resultaten van de zoekopdracht te kunnen begrijpen)

De technische partner heeft op basis van deze aanbevelingen, de opgelijste noden van BESOCIAL (bv. NLP) en de juridische aanbevelingen (T4.1) een mock-up van het toegangsplatform gemaakt (T4.2).

Algemeen kunnen we besluiten dat de resultaten van het BESOCIAL project een eerste belangrijke stap zijn in de richting van de implementatie van een lange termijn archiveringsstrategie voor sociale media in België.

Trefwoorden

Digital Humanities, Sociale-Media-Archivering, Born-digital Collecties, Culturele Erfgoedinstellingen, Digitale Bewaring