BESOCIAL

Towards a sustainable social media archiving strategy for Belgium

Contract - B2/191/P2/BESOCIAL

Summary

Context

Social media archiving (SMA) is neither new nor unknown. Over the past decade, the output of social media, being big data, has become one of the major sources for quantitative (and qualitative) research. It has been crucial to reveal fascinating insights into a variety of questions of human behaviour. In addition to collecting social media data solely for research purposes, cultural institutions such as national libraries are also finding their way to archiving and preserving this rapidly changing data type, by creating a framework for web and/or social media archiving.

From July 2020 until September 2022, the Royal Library of Belgium (KBR) engaged together with UGent, UNamur, and UCLouvain in a partnership with the aim of developing a sustainable strategy for archiving and preserving social media in Belgium. This BESOCIAL research project was funded by the Belgian Science Policy Office (BELSPO) as part of its BRAIN.be programme. The Royal Library of Belgium (KBR) was the coordinator of this project that was managed in close collaboration with CRIDS (Research Centre in Information, Law and Society) at the University of Namur, CENTAL (Centre de traitement automatique du langage), at the UCLouvain, IDLab (Internet Technology & Data Science Lab), GhentCDH (Ghent Centre for Digital Humanities), and MICT (Research Group for Media, Innovation and Communication Technologies), at the Ghent University. The interdisciplinarity of the research network ensured that technical, legal and operational aspects were covered as well as aspects related to user requirements and fostered cross-fertilisation within the project.

Objectives

The BESOCIAL research project was divided into a number of work packages (WP) and Tasks (T) outlining a step-by-step approach for the development of a sustainable social media archiving strategy for Belgium.

The first objective within the project “reviewing existing social media archiving projects in Belgium and abroad” was linked to WP1, where four dedicated tasks aimed to provide a concise (inter)national state-of-the art of social media archiving (SMA). Two additional objectives in the BESOCIAL project were to set up pilots for social media archiving and to provide access to the social media archive. WP2 functioned as the preparation phase for setting up a pilot for social media archiving in WP3, and a pilot for providing access to the social media archive in WP4.

The final results of these objectives were written down in a 6-part recommendation section (WP5) with the legal framework, the technical and functional requirements, a business model, a definition for SMA in Belgium, institutional embedding of SM(A) in KBR, and the definition of procedures.

Throughout the project, (preliminary) results and processes were shared nationally and internationally in the form of attending and holding conferences, giving presentations, and publishing articles (WP6). Another crucial ongoing task was having a legal helpdesk where BESOCIAL's legal partner offered ongoing consultancy regarding the main and relevant questions about privacy and ICT law encountered during the project. This led to an internal FAQ document summarising these SMA questions and answers.

Methodology

Objective 1: Review of existing social media archiving projects in Belgium and abroad: to investigate how the Belgian and international landscape of SMA is represented, we took a descriptive research approach. This entailed a number of steps: finding initiatives that harvest social media data via secondary research, collecting data of these initiatives via desk-research, surveys and semi-structured interviews, and analysing and interpreting the data, using a qualitative thematic analysis approach.

Objective 2 and 3: Set up pilots for social media archiving, and to provide access to the social media archive: for the preparation phase (WP2) and execution phase (WP3 and WP4) we used a mixed method approach. A legal framework was created using desk-research. Semi-structured interviews were used to get more insight into the requirements of the users. We also experimented with certain tools during the feasibility studies in order to start the harvesting of the data. For the creation of a pilot access platform an experimental method was also implemented.

Conclusions

Objective 1: Review of existing social media archiving projects in Belgium and abroad: Mapping the Belgian and international SMA landscape is a first step towards gaining insights in and identifying overlapping characteristics of the different approaches of social media archiving applied. Our findings showed that many institutions are engaged in SMA, but that the stage and efforts vary both in size and scope.

Overall, we can conclude that the choices and different steps taken in the social media archiving process are resource-dependent. There are three stumbling blocks: i) time to explore SMA next to the ongoing tasks within cultural institutions, ii) (technical) in-house knowledge and iii) limited budget to, for example, switch to commercial tools and thereby gaining time. In addition, the legal framework also plays an important role in SMA processes. When it comes to legal considerations that constitute obstacles to the evolution of the digital society, we can say that this evolution has happened at such a fast pace that it has been difficult for the law to keep up with the latest developments.

Internationally the following common trends were identified: i) Twitter is the social media platform most often archived; ii) collections of accounts and/or hashtags focus on important people, organisations, and events; iii) priority is given to selective harvests; iv) open source tools are mostly used; v) the WARC data format is most often used, and finally, vi) a clear lack of a common understanding of preservation concepts (e.g. 'preservation formats' or 'preservation standards') was observed.

On a Belgian level we concluded that Facebook is most often archived, followed by Twitter and Instagram. This preference is driven by the thematic aspect; organisations find that the accounts they want to capture are more likely to be found on Facebook. We also notified that the number of accounts and/or hashtags is well balanced between the number of social media platforms. In case that one wishes to archive a lot of accounts and hashtags, organisations opt for fewer platforms. Our findings showed that most of the harvesting is executed by means of open source tools with a focus on capturing the look and feel of an account.

Over time, the concept of social media and the way society uses these platforms will without doubt considerably change. Whether institutions will be able to monitor and follow-up on these changes has to be seen. Based on these conclusions, practitioners should consider conducting follow-up research by updating the Belgian and international overview to better understand the implications of certain SMA decisions for the future.

Objective 2: Set up pilots for social media archiving: By conducting several feasibility studies in WP2, we found that the focus of the BESOCIAL project should be on text, mainly to avoid legal and technical implications. Twitter and Instagram were put forward as the platforms to harvest with the help of the tools Social Feed Manager and Instaloader. The selection of the corpus was determined through a dual policy (T2.1): i) the BESOCIAL team created several seed lists, and ii) the input of the public was also taken into account. For the latter, a crowdsourcing campaign was set up, which led to many suggestions as well as publicity for BESOCIAL, social media archiving in Belgium, and KBR in general. The theme for the collections focused on Cultural Heritage in Belgium.

In the harvesting phase (WP3) guidelines were created to overcome the technicality of the tools used (T3.1). Also more insight was gained into the needs and requirements (T2.2) of a wide range of stakeholders when using a social media archive. The following conclusions emerged: i) there is a lack of awareness of the existence of (social media) web archives, ii) there is a need to include the academic field in selection decisions and policies, iii) we need an agreement on how archived content should be searchable, and iv) we need to open up references to particular methodologies or particular software or tools. These insights were taken into account when controlling the quality of the harvested data content (T3.2). Our analysis shows that i) prior experience with .csv or .json-files, or more generally, data literacy is key and vital for managing, accessing and critically analysing data and the data-collection process. ii) more contextual information should be provided in advance on what the data set is about in order to identify the specific domain knowledge needed, and iii) that the criterion 'license', the criterion 'terms of use' and the criterion 'prototypes and documentation' can be improved further.

All these insights were taken into account when creating a preservation plan and workflow in an ideal scenario at KBR (T3.3). This included recommendations for KBR, such as providing better documentation in order to archive and preserve our corpus in a sustainable way for the long term.

Objective 3: Pilot to provide access to the social media archive: In WP4 an analysis was conducted based on the needs of users (T4.3) when searching on a social media interface (based on the CENTAL's existing interface). We identified that the archived content should be query-able, and that the idea of classic search interface would not suffice for social media research. The following criteria must be taken into account:

- Orientating (e.g. a new user that visits the website and user interface without any prior knowledge needs contextual information on what he/she can do here, on what data is available through the user interface etc)
- Auditing (e.g. the user should be assisted in creating his or her search query)
- Constructing (e.g. having bar chart and pie chart visualisations to quickly grasp the search query results as a whole)

An access platform mock-up was created (T4.2) by the technical partner based on these recommendations, the BESOCIAL needs (e.g. NLP), and the legal recommendations (T4.1).

Overall we can conclude that the results of the BESOCIAL project are a first major step towards implementing a long-term social media archiving strategy for Belgium.

Keywords

Digital Humanities, Social Media Archiving, Born-digital Collections, Cultural Heritage Institutions, Digital Preservation