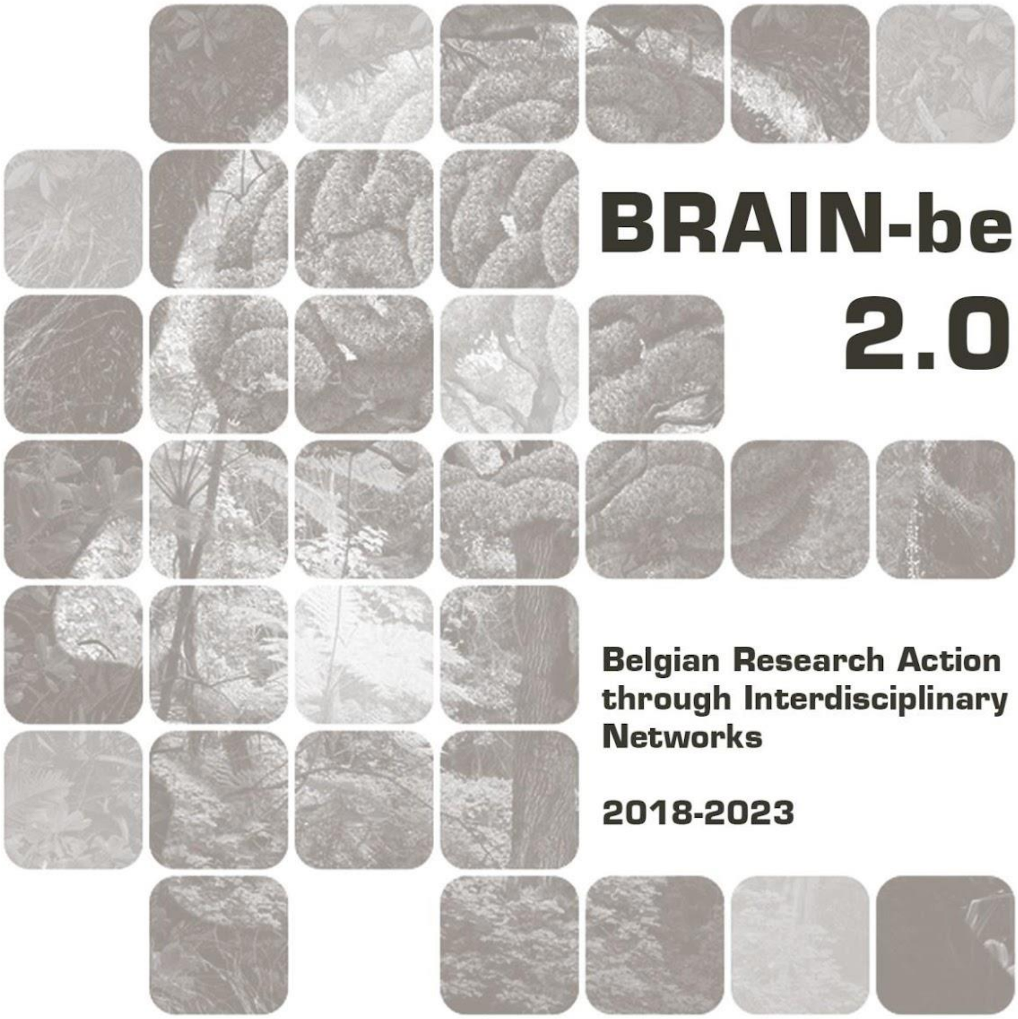


BRAIN-be 2.0

Belgian Research Action through Interdisciplinary Networks

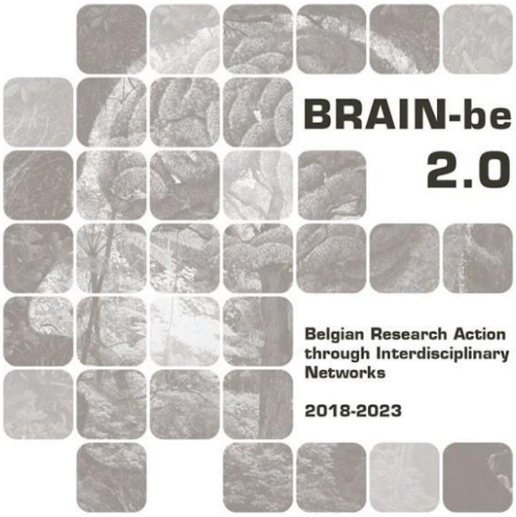
2018-2023

BESOCIAL

Towards a sustainable social media archiving strategy for Belgium

BIRKHOLZ, J. M. (KBR and UGhent), BUDULAN, D. (UCLouvain), CHAMBERS, S. (KBR and UGhent), DENIS, L.-A. (UNamur), GEERAERT, F. (KBR), HEYVAERT, P. (UGhent), MECHANT, P. (UGhent), MESSENS, F. (KBR and UGhent), MICHEL, A. (UNamur), ROLIN, E. (UCLouvain), VANDEPONTSEELE, S. (KBR), VLASSENROOT, E. (UGhent) and WATRIN, P. (UCLouvain)

Pillar 2: Heritage science



NETWORK PROJECT

BESOCIAL

Towards a sustainable social media archiving strategy for Belgium

Contract - B2/191/P2/BESOCIAL

FINAL REPORT

PROMOTORS:

Vandepontseele, Sophie (KBR)
de Terwangne, Cécile & Michaux, Benoît (UNamur)
Watrin, Patrick (UCLouvain)
Mechant, Peter (UGhent)

AUTHORS:

Birkholz, Julie M. (KBR and UGhent)
Budulan, D. (UCLouvain)
Chambers, Sally (KBR and UGhent)
Denis, Lise-Anne (UNamur)
Geeraert, Friedel. (KBR)
Heyvaert, Pieter (UGhent)
Mechant, Peter (UGhent)
Messens, Fien (KBR and UGhent)
Michel, Alejandra (UNamur)
Rolin, Eva (UCLouvain)
Vandepontseele, Sophie (KBR)
Vlassenroot, Eveline (UGhent)
Watrin, Patrick (UCLouvain)

Published in 2022 by the Belgian Science Policy Office
WTCIII
Simon Bolivarlaan 30 bus 7
Boulevard Simon Bolivar 30 bte 7
B-1000 Brussels
Belgium
Tel: +32 (0)2 238 34 11
http://www.belspo.be
http://www.belspo.be/brain-be

Contact person: Georges Jamart
Tel: +32 (0)2 238 36 90

Neither the Belgian Science Policy Office nor any person acting on behalf of the Belgian Science Policy Office is responsible for the use which might be made of the following information. The authors are responsible for the content.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without indicating the reference:

Birkholz, J. M, Budulan, D., Chambers, S., Denis, L.-A., Geeraert, F., Heyvaert, P., Mechant, P., Messens, F., Rolin, E., Vandepontseele, S., Vlassenroot, E. , Michel, A. and Watrin, P. ***BESOCIAL – Towards a sustainable social media archiving strategy for Belgium*** Final Report. Brussels: Belgian Science Policy Office 2022 – 42 p. (BRAIN-be 2.0 - (Belgian Research Action through Interdisciplinary Networks))

TABLE OF CONTENTS

ABSTRACT

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 objectives of the project were to: 1) review existing social media archiving projects in Belgium and abroad, 2) set up a pilot project for social media archiving and 3) set up a pilot project to provide access to the social media archive. The main deliverables comprise a comprehensive report on best practices within the field of social media archiving, in-depth legal analyses of the Belgian legal framework surrounding social media archiving and providing access to the collections, an analysis of user requirements, and recommendations for a sustainable social media archiving service in Belgium.

1. INTRODUCTION

The aim of the BESOCIAL research project was to develop a sustainable social media archiving strategy for Belgium. Corpora of archived social media content holds huge research potential for social and political scientists, historians, experts in communication studies, linguists etc. However, social media content is very ephemeral and is often not archived thereby creating serious challenges for (digital) scholars who want to use archived social media content as a data resource. A significant number of heritage institutions abroad have already set up programmes to archive social media, but social media archiving is still an emerging topic for which the standards for capturing and archiving are yet to be consolidated.

The first objective within the project is to undertake a **thorough review of existing social media archiving projects in Belgium and abroad**. Two additional objectives are to **set up pilots for social media archiving and for providing access to the social media archive**.

The obtained research results fed into recommendations for sustainable social media archiving in Belgium - the overall objective of the project. The network is composed of one federal scientific institution and three universities from the different language communities: KBR (Royal Library of Belgium), CRIDS (Research Centre in Information, Law and Society) at Namur University, CENTAL (Centre de traitement automatique du langage) at Université Catholique de Louvain, ID Lab (Internet Technology & Data Science Lab), Ghent CDH (Ghent Centre for Digital Humanities) and imec-mict (Research Group for Media and ICT) at 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. The results of the project are a first major step towards implementing a long-term social media archiving strategy for Belgium.

2. STATE OF THE ART AND OBJECTIVES

The first objective within the project was to undertake a **thorough review of existing social media archiving projects in Belgium and abroad.** WP1 tackled this objective, where four dedicated tasks aimed to provide a concise international state-of-the art of social media archiving (SMA). A detailed report of this internal overview can be found on: https://orfeo.kbr.be/handle/internal/7741 . Research results from WP1 were also valorized as an article in a special issue of the peer-reviewed journal 'International Journal of Digital Humanities', see https://link.springer.com/article/10.1007/s42803-021-00036-1. WP7 can also be linked to this objective, where BESOCIAL's legal partner CRIDS offered ongoing consultancy (in the form of a help desk) regarding the main and relevant questions about privacy and ICT law encountered during the project. This help desk led to an internal FAQ document summarising these SMA questions and answers.

Two additional objectives in the BESOCIAL project were to **set up pilots for social media archiving** and for **providing 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 findings of these objectives were written down in a 6-part strategy (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 in WP6 were shared nationally and internationally in the form of attending and holding conferences, giving presentations, and publishing articles.

State of the art: Social media

Coined in the nineties, it took until the mid-2000s for the phrase 'social media' to enter common parlance (Ortner, Sinner & Jadin, 2018). Although the exact meaning of the phrase is subject to ongoing discussions due to the variety of evolving stand-alone and built-in social media services, generally 'social media' refers to Information and Communication Technologies (ICT) that enable social interaction (Treem & Leonardi, 2012) that allows "the creation and exchange of user-generated content" (Kaplan & Haenlein, 2010, p. 61). Social media thus encompasses interactive computer-mediated technologies that facilitate the creation or sharing of information and other forms of expression via online communities and networks (Kietzmann, Hermkens, McCarthy and Silvestre, 2011; Obar & Wildman, 2015). However, the term "social" does not account for technological features of a platform alone, its level of 'sociability' is clearly determined by the actual performances and interactions of the social platform's users (Ariel & Avidar, 2015). Social media platforms such as Facebook, Twitter or YouTube are representative of the growing "networked information economy" (Benkler, 2006), marking a shift from an industrial information economy (content centrally produced and distributed by commercial entities) to an economy in which individuals and groups of citizens create, annotate, and distribute media de-centrally (Marwick, 2010). Social media platforms embody a key aspect of today's Internet, namely the rise of online user participation and interaction. Websites have evolved from a collection of online static pages to continually-updated platforms that invite users not only to consume (read, listen, watch), facilitate (tag, recommend, filter) and communicate (send messages, post comments, rate, chat) but also to create (personalise, aggregate, contribute) and share (publish, upload) content.

State of the art: born-digital heritage

Consequently, records of our social history can also be documented from online sources, in addition to traditional materials. Digital heritage and the importance of its active preservation was formally recognised with the adoption of the UNESCO Charter on the Preservation of the Digital Heritage (UNESCO, 2003). The charter recognises born digital resources, as those resources existing in “no other format but the digital original”, and as “part of the world’s cultural heritage” and therefore “constitute a heritage that should be protected and preserved for current and future generations”. Even though the charter was adopted prior to the large-scale advent of social media, UNESCO’s Concept of Digital Heritage (UNESCO, s.d.) recognises that “this digital heritage is likely to become more important and more widespread over time. Increasingly, individuals, organisations and communities are using digital technologies to document and express what they value and what they want to pass on to future generations. New forms of expression and communication have emerged that did not exist previously”.

State of the art: History of social media archiving

As the web evolved, web archiving evolved with it and the creation of social media platforms gave rise to SMA initiatives. One of the pioneer projects in SMA is the Occasio project, launched in 1995 that aimed to preserve political and social conversations posted between 1988 and 2002 on online discussion groups (IISH, 2020). During this period (national) libraries and archives also broadened the scope of their collections to include the web. At the National Library of New Zealand, the first Twitter archive was added to the collections in 2009 (Macnaught, 2018). The British Library started archiving social media systematically in 2010 but limited Twitter, Facebook and Youtube content was captured previous to this date, where the UK National Archives has archives of Twitter accounts dating back to 2008 in its collections (Hockx-Yu, 2014; Espley, Carpentier, Pop & Medjkoune, 2014). In 2010, a partnership between Twitter and the Library of Congress was initiated, in order to archive public tweets published on the platform (Zellier, 2018). Since 2017 this initiative has reduced in its capacity. The change to selective collecting was prompted by the changing nature of Twitter (increased length of tweets or increasing video, images or linked content for example) and constituted an alignment with the collection policies of the Library of Congress (Library of Congress, 2017). The Bibliothèque nationale de France has archived Facebook data since the creation of its web archive in 2006, but technological changes within Facebook forced the library to stop systematically archiving it in 2010 (Le Follic & Chouleur 2018). These last two examples clearly illustrate that collection development plans are directly influenced by (technological) changes in the social media landscape.

State of the art: Importance of preservation for the right to information and legal constraints

Social media archives and allow us to document the past in ways we have never previously had the ability, as well as ease to archive. They are an invaluable resource for researchers to study human behavior and history as they provide clear records of communication (Ruth & Pfeffer, 2014). The logs, social media posts and related metadata allow us to document the past in ways we have never previously had the ability, as well as ease to archive. Social media platforms and the web in general provide an essential tool for the freedom of expression and the right to information for citizens of all ages and backgrounds. The European Court of Human Rights frequently supports this observation in its case law.1 In such a context, the preservation of social media content and its availability for the research community and the general public are major societal challenges. Indeed, these SMA initiatives, more specifically the log files, content of social media posts and related metadata, allow people to search and access a multitude of content that can be considered as born-digital heritage and of cultural, societal, historical or scientific interest. In doing so, archiving institutions play the role of "facilitator" in the exercise of the fundamental rights conferred by Article 10 of the European Convention on Human Rights and, more particularly, the right to information. This fundamental right protects both the communication of ideas, opinions and information and their reception. Furthermore, the European Court of Human Rights had the opportunity, in 2012, to consider that the constitution of archives on the Internet fell under the umbrella of Article 10 of the Convention. It was in a Times Newspapers Limited v. the United Kingdom judgment concerning the establishment of a web archive of press articles that the Court for the first time specified that "[...] Article 10 guarantees not only the right to impart information but also the right of the public to receive it. In the light of its accessibility and its capacity to store and communicate vast amounts of information, the Internet plays an important role in enhancing the public's access to news and facilitating the dissemination of information in general. The maintenance of Internet archives is a critical aspect of this role and the Court therefore considers that such archives fall within the ambit of the protection afforded by Article 10". In particular, the Court added that providing citizens with Internet archives afford a substantial contribution for the preservation and the making available of news and information and constitutes also a valuable source for education and historical research.

However, even if SMA initiatives have a particular resonance in terms of fundamental rights' protection, they still involve competing interests that should be considered. Alongside the interest of scientists, researchers and society at large in accessing archived contents, there are the interests of other stakeholders such as copyright holders, people involved in producing, the owners of websites or social media pages or (national) cultural heritage institutions. Implementing SMA initiatives obviously involves the same legal considerations as the web; however, they go a step further by raising additional legal issues compared to those of web archiving. Here, we can think of the ambiguous relationship between social media and the right to privacy protected by Article 8 of the European Convention on Human Rights. In that respect, when it comes to archiving social media, we must be attentive to the question of whether the content posted on social media belongs to the private or public sphere. This question, which is at the heart of many controversies in jurisprudence, is crucial to assess a possible violation of the privacy of persons targeted by publications on social media. In addition, the right to privacy is a greater concern for social media than web pages; specifically aspects related to image right or e-reputation are much more sensitive on social media than on web pages.

State of the art: Metadata standards for effective data management

Archiving and mastering the volume, variety and velocity of data on social media platforms demands high-quality metadata to, among others, allow effective (research) data management. The National Information Standards Organization (NISO) defines several types of metadata: descriptive metadata to find and understand resources, administrative metadata which can be of technical, preservation or digital rights nature, structural metadata to describe relationships between resources and markup languages which integrates content with metadata to express other structural or semantic features (Riley, 2017). There is a strong need for provenance metadata on different levels for archived web content (Venlet et al., 2018) for both basic users and scholars (Vlassenroot et al., 2019; Littman et al., 2018); this is often in contrast to the needs of practitioners (Venlet et al., 2018). In case of social media this metadata can be provided via Application Programming Interfaces (APIs). Several metadata standards exist from which a common subset can be distilled, however, most tools which create metadata define descriptive metadata differently and mostly collect technical metadata. NISO lists 11 metadata standards in the cultural heritage field ranging from the storage efficient machine readable MARC format family developed in 1968 to several XML-Schemas and OWL ontologies like DDI and PREMIS developed in recent years. Whereas these standards cover different types of metadata, Dooley et al. (2018) reviewed existing metadata standards with respect to descriptive metadata and recommended the use of 14 data elements. These elements are applicable both on collection and on item level. Although these 14 elements largely overlap with Dublin Core, they are meant to be standard-neutral. Although no minimum set is required, Title and URL are the absolute minimum and Collector, Creator, Date and Description are strongly recommended. In practice, descriptive metadata is defined differently by different platforms and tools, but that most tools provide technical metadata as WARC is an often used file format to store captured web content (Samouelian et al., 2018). Several commercial tools for social media harvesting exist, but also various open source solutions have been developed to monitor, capture and store social media content. For lists of social media research tools, including data collection and archiving tools, curated by researchers see: the 'Social Media Research Toolkit', and the wiki 'Social media data collection tools'. A list of general web harvesting tools were collected by the Data Together initiative in 2018 in the form of a collaborative spreadsheet (Hucka, 2017).

3. METHODOLOGY

The BESOCIAL project is 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. Under section 4 an overview is included, of the results and the methodological steps taken.

4. SCIENTIFIC RESULTS AND RECOMMENDATIONS

Work package 1: Review of existing social media archiving projects & corpora in Belgium & abroad (M1-M6)

In WP1, four dedicated tasks aimed to provide a concise international state-of-the art of social media archiving (SMA). The aim of Task 1.1. was to create an overview of international best-practices for preserving and archiving social media. Task 1.2. had the purpose to analyse potential foreign legal frameworks allowing SMA, inter alia, by national libraries. The analysis focused on selected European countries but also on non-EU countries and regions that may be of interest (e.g. New-Zealand, Canada or Quebec). Task 1.3. surveyed existing tools, standards and techniques relevant to discover, collect and consolidate data from different social media networks for policy-aware long-term preservation. Finally, Task 1.4. aimed to obtain an overview of the preservation policies in place for social media content by studying the policies of institutions that are archiving social media in detail. For the first work package our methodology consisted of three different phases, which fed into one final report that integrated the research results of the four tasks described above. The main outputs of each task are detailed here below. A more detailed report of Work Package 1 can be found in the form of a report that is made available on the Orfeo Platform and communicated to the follow up committee: https://orfeo.kbr.be/handle/internal/7741.

Task 1.1. Analysis of selection and access policies (Lead: KBR; UGhent-MICT&GhentCDH, CRIDS) (M1-M6)

The purpose of this task was to provide an overview of current practices by archiving institutions for selection and access. A secondary research approach (also known as desk research) was implemented which involved summarization, collation and synthesis of existing projects relating to SMA. In addition to the desk research related to SMA projects, a second desk research study was also carried out by computer engineers reviewing data collection documentation and information in technical GitHub repositories, see Task 1.3 below. With regards to the selection of our sample of web archiving initiatives, a number of characteristics were taken into account:

- Web archiving initiatives that were included in the BRAIN-be PROMISE-project (2017-2019), the web archiving initiative of the Royal Library of Belgium and the State Archives of Belgium;
- Established web archiving initiatives;
- Convenience sampling (also known as grab sampling, accidental sampling, or opportunity sampling), a type of non-probability sampling that involves the sample being drawn from that part of the population that is close to hand. This type of sampling is most useful for pilot testing or exploratory research;
- Initiatives that are archiving or do not yet archive social media.

A number of archiving initiatives were selected and analysed in depth, see Table 1 for the details.

Country	**Institution**	**Name**	**Abbreviation**
Canada	National Library	Library and Archives Canada	LAC
Canada	Regional Library	Bibliothèque et Archives nationales du Québec	BAnQ
Denmark	Royal Danish Library	Netarkivet	Netarkivet
Estonia	National Library	Eesti Veebiarhiiv	Eesti Veebiarhiiv
France	National Library	Bibliothèque nationale de France	BnF
France	National Audiovisual Institute	Institut national de l’audiovisuel	INA
Hungary	National Library	National Széchényi Library	NSL
Ireland	National Library	National Library of Ireland	NLI

Luxembourg	National Library	Bibliothèque nationale du Luxembourg	BnL
New-Zealand	National Library	National Library of New Zealand	NLNZ
Switzerland	National Library	Webarchiv Schweiz	Webarchiv Schweiz
The Netherlands	National Library	KB Webarchief	KB
The Netherlands	National Archive	Nationaal Archief	NA
UK	British Library	UK Web Archive	UKWA
USA	University Library	George Washington University Libraries	GWUL

Table 1. List of Social Media Archives, as recorded from desk research

In the second research phase, a questionnaire which ran from July 2020 to September 2020 was sent to representatives from the aforementioned institutions. The aim of this survey was to address issues or questions that could not be resolved fully by the desk research in phase 1. The third and final research phase encompassed further validation and synthesis by means of in-depth interviews. The information about the interviews are found here in detail below in Table 2.

Country	**Institution**	**Date**	**Interviewee**	**Interviewers**
Canada	Library and Archives (LAC)	18/11/2020	Tom Smyth	Sally Chambers, Sven Lieber, Jessica Pranger & Eveline Vlassenroot
Denmark	Royal Danish Library (Netarkivet)	24/11/2020	Anders Klindt Myrvoll & Tue Hejskov Larsen	Sally Chambers, Sven Lieber & Eveline Vlassenroot
France	Bibliothèque nationale de France (BnF)	27/11/2020	Alexandre Chautemps & Sara Aubry	Sally Chambers, Friedel Geeraert, Jessica Pranger & Eveline Vlassenroot
France	National Audiovisual Institute (INA)	18/11/2020	Thomas Drugean, Claude Mussou & Jérôme Thiève	Sven Lieber
Luxembourg	Bibliothèque nationale du Luxembourg (BnL)	18/11/2020	Ben Els & Yves Maurer	Sally Chambers, Sven Lieber,

				Jessica Pranger & Eveline Vlassenroot
New Zealand	National Library	24/11/2020	Gilian Lee, Ben O’Brien, Valerie Love & Ronda Grantham	Sally Chambers, Friedel Geeraert, Sven Lieber & Eveline Vlassenroot
Portugal	Arquivo.pt	10/12/2020	Daniel Gomes	Sally Chambers, Sven Lieber, Jessica Pranger & Eveline Vlassenroot
United Kingdom	British Library	30/11/2020	Nicola Bingham	Friedel Geeraert, Sven Lieber, Eveline Vlassenroot & Sally Chambers
United Kingdom	National Archives	4/02/2021	Tom Storrar, Claire Newing & Sarah Dietz	Sally Chambers, Sven Lieber, Fien Messens & Eveline Vlassenroot

Table 2. Details of the interviews

Our findings show that many institutions are engaged in SMA, yet the stage and efforts vary in size and scope. Archiving social media happens through selective crawls that most often focus on specific events, manifestations or even emergencies and to a lesser extent through crawls on specific themes. To mitigate the fact that it is very difficult or nearly impossible to anticipate or plan for certain major events (e.g. covid19), some institutions in our sample (e.g. the National Library of France or the National Library of Canada) shifted their strategy to a continuous automated collection process of news and social media content supplemented with the archiving of curated content. Given that it is not feasible to archive the entire social web, selections must be made. These selections are often based on a specific topic; a hashtag (#) or keywords, a limited time period, or a crawl on one specific platform. Twitter is the social media platform most often archived by the institutions in our sample, followed by Facebook and Instagram.

These selections are a challenge for SMA initiatives as despite efforts to be transparent about their crawling activities and including known limitations, there is also the limitation of the tools. Our findings show that much of the crawling is done through application programming interfaces (APIs). APIs limit the information that can be collected (i.e. maximum request per day) and limit its reuse, which further limits access and knowledge to access questions of quality. There is a concern that this may result in implicit unrepresentative sampling which influences the external validity of data samples. External validity refers to when the conclusions drawn from the sample are applicable for an entire population; thus, giving a user or researcher an indication of the generalisability of the research to a specific population. Without explicit knowledge of these selection criteria, researchers are limited in drawing conclusions and testing theories.

Moreover, another challenge is how to provide and ensure access to archives and under which copyright conditions. In particular, for researchers using these broad and large archives there remain questions around the quality of these data; largely this is about the (in)completeness of the material and related metadata. Despite these known challenges that are inherent to current SMA processes, social media archives remain an under-utilised resource for humanities and social science researchers in particular.

Task 1.2. Analysis of existing foreign legal frameworks (Lead: CRIDS) (M1-M6)

The purpose of this task was to analyse potential foreign legal frameworks allowing SMA, inter alia, by national libraries. An extensive analysis of SMA on an international level can be found in the WP 1 Report, referenced above. This analysis also fed the reflection and the recommendations work document concerning SMA and the potential revision of Belgian legal deposit law (see task 5.1).

The results of this task provides an update of the legal aspects of the national initiatives studied during the PROMISE research project, with a specific emphasis on provisions related to social media archiving. In a second part, analysis of new (e.g. projects that were not considered in the web archiving project of PROMISE 2017-2019) national initiatives has been analysed.

During the conducted interviews and questionnaires we found out that there is the challenge of how to provide and ensure access to archives and under which copyright conditions this can be done (Zimmer, 2015; George Washington University Libraries, 2016; McCreadie, Soboroff, Lin, Macdonald, Ounis & McCullough, 2012). The ways in which institutions provide data level access to their social media collections varies from scope and size. The national legal frameworks for accessing social media information varies as well.

Task 1.3. Analysis of technical solutions for social media archiving and preliminary testing of tools and quality control (Lead: UGhent-IDLab; CENTAL) (M1-M6)

In the elaboration of this task, more insight was gained in the field of existing tools, standards and techniques relevant to discover, collect and consolidate data from different social media networks for policy-aware long-term preservation. An extensive analysis of SMA tools on an international level can be found in the WP 1 report, and a summary table online available under an open CC BY 4.0 license.

From the discussed questionnaire answers and from our own tool testing activity we conclude for the following points for our use case of a sustainable social media archiving in Belgium:

A combination of large scale API harvesting and subsequent, more selective and time-intensive, look and feel harvests seem to be a solution with an appropriate trade-off. Standard web archiving tools have shown to be not always reliable with social media content, similarly some social media archiving tools are also error prone, due to changes from social media providers, or slow as harvesting is performed live while human users browse. Most tools can quickly be set up, but the question is if their output sufficiently addresses the use case, e.g. which data needs to be archived? How much data should be harvested and is the preservation of the look and feel important? Who are the users? What resources are available for harvests and curation? Answers to such questions already limit possible choices. Harvesting social media while keeping a provided look and feel seems to be the simplest choice which is also accessible by regular users. However, it is still error prone or slow and there is not necessarily a single original look and feel as social media content is visualized slightly differently for different clients. Neglecting any provided look and feel and focusing on data only, reduces storage costs and provides more metadata while still archiving the actual content probably serving a high percentage of possible use cases. However, lots of implicit information valuable for future use or research will be lost, i.e. use of colors, placement of content such as comments or ads etc. Harvesting social media data from APIs seems to be the more reliable first choice as relevant data can be harvested, annotated and further processed on a large scale. Customized visualizations may provide limited look and feel. Such initially harvested and possibly further processed data can inform subsequent more selective harvests with different tools by manual curators also taking the look and feel into account to accompany harvested data.

Collected API metadata together with the metadata of tools like SFM (Social Feed Manager) have the potential to automate data stewardship tasks and improve the work of archivists and the experience of users. Data stewardship for social media archives entail diverse tasks such as the description of archived social media content according to different metadata standards or the removal of content from the search index due to take down policies. These are often manual tasks, but FAIR archives can support archivists in such tasks with high quality metadata. Similarly, users are supported in exploring and accessing the archive's content, because applications using the FAIR archives can be built. The key to FAIR social media archives are provenance information of the harvesting process, archived files and their format and of course the content itself. Tools like SFM which focus on providing standardized harvesting workflows and metadata on top of different harvesters for different social media providers APIs are appropriate choices to reach the goal of FAIR archives.

Task 1.4. Analysis of preservation policies (Lead: KBR; GhentCDH) (M1-M6)

The aim of this task was to obtain an overview of the preservation policies in place for social media content by studying the policies of institutions that are archiving social media in detail.

Preservation practices proved to be diverse among those surveyed. A long-term view on the usability of archived content implies a number of challenges, which KBR will have to consider, some of which were highlighted by participants in the BESOCIAL survey. Format migration is still a nascent field, even though the native file formats will at some point in the future have to be migrated to preservation formats. It would also seem that there were several interpretations of what is understood as 'preservation procedures'; with harvesting and archiving being closely connected processes and sometimes used interchangeably. The WARC standard was named as a preservation standard even though the format was developed as a storage format, rather than specifically for long-term preservation. This may demonstrate a lack of common understanding of what is considered as digital preservation procedures or formats, and therefore a need for raising awareness about preservation of social media content.

Work Package 2: Preparation of pilot for social media archiving (M4-M15)

A more detailed report of Work Package 2 and 3 can be found in the form of a report that is made available on the Orfeo Platform and communicated to the follow up committee: https://orfeo.belnet.be/handle/internal/10034.

Task 2.1. Development of methodology for selection of content for real-time archiving of social media (Lead: KBR; CENTAL, UGhent-All) (M4-M9)

A selection policy has been drawn up within the KBR to guarantee the representativeness of the data. During the creation of this policy, careful attention was paid to ensure that it complies with the values and standards of the KBR and its legal deposit.

Linked to this task a mini-pilot was also currently running as a form of feasibility study surrounding the theme COVID-19 in Belgium on Twitter. The mini pilot results in a test crawl of two corpora from two different tools (a tool developed by the partner CENTAL's and the open source Social Feed Manager), harvesting a list of hashtags and accounts that were related to the theme COVID. The harvesting organised by CENTAL and IDLAB started from 25 January 2021. An initial analysis was performed of the content collected from 25 January to 16 February 2021. Table 3 lists the categories of the prepared seedlist as well as the number of accounts per category and the number of harvested tweets per category.

Seed list categories	**Number of accounts in category**	**Number of harvested tweets**
Accounts directly related to coronavirus	4	13,621
Accounts of artists/artistics events	4	4,301
Accounts of charitable institutions	2	861
Accounts of commercial products	1	1050
Accounts of governmental institutions	4	10,464
Accounts of health/medical institutions/communities	5	9,424
Accounts of hospitals/hospital departments	3	1,563
Accounts of journalists	1	3,663
Accounts of media channels (newspapers, magazines, online news, etc.)	18	56,012

Accounts of politicians	13	22,984
Accounts of scientists and scientific institutions	11	12,230
Accounts of social/political institutions/movements	8	20,361
Accounts related to transport	4	15,159

Table 3: Number of Tweets collected per seed list category, numbers from the harvesting period 2021-01-25 to 2021-02-16

In an update meeting, discussing the advantages, disadvantages and results from the two tools, the team selected to implement Social Feed Manager as the crawling tool for BESOCIAL. Social Feed Manager preserves harvesting provenance by wrapping API requests in WARC files, thus for the account collection we have a daily WARC file combined 64 MB and for the hashtag collection combined around 6 MB.

At the time of publication of this report, we have collected more than 160k tweets for the account collection from 78 Twitter handles, and more than 10k tweets for the hashtag collection from 28 COVID-19 related hashtags (see seedlist). Extracted for the analysis we end up with a 729 MB JSON file for the account collection and a 58 MB JSON for the hashtag collection.

Figure 1, below, shows the temporal distribution of different types of tweets from this account collection. A few Tweets date back to 2011 because for a so-called timeline harvest, where tweets from a specific account are obtained, Twitter returns the last 3,200 tweets and some accounts in this collection are not that active resulting in the fact that we captured their Tweets almost entirely.

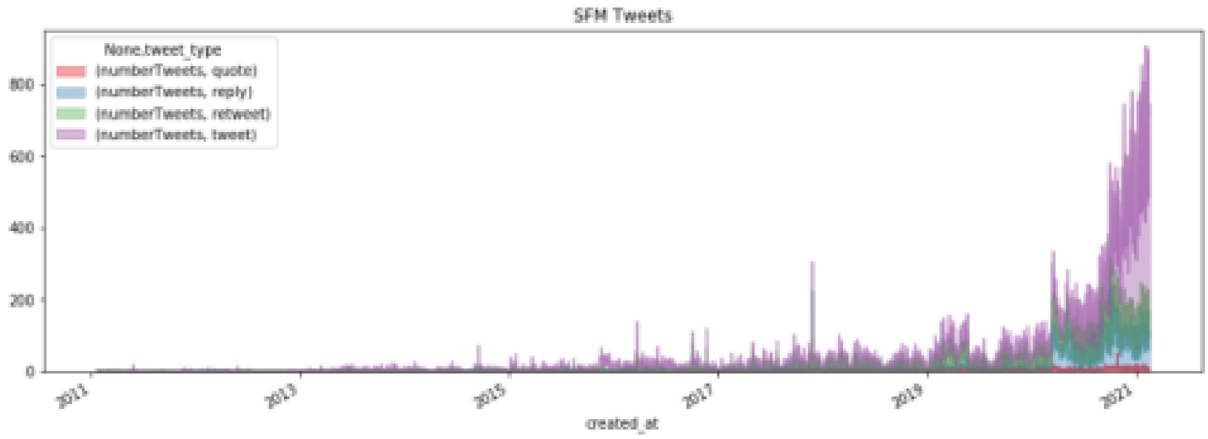


Figure 1: the temporal distribution of harvested account collection tweets.

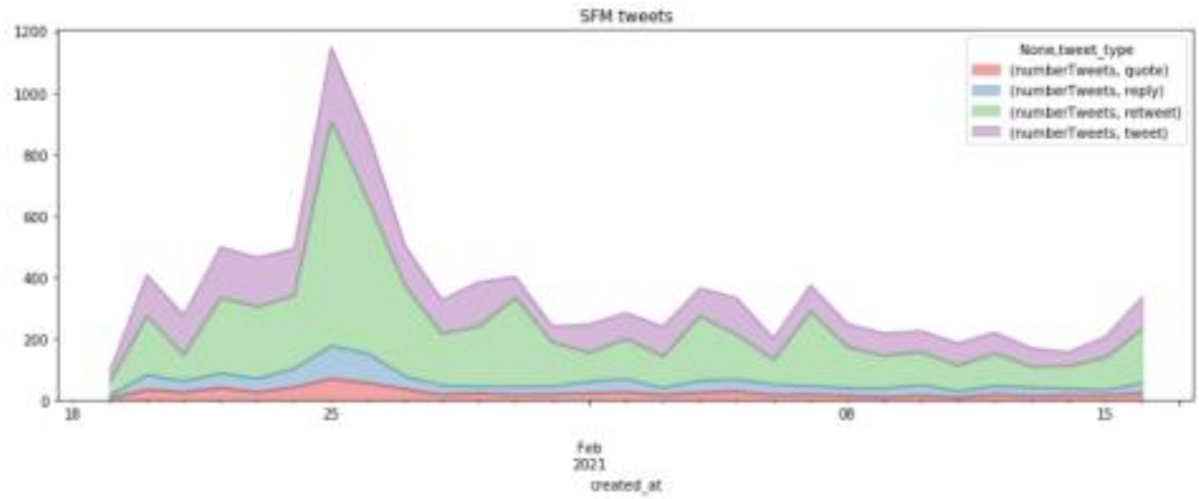


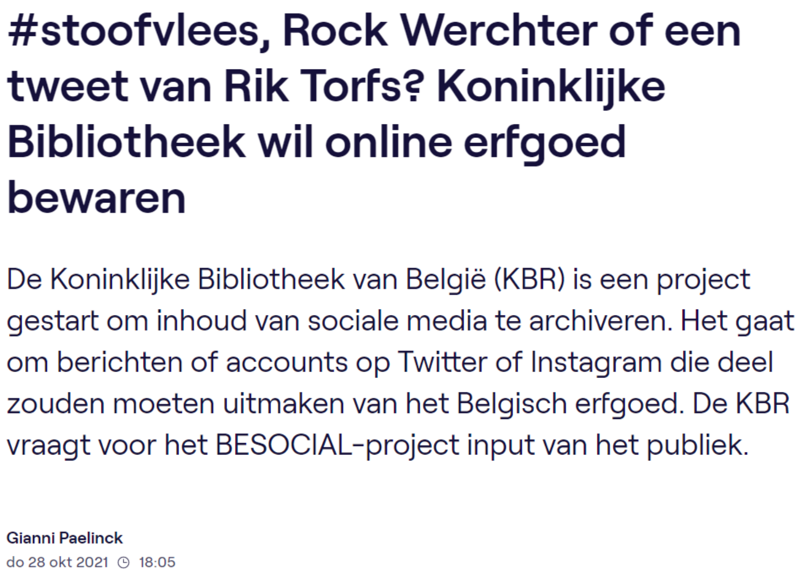
Figure 2: the temporal distribution of harvested hashtag collection tweets.

For the hashtag collection, on the other hand, figure 2 shows that hashtags are only captured a few days to the past. Interestingly retweets are a large part of this collection. A deeper analysis of co-occurring hashtags, excluding the ones we looked for, can inform future adaptations of this collection, i.e. including identifying and adding other relevant hashtags.

The WP2 and WP3 report documents the findings from the mini-pilot with the potential problems and disadvantages. With this crawl, the team can determine the feasibility of the harvesting, test the appropriateness of data collection, and test the process (timing) and obtain preliminary data. All this with the aim to limit possible pitfalls.

In addition to setting up a mini-pilot, the selection component also includes ensuring the representativeness of the final seed list that will be created. Based on this seed list, a choice is made of which social media (hashtags and accounts) will be harvested or not. That requires the development of a specific selection policy that details what cultural heritage of Belgium is on social media in the form of hashtags and handles/profiles, as well as what is legally allowed to be collected and stored.

The BESOCIAL team started early on in the project to create a list that would represent this heritage on social media in Belgium. It quickly became apparent that in order to ensure this validity, we could not just use our team to supplement our seed list with the culture-related hashtags and accounts, but that feedback from a broader audience was necessary. In an effort to ensure the validity of this list, a crowdsourcing campaign was developed with KBR’s communication team, where the public (expert or not) can contribute to the collection by way of suggesting hashtags and handles from social media in Belgium on the topic of cultural heritage.The campaign succeeded not only in appealing to the Belgian public, but also in suggesting as many hashtags and accounts as possible. At the beginning of September 2022, the number of suggested hashtags and accounts is around 700. Indirectly, BESOCIAL and social media archiving were put in the picture in the press. This was also a way for KBR to put forward this innovative project.



#stoofvlees, Rock Werchter of een tweet van Rik Torfs? Koninklijke Bibliotheek wil online erfgoed bewaren

De Koninklijke Bibliotheek van België (KBR) is een project gestart om inhoud van sociale media te archiveren. Het gaat om berichten of accounts op Twitter of Instagram die deel zouden moeten uitmaken van het Belgisch erfgoed. De KBR vraagt voor het BESOCIAL-project input van het publiek.

Gianni Paelinck
do 28 okt 2021 ◷ 18:05

Figure 3: Extract from article on crowdsourcing campaign as published on https://www.vrt.be/vrtnws/nl/2021/10/28/online-erfgoed/.

Task 2.2. Analysis of user requirements (Lead: UGhent-MICT; KBR) (M4-M9)

Future users of the BESOCIAL archive can be diverse, from the public to researchers exploring the past on social media. In order to attempt to speculate on these needs, a list of personas was created to verify the needs of a broad range of stakeholders (researchers, cultural heritage professionals, publishers, policy-makers and other potential end users). These five personas (i.e.: Pierre- a 42-year old postdoc researcher in communication sciences; Febe- a 24-year old PhD student in computational linguistics; Ben- a 35-year old journalist working for the newspaper De Standaard; Manou- a 28-year old scientific researcher at KBR; and Marcel- an 68-year old former team coordinator at a car manufacturer, now retired) are concisely described in terms of their background, expertise and goals for using a social media archive. Next, a preliminary interview guide was constructed. Using these five persona as an 'ideal' we explored our personal and professional networks to find suitable interviewees and contacted them using email. In total six interviews were conducted in Spring 2021 with individuals who fit this criteria, to sketch the needs of users.

Although our results are based on a small number of interviews we believe that they provide some interesting insights that can partially remediate the antithetical nature of web archives to research as argued by Ogden and Maemura (2021). In terms of selecting which social media content to archive (and which not) we note the need for (a stronger) inclusion of the academic field in selection decisions and policies. Despite the fact that diverse interests in what should be archived exist, we also noted that most interviewees agree on how the archived content should be query-able. However, the idea of classic search interface would often not suffice and further development of interfaces that open up and make searchable archived social media should be explored (e.g. an interface based on visualizations).

In general interviewees stated that 'one should aim to provide as much information on any given collection as possible'. Interestingly for some this should also include references to certain methodologies or certain software or tool sets that can be used to analyse the collection as well as to articles or research papers that (partially) used the collection in the past. As such, KBR Digital Research Lab could play an important role in involving the research community by encouraging researchers to make use of the resources contained in web archives and by supporting researchers to the fullest extent possible. This is important as researchers often lack the necessary (digital) skills or domain knowledge or neglect (or are unaware) of the big data and legal features of most social media archives. The preliminary results were validated via a pre-conference workshop at RESAW titled 'User requirements for working with social media web archives', on 14 June 2021.

Task 2.3. Analysis of existing legal frameworks in Belgium for selection of content for social media archiving (Lead: CRIDS; KBR) (M4-M15)

During the first months of the project, in terms of legal aspects impacting the selection of content, the research focused on copyright and data protection aspects. On the one hand, under copyright law, collecting and harvesting copyrighted social network content constitutes an "act of reproduction" requiring, in principle, the authorisation of the right holders. The principle of such prior authorisation and the possibility of using an exception were analysed. On the other hand, from the point of view of data protection law, the selection, harvesting and archiving of social network content containing personal data constitute "data processing" involving the application of the GDPR. In this respect, the GDPR provides for interesting derogatory regimes for archiving in the public interest and for scientific and historical research. The scope and implications of these derogatory regimes have been analysed in order to identify the implications for the BESOCIAL project. The results relating to the implications for SMA and web archiving of the derogatory regime for archiving in the public interest have been valorised through a book chapter that was published in March 2021 (Michel, 2021) (see Section 6).

In parallel to these analyses, results relevant for Task 4.1 "legal considerations concerning access to social media archive" were highlighted with respect to copyright, privacy and data protection. On the basis of the desired options developed in the framework of the Mini-Pilot, the CRIDS pointed out areas of concern and proposed solutions to comply with legal requirements. In particular, the rules stemming from the Belgian law of 30 July 2018 for the communication to identified third parties and the dissemination to unidentified third parties of personal data processed for archiving in the public interest purposes and for scientific research purposes were presented and implemented in the context of access to archived tweets for the public.

A case law chronicle of interesting decisions of European and Belgian courts and tribunals in media law is also in progress. This analysis provides interesting considerations for BESOCIAL regarding the right to be forgotten and of dereferencing, the accessibility of online archives (heritage and press), the balancing of privacy and data protection rights against freedoms of expression and information, content moderation, content deletion, access blocking, etc.

Task 2.4. Definition of the technical and functional requirements based on the OAIS model (Lead: KBR; GhentCDH) (M4-M9)

The Royal Library of Belgium (KBR) created a “wish” list of functional and technical requirements with a focus on long term preservation in the framework of the BESOCIAL project. The document identifies and describes the technical and functional requirements for the pilots (WP3 pilot for social media archiving, and WP4 pilot for access to the social media archive). The requirements within this document are structured according to the core functions: selection and harvesting. Task 5.2 finalised this "wish" list into an operational document of requirements.

Work Package 3: Pilot for social media archiving (M7-24)

This work package aimed to outline the policy for the harvesting of social media data and collecting the content, as well as quality control and plan for the future. A more detailed report of Work Package 2 and 3 can be found in the form of a report that is made available on the Orfeo Platform and communicated to the follow up committee: https://orfeo.belnet.be/handle/internal/10034.

Task 3.1. Development of a social media harvester and harvesting social media content (Lead: UGhent-IDLab; CENTAL) (M7-M24)

The social media harvester setup of IDLab integrates the targeted social media platforms, Twitter and Instagram, unifies their messages through semantic annotation to embed an overlay, uniform vocabulary, and adds extra metadata on versioning, origins and applied policy. IDLab created documentation for the use and deployment of Social Feed Manager and Instaloader. For both tools it includes how to provide API or login credentials, how to set up harvests for accounts and hashtags, how to start and monitor harvests and how to export these harvests afterwards. For Instaloader we paid special attention to how to set up periodic harvesting as this is not supported out of the box.

IDLab provided semantic annotations for the different harvests through the use of RDF Mapping Language (RML) rules. RML allows defining declaratively how semantic annotations are added to existing data. The benefit of RML rules is that it detaches defining the rules from executing them, which in turn allows for more flexibility in choosing which tools are used to define them and which tools are used to execute them. All semantic annotations together are also called a knowledge graph.

The data model used for the semantic annotations follows the Europeana Data Model (EDM). It is based on standards such as Lightweight Information Describing Objects for museums, Encoded Archival Description for archives or Metadata Encoding & Transmission Standard for digital libraries. IDLab created the RML rules based on the aforementioned data model. It used the open-source tool RMLMapper to execute these rules and generate the knowledge graph. The knowledge graph only contains the semantic annotations about the content harvested so far. If new content is harvested the RML rules have to be executed on the new content only.

Within the context of web archiving, different kinds of users may want to interact in different use cases with the data.

> *A data scientist wants to perform analysis and may require machine readable data facilitating analysis.*

> *A social scientist wants to perform a study and may require the original look and feel of harvested data.*

The needs of these different user roles for different use cases can be represented using constraints on our knowledge graph expressed in separate data shapes. Different user roles need the data in different formats and this need can be represented as constraints. A validation with these constraints can inform users by saying how relevant the data are for them, i.e. available in an appropriate format and license and inform further harvests, i.e. indicate for which collection elements (tweets) a harvest in a different format is needed. SHACL shapes can be used to validate if an element is available in a specified view, i.e. if a tweet is available in HTML, useful for social scientists and/or in JSON, useful for data scientists.

Within archiving metadata records, summarizing collections exist and are called “finding aids”. Such summarizations are smaller than all the actual data and aim for human consumption. A SHACL validation on the elements of a collection can inform a summarized metadata record of the collection. The result is a SHACL validation report indicating how many elements, e.g. tweets, adhere to the shape. A SPARQL query getting the number of elements and number of violations can then generate a percentage which may lead to an information such as "83% of this election tweet collection preserved the look and feel" or "100% of this election tweet collection exist in JSON format".

Task 3.2. Quality control of harvested content (Lead: UGhent-MICT&GhentCDH; KBR) (M10-M24)

This task focused on assessing the quality of the harvested content in BESOCIAL. In specific, we looked at the content that was harvested during BESOCIAL’s ‘mini-pilot’ (Twitter COVID19 related content), and at how this content was opened up as ‘data dump’, i.e. a zipped folder containing various subfolders with the harvested data as .json and .csv, including the README.txt file. In order to do this, we took a three folded approach. A first line of inquiry into the ‘quality’ of the harvested content involved a qualitative research design that started from the different personas that were developed in WP2 ((i.e.: Pierre- a 42-year old postdoc researcher in communication sciences; Febe- a 24-year old PhD student in computational linguistics; Ben, a 35-year old journalist working for the newspaper De Standaard; Manou, a 28-year old scientific researcher at KBR; and Marcel (an 68-year old former team coordinator at a car manufacturer, now retired). We created narratives – based on in-depth interviews – on how the persona would work with the provided harvested content and did a SWOT analysis using the perspectives of the personas.

Next to this persona-driven research approach, we also took a more computational approach to assess the usability and quality of the content of the BESOCIALs’ harvesting ‘mini-pilot’ by experimenting and exploring the harvested data in Tableau as well as Jupyter notebooks, two out-of-the-box or off-the-shelf software tools that are suitable for exploring and visualizing data. The third and last line of inquiry in our three folded approach involved using the benchmark of datasets for computationally-driven research developed by Candela et al. (2021).

Our first persona-driven approach showed some clear results. Most notably, it highlighted the practical and down-to-earth issues that would impede persona such as Manou and Jan (both with rather low computer and coding skills) in working with the harvested content and it underlined how straightforward and easy this job would be for a persona such as Febe (who has high computer and coding skills). An analysis from the perspective of Manou and Jan showed for example that the sheer size of the data limits their possibilities of exploring the data. This is also the case for the very cryptic and complicated filenames or the lack of clear and exhaustive documentation about the harvested corpus. While the lack of good documentation was also noted by the persona Febe, contrary to the other personas, the json-dump of the COVID19-collection harvested during BESOCIAL felt very 'natural' to her as she had the tools and expertise to process, clean and analyse such data. Our analysis shows that 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.

The Jupyter notebooks-case showed that people who are proficient in Python and in working with the Jupyter environment do not encounter many hurdles working with the harvested data. This case however also showed that, for researchers with no experience with Twitter or for researchers not proficient in the languages included in the database, more contextual information should be provided in advance on what the data set is about in order to identify the specific domain knowledge needed. The Tableau-case showed similar results as it demonstrated that potential users with skills in working with Tableau can easily create basic visualisations of the harvested content.

In our third and last line of inquiry we tried to apply the benchmark developed by Candela et al. (2021) to the content harvested during BESOCIALs' 'mini-pilot' (and to how this content was presented as .json and .csv files). This exercise resulted in a total score of 5. It showed that the criterion 'license', (closed licenses which are less permissive and limit the usage), the criterion 'terms of use' and the criterion 'prototypes and documentation' can be improved further (e.g. by providing examples of code of scripts that can be applied to the database in order to clean, search or analyse the data or by providing examples of use, e.g. a published article that uses the database as a data resource).

Task 3.3. Development of a preservation plan for archived social media (Lead: KBR; GhentCDH) (M7-M18)

This task focused on the development of a preservation plan for archived social media linked to the results of prior research in earlier work packages within the BESOCIAL project (T1.4 Analysis of Preservation Policies) and the first tests of collecting social media content (T3.1 and T3.2) that provided us with the necessary information to develop a preservation plan for archived social media. The plan covers aspects such as the definition of preservation formats, the description of the quality control and the different stages files go through etc. The adaptability of KBR's current preservation workflow was also studied in order to define the necessary preservation actions for social media archives after the end of the project.

The document is divided into three parts. Firstly, Get to know your organisation and your data: in this phase we painted the scene of the readiness for digital preservation at KBR. An extra step here is outlining the corpus of BESOCIAL. The second level is writing down the bit-level preservation for BESOCIAL. This phase is the minimum level of preservation that is needed and is mostly linked to storage and risk management. And thirdly, Long Term Preservation Plan: in the last and the most important phase we identified a framework using the OAIS model for preserving the WARC and JSON files for a sustainable long-term platform.

Scientific data related to this task can be found here: https://doi.org/10.34934/DVN/9VCZPB.

Work Package 4: Pilot for access to social media archive (M16-M21)

This work package focused on detailing and developing a plan for providing access to the social media archive via a dedicated access platform.

Task 4.1. Legal considerations concerning access to social media archive (Lead: CRIDS; KBR) (M16-M21)

This task has been achieved within the global legal report issued in December 2021, alongside the legal considerations on the selection and preservation of content (task 2.3.). This report exploits previous analyses that were made during the first year of the BESOCIAL project and contains new analyses.

Preservation and access questions have been analysed together in that report, following a structure divided between the different legal matters at stake: privacy aspects of social media archiving (including the need to balance the freedom of information and the right to privacy as well as an analysis related to personal data protection), copyright aspects and some open data considerations. The choice to analyse these two aspects together was motivated by the fact that both analyses were often based on the same legal reasoning and that it would be counterproductive to try and divide them.

Two recapitulative sections were included at the end of the report: the first one provides a summary of the recommendations for the modification of the Belgian law on legal deposit that were made throughout the report (which are meant to nourish task 5.1) and the second one contains a summary of the legal aspects for the archiving of social media content and the access to archived social media content. In this last part, we therefore divided the recommendations made in the report between preservation and access in order to bring more coherence regarding the division of the tasks of the project (2.3 and 4.1).

In this last section and regarding access to social media archives specifically, we highlighted several main recommendations that have to be taken into account regarding privacy and copyright considerations. First, the right to privacy as well as personal data protection entails that the pseudonymisation of private information and personal data should be envisaged in some cases when giving access to the archives. Second, the application of copyright law implies that the archives protected by a copyright will have to be accessible on KBR's premises only.

Another report was issued in May 2022 on the application of the image right to social media archiving. Our main findings were that it is difficult to articulate both of these notions. It however confirmed the need to minimise the amount of images collected as part of the archives.

A report on the archiving of unlawful content was also produced.

Task 4.2. Development of access platform and search environment (Lead: CENTAL; UGhent-IDLab) (M7-M21)

The objective of this task was to develop a platform to access the collected data but also to enrich this data with semantic data such as named entities, similarity vectors, etc.

We decided to create this interface in partnership with the MiiL (Media Innovation & Intelligibility Lab) given the scope of the task. We are also using their data to test the interface, which gathers thousands of tweets. Once the test phase is completed, we will inject the BESOCIAL data. In terms of the technologies used to create the interface, we have chosen to use ElasticSearch, which is a software that allows us to index and search a large amount of data, as well as the javascript language for back-end and front-end development.

This allowed us to create an interface with a number of functions, including the following: a simple search in the index, a more complex search by choosing either the exact form or the verbal or adjectival form, etc. of the occurrence(s) searched for, or a search by hashtag. In terms of results, we find all sorts of graphics and a word cloud as well as the possibility of exporting the results searched in csv format. In parallel, we have added an authentication system so that users can receive the exported files by email. In addition, we have also created named entity models for the different languages concerned on noisy data because the classic entity recognisers do not work on data with spelling variations, which do not start with a capital letter and do not end with a dot, etc. We have obtained good results and we are in the process of integrating them into the platform.

We updated our first version according to the recommendations made in *task 4.3. Evaluation of the Belgian pilot social media archive* to make the interface more user-friendly and intuitive for a first version (change the position of the search button, add more explanatory text, etc).

A more detailed report can be found in the form of a report that is made available on the Orfeo Platform and communicated to the follow up committee:

Task 4.3. Evaluation of the Belgian pilot social media archive (Lead: UGhent-MICT) (M7-M21)

The goal of this task was to evaluate, gain insights and provide feedback on the functionalities, look and feel and usability of the Belgian pilot social media archive-interface at http://130.104.253.27/. These insights are then taken into account when the interface to the social media archive is redesigned or further elaborated on. After various meetings with the involved consortium partners in March 2022 it was jointly decided to tackle this task by means of an internal expert-review. This internal expert-review of the Belgian pilot social media archive-interface was conducted by 3 professionals during 4 workshops in May and June 2022. We decided to first focus on providing feedback on the various aspects of the current interface. Next we elaborated on how this current interface can be modified or adapted to even better serve the needs and requirements of potential users in terms of functionality and usability.

Work Package 5: Recommendations for sustainable social media archiving in Belgium (M16-M24)

This work package focused on developing the documentation and recommendations for sustainable social media archiving. This includes legal, technical and functional aspects as well as developing a business model and embedding social media archiving in KBR as an institution.

Task 5.1. Legal recommendations concerning social media archiving (Lead: CRIDS; KBR) (M22-M24)

This task has been partly achieved within the global legal report issued in December 2021. In the meantime, an analysis of image rights was added to this report as well as recommendations to KBR. These recommendations will be based on the different legal analyses carried out during the research project. The operational procedures that will be developed in T5.5 will be based on these legal recommendations.

Task 5.2. Technical and functional requirements for an operational social media archiving system in Belgium (Lead: KBR; GhentCDH, CRIDS) (M16-M21)

The purpose was to ensure that the results of the technical and functional requirements outlined in T2.4 will be reworked based on the research results obtained during the pilot (WP3 Pilot for SMA and WP4 Pilot for access to social media archive). The operational link was made with the IT infrastructure at KBR.

The purpose of this document was to outline the technical and functional requirements of the BESOCIAL corpus, based on the OAIS model. In Task 2.4 a start was made in a first version where a wishlist of functional and technical requirements was created with a focus on long-term preservation in the framework of the BESOCIAL project. This report focused more on what is **feasible** within KBR on the **operational level** of the technical and functional requirements. The knowledge gained in various tasks (e.g. Task 3.3 - preservation plan, and Task 4.3 - Evaluation of the access platform) was included in this final list of requirements.

In the document, we explained the responsibilities for Social Media Archiving (SMA) for each requirement in 3 scenarios:

- **Scenario A**: During the BESOCIAL research project (2020-2022)

  After the ending of the BESOCIAL research project
- **Scenario B**: a specific person and team will be responsible for embedding social media archiving within KBR. This will likely take the form of a BRAIN-BE funded project for 3 years.
- **Scenario C**: a third-party service provider will be responsible for social media archiving for KBR. We will make this exercise with the most optimistic outcome in mind. The service provider will manage most phases of the social media archiving (SMA), except from selection and linking the seed lists to the chosen harvest tool(s). At the end of the SMA cycle the service provider will transfer the data to KBR to ingest it in the LTP (Long Term Preservation platform) of BELSPO. This to assure sustainable preservation of the captured data.

Task 5.3. Business model development (Lead: UGhent-MICT & GhentCDH; KBR) (M16-M21)

The business model framework that was chosen for this study was the Service Dominant Business Model since it is better suited to model social media archiving activities than classic business models based on the manufacturing economy. The following stakeholders were identified in this study: KBR, (domain) experts, external provider for web archiving services, Belgian heritage institutions, users, the Belgian Science Policy Office (BELSPO), and the Belgian society at large.

The value-in-use proposition for the users was defined as ‘sustainable & easy access & use of KBR social media archive collection’ and two business scenarios were prepared prior to a workshop. In scenario A, KBR manages the social media archive from start to finish (including archiving and providing access). In scenario B, an external third-party service provider manages the social media archive (from start to finish).

A SDBMR workshop was organized on Friday September 2nd 2022 involving three external experts. The participants of the workshop, focused in several iterations, and in small groups, on answering the different questions related to the SDBMR methodology. This exercise clearly showed the huge task load of KBR in scenario A, in which KBR, as orchestrator, not only needs to manage the social media archiving process from start to finish but also needs to put time and effort in processes such as promotion, policy creation, lobbying, and communication about SMA. Some win-win scenarios were detected to minimize costs (e.g., by not storing and preserving the same material) for Belgian heritage institutions. Still, the costs for KBR are significant (e.g., infrastructure & personnel costs, potential legal costs & advice, cost of running SMA programme, (potential) cost for offsetting environmental costs, …) in scenario A and will need to be covered by additional long term structured funding as no monetary compensation will be asked from users of the social media archive. As a final outcome of the workshop a business model radar was also developed for scenario A (see Figure **x**). The model points to the financial as well as non-financial costs and benefits of each actor involved in the value proposition as well as their related activities and value proposition to serve the co-created value-in-use that was defined as 'sustainable & easy access & use of KBR social media archive collection'.

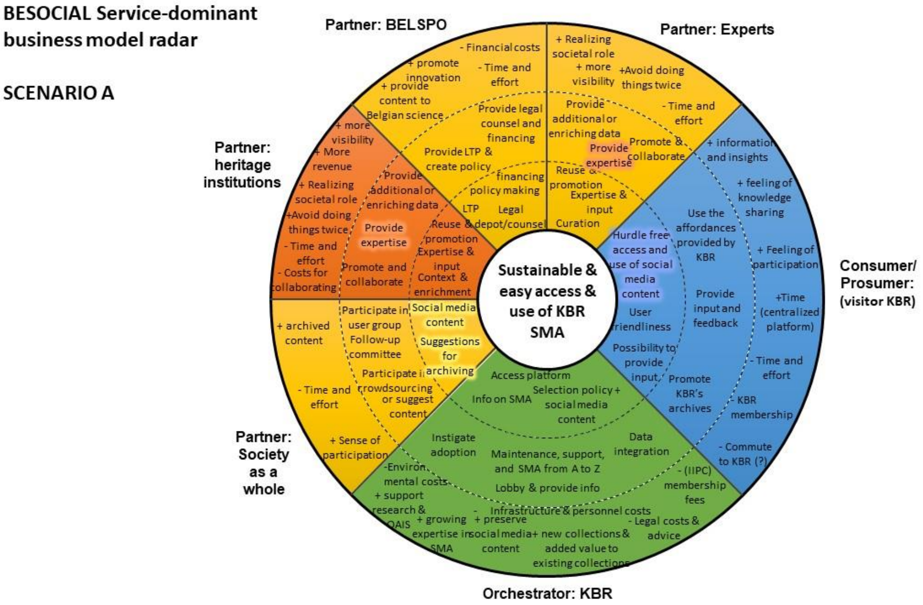

BESOCIAL Service-dominant business model radar
SCENARIO A
Partner: BELSPO
- Financial costs
+ promote innovation
- Time and effort
+ provide content to Belgian science
Provide legal counsel and financing
Provide LTP & create policy
financing policy making
LTP
Legal depot/counsel
Partner: Experts
+ Realizing societal role
+ more visibility
+Avoid doing things twice
- Time and effort
Provide additional or enriching data
Provide expertise
Promote & collaborate
Reuse & promotion
Expertise & input
Curation
Partner: heritage institutions
+ more visibility
+ More revenue
+ Realizing societal role
+Avoid doing things twice
- Time and effort
- Costs for collaborating
Provide additional or enriching data
Provide expertise
Promote and collaborate
Reuse & promotion
Expertise & input
Context & enrichment
Consumer/ Prosumer: (visitor KBR)
+ information and insights
+ feeling of knowledge sharing
+ Feeling of participation
+Time (centralized platform)
- Time and effort
- KBR membership
- Commute to KBR (?)
Use the affordances provided by KBR
Provide input and feedback
Promote KBR's archives
Hurdle free access and use of social media content
User friendliness
Possibility to provide input
Sustainable & easy access & use of KBR SMA
Partner: Society as a whole
+ archived content
- Time and effort
+ Sense of participation
Participate in user group
Follow-up committee
Participate in crowdsourcing or suggest content
Social media content
Suggestions for archiving
Orchestrator: KBR
Access platform
Info on SMA
Selection policy + social media content
Instigate adoption
Maintenance, support, and SMA from A to Z
Data integration
Lobby & provide info
-Environmental costs
+ support research & OAIS
+ growing expertise in SMA
+ preserve social media content
- Infrastructure & personnel costs
+ new collections & added value to existing collections
- Legal costs & advice
- (IIPC) membership fees

Figure 4 Abstracted SDBM radar visualization scenario A

After resolving the radar for scenario A, additional time was spent during the workshop in mapping the differences and creating a separate radar for scenario B (see Figure 5). This showed potentially different types of cost for KBR in scenario B. Benefits on the other hand include a speedier process, the availability of a copy of the data, ... However, an external third-party SMA provider setting certain limitations or constraints, might also lead to certain costs for other network actors such as visitors or heritage institutions.

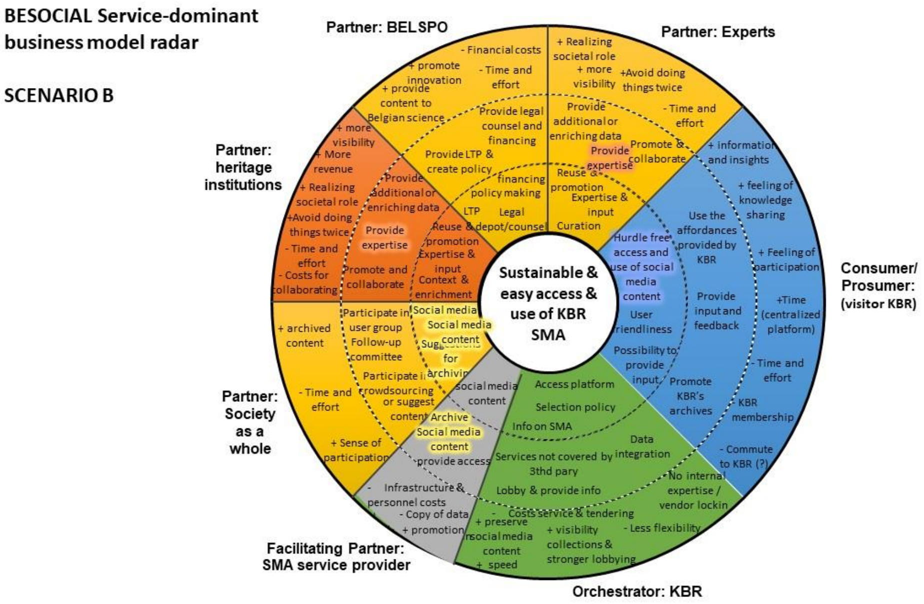

BESOCIAL Service-dominant business model radar
SCENARIO B
Partner: BELSPO
Partner: Experts
Partner: heritage institutions
Consumer/ Prosumer: (visitor KBR)
Partner: Society as a whole
Facilitating Partner: SMA service provider
Orchestrator: KBR
Sustainable & easy access & use of KBR SMA
- Financial costs
+ promote innovation
- Time and effort
+ provide content to Belgian science
Provide legal counsel and financing
Provide LTP & create policy
financing policy making
LTP
Legal depot/counsel
+ Realizing societal role
+ more visibility
+Avoid doing things twice
- Time and effort
Provide additional or enriching data
Provide expertise
Promote & collaborate
Reuse & promotion
Expertise & input
Curation
+ more visibility
+ More revenue
+ Realizing societal role
+Avoid doing things twice
- Time and effort
- Costs for collaborating
Provide additional or enriching data
Provide expertise
Promote and collaborate
Reuse & promotion
Expertise & input
Context & enrichment
+ information and insights
+ feeling of knowledge sharing
+ Feeling of participation
+Time (centralized platform)
- Time and effort
- KBR membership
- Commute to KBR (?)
Use the affordances provided by KBR
Provide input and feedback
Promote KBR's archives
Hurdle free access and use of social media content
User friendliness
Possibility to provide input
+ archived content
- Time and effort
+ Sense of participation
Participate in user group Follow-up committee
Participate in crowdsourcing or suggest content
Social media
Social media content
Suggestions for archiving
social media content
Archive Social media content
provide access
- Infrastructure & personnel costs
- Copy of data
+ promotion
Access platform
Selection policy
Info on SMA
Services not covered by 3thd pary
Data integration
Lobby & provide info
- Costs service & tendering
No internal expertise / vendor lockin
+ preserve social media content
+ speed
+ visibility collections & stronger lobbying
- Less flexibility

Figure 5 Abstracted SDBM radar visualization scenario B

Task 5.4. Definition of a Belgian strategy for social media web archiving including cost calculation and Governance (Lead: KBR; UGhent, CRIDS) (M19-M24)

The research results of the previous work packages fed into this strategy so that the technical, legal, operational, organisational and user-related elements were taken into account. Attention was also paid to governance, for example how to organise the collaboration with external partners for the selection of content etc. and how the KBR scientific committee can provide advice for the social media archive. A cost calculation was developed based on three different scenarios: 1) KBR can hire a social media expert, 2) social media archiving becomes part of the tasks of the expert in web archiving or 3) KBR does not continue with social media archiving but stays up to date about new developments in the field by attending conferences. This cost calculation was submitted to the KBR Board of Directors along with a legal analysis about the risks involved to make an informed strategic decision. The Board of Directors decided that KBR would continue with social media archiving and would integrate it in the public procurement procedure for web archiving services that is currently in development.

Task 5.5 Institutional embedding of social media archiving at KBR (Lead: UGhent-GhentCDH) (M19-M24)

The purpose of this task was to ensure that the results and recommendations of the BESOCIAL project (2020-2022) are fully embedded into the daily operations at KBR. This includes ensuring that selection, access and preservation policies are developed in close liaison with the relevant departments within KBR. Similarly, the technological aspects of the project are embedded in close collaboration with KBR's IT team.

To execute this embedding exercise, three scenarios were developed which present three operational possibilities for sustainably embedding social media (archiving) in KBR in the long term. These scenarios will be analysed and synthesised from a helicopter perspective in the conclusion.

The scenarios we developed consist of 3 possible directions in which KBR can embed the concept of social media (archiving) in the long and short term:

- **Scenario A**: a specific person and team will be responsible for embedding social media archiving within KBR. This will likely take the form of a follow-up project at KBR on Social Media Archiving.
- **Scenario B**: social media archiving is considered to be part of KBR's web archiving activities. The person responsible within KBR for web archiving will, in addition to his/her already existing tasks, keep social media archiving active by mutual agreement and to the extent possible. This will take up a maximum of 30% of his/her range of duties.
- **Scenario C:** Passive embedding of social media within KBR; do nothing, but keep a passive trace of KBR's social media (executed by the communication department at KBR). This can take the form of renewing the IIPC subscription every year and attending their conferences.

The format of this document is a short, practical and action-driven report, outlining concrete actions for how to move forward with social media archiving at KBR after the BESOCIAL project. This report provides context and guidelines for users internal and external to KBR. Internal users are interpreted as archivists, librarians and administrative collaborators while the external stakeholders are web content creators and administrators, end users of the social media archive (e.g. researchers) and the general public.

In conclusion we can say that If we look at the various action points from a helicopter perspective, there is a need within KBR to develop documentation on the process of preserving social media data. Especially the operational side of the infrastructure at KBR needs to be documented in an understandable way (e.g. implementation of the OAIS model). In addition, we also see the internal need for working groups that in their turn bring expertise together to guarantee this embedding (e.g. Digital Data Working Group). By maintaining and expanding the national and international born-digital data network (e.g. setting up expert groups around certain themes), we can also use their knowledge to reach our goals.

The BESOCIAL research project has already taken several steps towards the preservation of social media within KBR. Now, the main goal is to ensure that this knowledge is not lost, and can be transformed into a long-term vision for sustainable archiving and preservation of Belgian social media.

Task 5.6. Definition of procedures (Lead: KBR) (M19-M24)

This task focused on providing procedures and guidelines for users internal and external to KBR. Internal users are interpreted as archivists, librarians and administrative collaborators while the external stakeholders are web content creators and administrators, end users of the social media archives and the broad public. These procedures and guidelines have been drafted taking into account the research results obtained within the project during the previous work packages in order to integrate them in the operational, technical and legal contexts within which the KBR function.

This document consolidates all findings of WP5 into clear procedures for Social Media Archiving (hereafter referred to as SMA) at KBR. The entire workflow of SMA was taken into account as well as the legal, technical, operation and user-oriented perspectives of the research. The procedures covered the selection, harvesting, ingest, data management, storage, preservation, access, administration and strategic management.

A separate annex was created to document various guidelines for the use of certain tools (Social Feed Manager, Instaloader, Conifer, and Heritrix).

Work Package 6: Coordination, dissemination and valorisation (Lead: KBR) (M1-M24)

Work package 6 concerns the coordination, dissemination and valorisation of the project, which occurs throughout the entire project.

Task 6.1 Project management and coordination (Lead: KBR) (M1-M24)

Monthly meetings with the BESOCIAL team have been organised, as well as bi-lateral meetings in the context of the specific tasks. The second, third and fourth meeting of the follow-up committee took place on 22 June 2021, 25 January 2022, and 15 September 2022. The output of the different work packages and tasks is regularly sent to the scientific follow-up committee.

Task 6.2 Dissemination activities (Lead: KBR ; UGhent-All, CENTAL, CRIDS) (M1-M24)

The internal communication within the project team and the management is handled by email, shared documents in a dedicated folder on Google Drive and in online meetings. A dedicated webpage has been created on the website of the KBR for the project: https://www.kbr.be/en/projects/besocial/.

We communicate information for social media through KBR's communication office on the social media accounts of KBR and of the other project partners.

Task 6.3 Valorisation activities (Lead: KBR; UGhent-All, CENTAL, CRIDS) (M6-M24)

See Section 'Dissemination and Valorisation'' for a detailed list of publications and outcomes of the projects.

Work Package 7: Helpdesk for legal enquiries (Lead: CRIDS) (M1-M24)

A working document was set up where CRIDS provides ongoing consultation throughout the project which is referred to as the helpdesk. This relates to relevant questions about privacy, data protection and ICT law encountered by the partners in their tasks.

This helpdesk has led to a first version of an internal FAQ document. This document both summarises these questions and answers and contains the main conclusions of the global legal report presented in a question and answer format, with more to the point legal analyses to facilitate their application. The first version contains three parts: the first one for questions related to data protection law and the right to privacy, the second one for questions related to intellectual property law and the third one for questions related to the modification of the law on legal deposit. The second and final version of the FAQ document has been issued in August and includes an additional part containing questions and answers related to open data matters, coming from both the helpdesk and our own analysis.

5. DISSEMINATION AND VALORISATION

5.1 Conferences attended:

- (2020, August 26 - 27). IIPC RSS WEBINAR: WEB ARCHIVING SOCIAL MEDIA AND NEWS WEBSITES, online.
- (2021, April 20 - 22). WARCnet AARHUS MEETING, online.
- (2021, April 22). DH Virtual Discussion Group for Early Career Researchers, online.
- (2021, April 29). LIBER Citizen Science Working Group workshop 'Citizen Science: Framing the roles of libraries, online.
- (2021, May 27). DH Virtual Discussion Group for Early Career Researchers, online.
- (2021, June 2-4). DH BENELUX 2021 Leiden, online.
- (2021, June 15-16). IIPC WEB ARCHIVING CONFERENCE, online.
- (2021, June 8). KU Leuven Bibliotheken: Onderzoeksseminarie Digital Humanities, online.
- (2021, September 6 - 9). SEMANTICS 2021 -17th International Conference on Semantic Systems, online.
- (2021, June 2 - 4). DH Benelux 2021 Leiden, online.
- (2021, June 15 - 16). IIPC Web Archiving Conference, online.
- (2021, September 6 - 9). Semantics 2021 -17th International Conference on Semantic Systems, online.
- (2021, October 25, May 23) DH Early Career Discussion Group, online.
- (2022, January 19) IIPC Social Media Archiving Workshop, online.
- (2022, March 24) Study day social media archiving, Liège.
- (2022, April 25) DH lunch at the University of Texas at Austin, online.
- (2022, May 23 - 25) IIPC Web Archiving Conference, online.
- (2022, June 1-3) DH Benelux, online and in Luxembourg.

5.2 Presentations given

- Sally Chambers and Jessica Pranger (2020, August 26 - 27). Towards a sustainable archiving social media archiving strategy for Belgium [Conference Presentation]. IIPC RSS WEBINAR: WEB ARCHIVING SOCIAL MEDIA AND NEWS WEBSITES, online.
- Fien Messens (2021, May 27). Introducing BESOCIAL - Towards a sustainable archiving strategy for social media in Belgium [Elevator Pitch]. DH Virtual Discussion Group for Early Career Researchers, online.
- Fien Messens et al. (2021, June 2-4). Introducing BESOCIAL- towards a sustainable archiving social media archiving strategy for social media in Belgium [Poster Presentation]. DH BENELUX 2021 Leiden, online.
- Fien Messens (2021, June 8). Introducing BESOCIAL [Conference Presentation]. KU Leuven Bibliotheken: Onderzoeksseminarie Digital Humanities, online.
- Friedel Geeraert (2021, June 14). Update Web Archiving at KBR + Update BESOCIAL [Update]. General Assembly IIPC WEB ARCHIVING CONFERENCE, online.
- Peter Mechant, Eveline Vlassenroot, Sally Chambers, Niels Brügger, Susan Aasman, Friedel Geeraert and Jessica Pranger. (2021, June 14) [pre-workshop]. User requirements for working with social media web archives, RESAW conference.
- Eveline Vlassenroot, Friedel Geeraert, Sally Chambers, Peter Mechant, Fien Messens and Julie M. Birkholz (2021, June 15-16). Unlocking web and social media archives for

humanities research: a critical reflection [Conference Presentation]. IIPC WEB ARCHIVING CONFERENCE, online.

- Sven Lieber, Dylan Van Assche, Anastasia Dimou, Sally Chambers, Julie Birkholz and Fien Messens (2021, September 6-9). BESOCIAL: A Sustainable Knowledge Graph-based Workflow for Social Media Archiving [Conference Presentation]. SEMANTICS 2021 - 17th International Conference on Semantic Systems, online.
- Sven Lieber (2021, October 25). Harvesting Social Media Heritage. DH Early Career Discussion Group.
- Fien Messens (2022, January 19). Let's talk selection. IIPC Social Media Archiving Workshop.
- Friedel Geeraert and Fien Messens (2022, January 19). [co-host together with Zagreb University and IIPC]. IIPC Social Media Archiving Workshop.
- Sally Chambers (2022, March 24). Preservation of digitised and born-digital collections: interconnections, policies and workflows. Open Preservation Foundation, IMPACT and IIPC workshop.
- Lise-Anne Denis & Fien Messens (2022, March 24). Study day social media archiving organised by AAFB (Association des Archivistes Francophones de Belgique).
- Julie M. Birkholz, Friedel Geeraert, Isabelle Gribomont and Fien Messens (2022, April 25). Born-digital collections at KBR. DH lunch at the University of Texas at Austin, online.
- Fien Messens, Pieter Heyvaert, Eva Rolin, Patrick Watrin, Lise-Anne Denis and Peter Mechant (2022, May 24). BESOCIAL: social media archiving at KBR in Belgium. Panel at the IIPC Web Archiving Conference 2022, online - https://netpreserve.org/ga2022/wac/abstracts/#Session_2
- Lise-Anne Denis & Fien Messens (2022, June 3). To harvest or not to harvest? The importance of legal advice in BESOCIAL. Poster presentation at DH Benelux 2022, Luxembourg, online.
- Fien Messens & Friedel Geeraert (2022, June 23). Context BESOCIAL [presentation] Kennismakingssessie Nederlandse collega's, Rijksarchief Antwerpen.
- Fien Messens, Friedel Geeraert & Lise-Anne Denis (2022, August 31). Legal context BESOCIAL [presentation] Kennismakingssessie Nederlandse collega's, online.

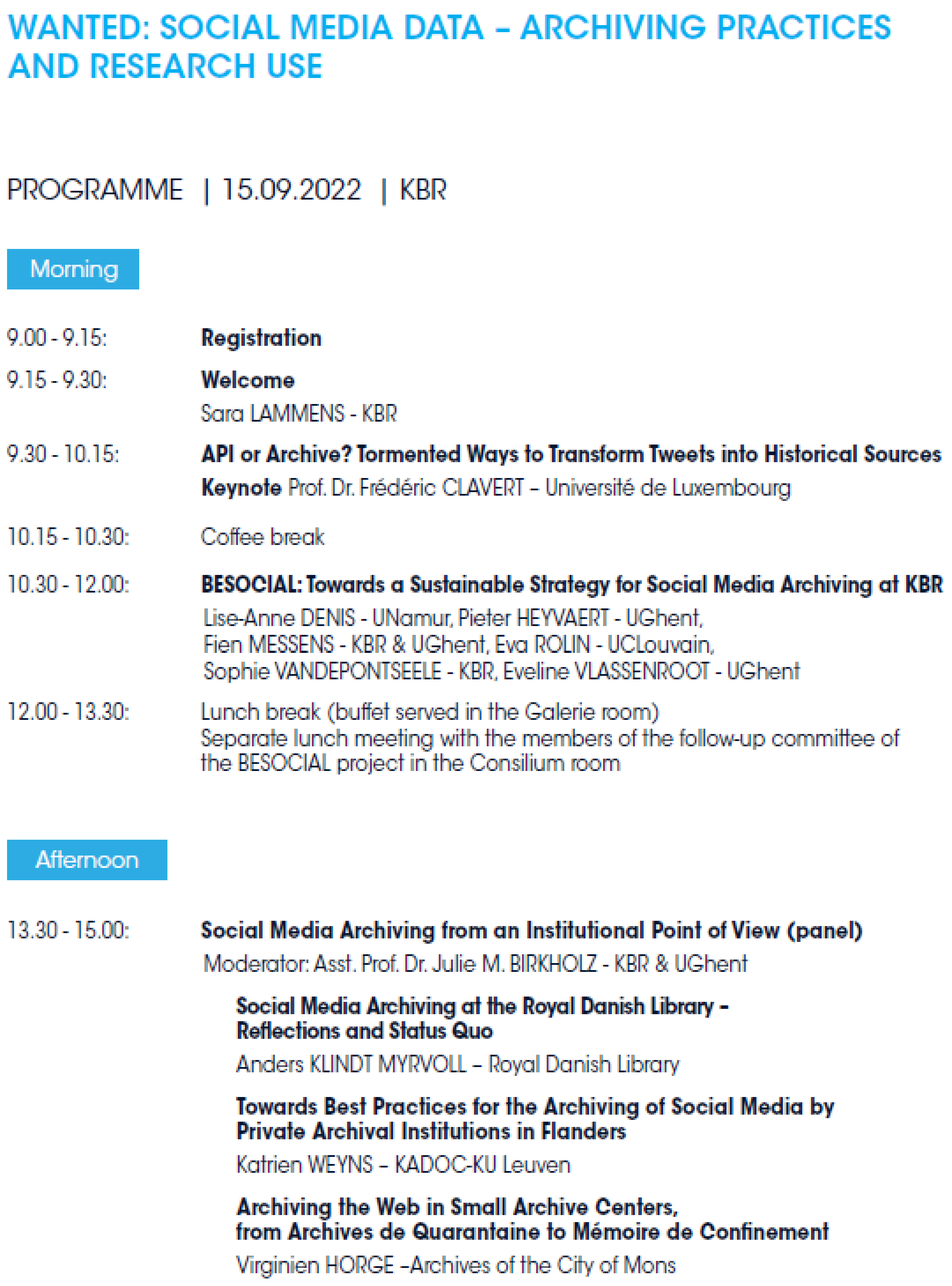
5.3 Press crowdsourcing campaign

- Stassart, Camille. 2022. "Tweets et Publications Instagram, Un Patrimoine Numérique à Part Entière." DAILY SCIENCE. January 11, 2022. https://dailyscience.be/11/01/2022/tweets-et-publications-instagram-un-patrimoine-numerique-a-part-entiere.
- Interview Radio Bruzz (November 2021)
- D.O.D. (2021, 29 oktober). KBR vraagt hulp voor archiveren sociale media. De Standaard;
- Van Huffel, Jozefien. (2021, 28 december). "Archief van Vluchtigheid." in Kerk & Leven. https://www.kerkenleven.be/uitgave/2152/artikel/archief-van-vluchtigheid.
- "Hoe Gaan We Binnen 200 Jaar Posts Op Sociale Media Terugvinden? KBR Wil Belgische Content Uit Sociale Media Archiveren." n.d. Radio 1. Accessed August 25, 2022. https://radio1.be/hoe-gaan-we-binnen-200-jaar-posts-op-sociale-media-terugvinden-kbr-wil-belgische-content-uit-sociale.

- Paelinck, Gianni. NWS, VRT. 2021. “#Stoofvlees, Rock Werchter of Een Tweet van Rik Torfs? Koninklijke Bibliotheek Wil Online Erfgoed Bewaren.” Vrtnws.be. October 28, 2021. https://www.vrt.be/vrtnws/nl/2021/10/28/online-erfgoed.
- Jaumotte, M. 2021. “Les Médias Sociaux Belges Archivés et Conservés Par La Bibliothèque Nationale.” RTBF. Accessed August 25, 2022. https://www.rtbf.be/article/les-medias-sociaux-belges-archives-et-conserves-par-la-bibliotheque-nationale-marion-jaumotte-10869305?id=10869305.
- “Réseaux Sociaux : Un Miroir de La Société à Préserver.” n.d. Athena. Accessed August 25, 2022. http://athena-magazine.be/magazine/le-magazine-n355/reseaux-sociaux-un-miroir-de-la-societe-a-preserver/.
- Corbeel, A. A. P. (2022, 17 februari). La KBR vous invite à sauver le patrimoine belge en ligne. RTBF. https://www.rtbf.be/article/la-kbr-vous-invite-a-sauver-le-patrimoine-belge-en-ligne-10937416.

5.4 Organisation of concluding colloquium: Wanted: social media data. Archiving practices and reuse on 15 September 2022

5.4.1 Programme



WANTED: SOCIAL MEDIA DATA – ARCHIVING PRACTICES AND RESEARCH USE

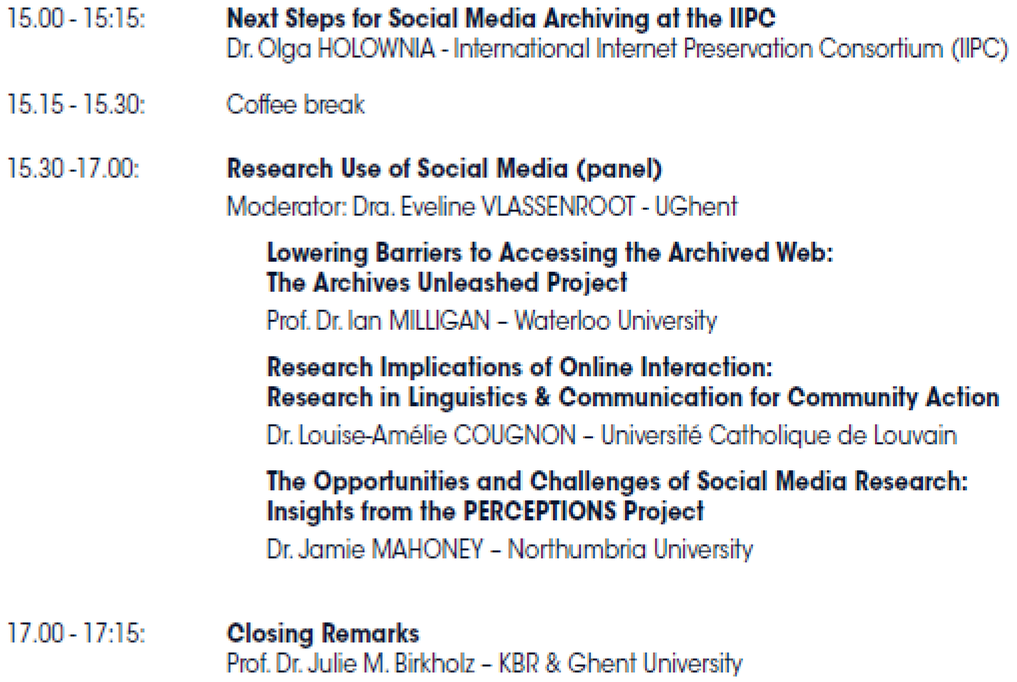
PROGRAMME | 15.09.2022 | KBR

Morning

9.00 - 9.15:	**Registration**
9.15 - 9.30:	**Welcome** Sara LAMMENS - KBR
9.30 - 10.15:	**API or Archive? Tormented Ways to Transform Tweets into Historical Sources** **Keynote** Prof. Dr. Frédéric CLAVERT – Université de Luxembourg
10.15 - 10.30:	Coffee break
10.30 - 12.00:	**BESOCIAL: Towards a Sustainable Strategy for Social Media Archiving at KBR** Lise-Anne DENIS - UNamur, Pieter HEYVAERT - UGhent, Fien MESSENS - KBR & UGhent, Eva ROLIN - UCLouvain, Sophie VANDEPONTSEELE - KBR, Eveline VLASSENROOT - UGhent
12.00 - 13.30:	Lunch break (buffet served in the Galerie room) Separate lunch meeting with the members of the follow-up committee of the BESOCIAL project in the Consilium room

Afternoon

13.30 - 15.00:	**Social Media Archiving from an Institutional Point of View (panel)** Moderator: Asst. Prof. Dr. Julie M. BIRKHOLZ - KBR & UGhent **Social Media Archiving at the Royal Danish Library - Reflections and Status Quo** Anders KLINDT MYRVOLL – Royal Danish Library **Towards Best Practices for the Archiving of Social Media by Private Archival Institutions in Flanders** Katrien WEYNS – KADOC-KU Leuven **Archiving the Web in Small Archive Centers, from Archives de Quarantaine to Mémoire de Confinement** Virginien HORGE –Archives of the City of Mons



15.00 - 15:15:	**Next Steps for Social Media Archiving at the IIPC** Dr. Olga HOLOWNIA - International Internet Preservation Consortium (IIPC)
15.15 - 15.30:	Coffee break
15.30 -17.00:	**Research Use of Social Media (panel)** Moderator: Dra. Eveline VLASSENROOT - UGhent **Lowering Barriers to Accessing the Archived Web: The Archives Unleashed Project** Prof. Dr. Ian MILLIGAN – Waterloo University **Research Implications of Online Interaction: Research in Linguistics & Communication for Community Action** Dr. Louise-Amélie COUGNON – Université Catholique de Louvain **The Opportunities and Challenges of Social Media Research: Insights from the PERCEPTIONS Project** Dr. Jamie MAHONEY – Northumbria University
17.00 - 17:15:	**Closing Remarks** Prof. Dr. Julie M. Birkholz – KBR & Ghent University

5.4.2 Short report on BESOCIAL colloquium

The Colloquium Wanted: social media data. Archiving practices and research use took place on 15 September 2022 at KBR (Royal Library of Belgium) in Brussels and online via Zoom Webinar.

Social media archiving is very important as everyone relies heavily on digital information in our daily lives. This has led to new challenges from the point of view of preservation since web content is particularly ephemeral. Social media is for many people the first source of news and entertainment However, how will researchers be able to study human behaviour in the beginning of the 21st century if this content is not preserved?

Cultural institutions such as national libraries have, over the last decades, started to archive and preserve this rapidly changing data type, by creating a framework for web and social media archiving. Assuring long-term availability of these born-digital collections is key to ensure their value to future generations of researchers. There are many threats to ensuring long-term availability of social media data, including technological, legal, financial, and organisational aspects, most of which were very familiar to the participants in the conference. In addition, using archived social media data for research poses many challenges.

The programme focused on two main themes: archiving practices of social media and research use of these born-digital collections. The first theme was explored through a presentation of the BESOCIAL team who highlighted the most important research results of their research project and during the panel on social media archiving from an institutional point of view. The research use of social media data was illustrated during the keynote and the presentations in the last panel of the day. In addition, the next steps for social media archiving at the IIPC, the International Internet Preservation Consortium were presented.

The keynote speaker Frederic Clavert presented the audience with a question: is social media archiving a new practice as our traditional practices of archiving are so different from what we need to do for social media data? Thus is the APIs that steer us and shape the ways in which we work or vice versa?

The audience was also introduced to three different examples of how institutions approach social media archiving: from a web archiving perspective at KBR (the Royal Library of Belgium), to links to the physical collections at KADOC – KU Leuven, and events from the Archives of the City of Mons. Then three different research projects showed very unique ways of using and investigating different trends on social media: The Archives Unleashed project at Waterloo University, the Université Catholique de Louvain and the PERCEPTIONS project at Northumbria University. The IIPC explained their role in social media archiving and the projects that they support that are focused on social media archiving.

6. PUBLICATIONS

Blog Posts

Pranger, J. (2020). The BESOCIAL project: towards a sustainable strategy for social media archiving in Belgium. Retrieved from https://netpreserveblog.wordpress.com/2020/09/23/the-besocial-project towards-a-sustainable-strategy-for-social-media-archiving-in-belgium/.

Geeraert, F., and Pranger, J. (2020). BESOCIAL. Retrieved from https://www.kbr.be/en/projects/besocial/.

(Peer-reviewed) Book Chapters

Michel, A. (2021). Web Archiving in the Public Interest from a Data Protection Perspective. in Larcier: Deep diving into data protection.

de Terwangne, C. & Michel, A. (2021). Processing of personal data for "journalistic purposes". in Larcier: Deep diving into data protection

Peer Reviewed Academic Articles

Vlassenroot, E. et al. (2021). Web-archiving and social media: an exploratory analysis. In International Journal of Digital Humanities (in press).

Messens, F (2022). Belgian social media archiving initiatives mapped. In In Monte Artium (awaiting final approval).

Peer Reviewed(Academic) Conferences Proceedings

Sven Lieber, Dylan Van Assche, Anastasia Dimou, Sally Chambers, Julie Birkholz and Fien Messens (2021). BESOCIAL: A Sustainable Knowledge Graph-based Workflow for Social Media Archiving. Will be presented at SEM21EU (SEMANTICS 2021 - 17th International Conference on Semantic Systems) September 2021.

Other

Messens, Fien; Vlassenroot, Eveline; Mechant, Peter; Watrin, Patrick; Rolin, Eva; Chambers, Sally; Birkholz, Julie M.;Geeraert, Friedel; Lieber, Sven; Michel, Alejandra, 2021, "BESOCIAL: Interview / Survey results WP1", https://doi.org/10.34934/DVN/RMKYKO, Social Sciences and Digital Humanities Archive – SODHA, V1.

Messens, Fien; Lieber, Sven; Chambers, Sally; Geeraert, Friedel, 2022, "Seed list mini pilot COVID-19 collection", https://doi.org/10.34934/DVN/SE8NUY, Social Sciences and Digital Humanities Archive – SODHA, V1

Messens, Fien, Birkholz, Julie. M, Chambers, Sally, Geeraert, Friedel, Michel, Alejandra, Mechant, Peter, Rolin, Eva. (2021). BESOCIAL: towards a sustainable strategy for archiving and preserving social media in Belgium. In DH Benelux 2021, Abstracts. Leiden, The Netherlands.

Michel, Alejandra, Pranger, Jessica, Geeraert, Friedel, Lieber, Sven, Mechant, Peter, Vlassenroot, Eveline, Chambers, Sally,Birkholz, Julie, & Messens, Fien. (2021). Towards a sustainable social media archiving strategy for Belgium: BESOCIAL: WP1 report: an international review of Social Media Archiving initiatives. s.n. https://orfeo.belnet.be/handle/internal/7741.

Alejandra Michel, Eva Rolin, Eveline Vlassenroot, Fien Messens, Friedel Geeraert, Julie Birkholz, Lise-Anne Denis, Patrick Watrin, Peter Mechant, Pieter Heyvaert, Sally Chambers, and Sven Lieber. (July 2022). BESOCIAL: Reports on WP2 Preparation of pilot for social media archiving and WP3 Pilot for access to social media archive. https://orfeo.belnet.be/handle/internal/10034.

Messens, Fien, & Denis, Lise-Anne. (2022). To harvest or not to harvest? The importance of legal advice in BESOCIAL. DH Benelux 2022 - ReMIX: Creation and alteration in DH (hybrid), Belval Campus, Esch-sur-Alzette, Luxembourg and online. Zenodo. https://doi.org/10.5281/zenodo.6572703.

Geeraert, Friedel & Messens, Fien. (2022). Country report: web and social media archiving in Belgium, Orfeo, https://orfeo.belnet.be/handle/internal/9980.

P. Mechant & E. Vlassenroot, BESOCIAL: Analysis of user requirements (Task 2.2), July 2021. Orfeo, https://orfeo.belnet.be/handle/internal/10011.

P. Mechant et al., BESOCIAL: Quality control of harvested content (Task 3.2), March 2022. Orfeo, https://orfeo.belnet.be/handle/internal/10012.

P. Mechant, Theys, T. & E. Vlassenroot, BESOCIAL: Evaluation of the Belgian pilot social media archive (Task 4.3), June 2022. Orfeo, https://orfeo.belnet.be/handle/internal/10035.

Lieber, Sven & Heyvaert, Pieter. (2022) BESOCIAL: User documentation of Social Feed Manager and Instaloader. Orfeo, https://orfeo.belnet.be/handle/internal/10013.

Denis, Lise-Anne. (2023). The importance of legal requirements for web archives studies in Belgian and French law. In WARCnet, 'Seeing the past and the present through web archives: A transnational perspective' [abstract awaiting final approval by publisher].

7. ACKNOWLEDGEMENTS

First of all, we would like to thank the members of the scientific follow-up committee for providing us with interesting ideas, feedback and guidance throughout the project.

- Walter Daelemans, University of Antwerp
- Frédéric Clavert, Université de Luxembourg
- Sara Aubry, Bibliothèque nationale de France
- Rolande Depoortere, Archives générales du Royaume et Archives de l'Etat dans les provinces
- Hannes Lowagie, KBR - Koninklijke Bibliotheek van België
- Silvia Sevilla, Publications Office of the European Union
- Herbert Van de Sompel, DANS (Data Archiving and Networked Services), Ghent University
- Anja Bechman, Aarhus University

We would also like to thank the representatives of social media archiving initiatives in Belgium and abroad whom we interviewed or to whom responded to questionnaires for establishing a state of the art. This was a very instructive phase for the entire research team and it also allowed us to maintain our network within the (inter)national web and social media archiving community:

- Jeroen Buysse, Liberas en Project Catching the Digital Heritage
- Katrien Weyns en Thomas Peeters, Best practices voor de archivering van sociale media in Vlaanderen en Brussel en KADOC-KULeuven
- Sarah Lessire, AAFB - Association des archivistes francophones de Belgique and project Quarantaine Archief
- Maarten Savels en Jeroen Fernandez-Alonso, Amsab – Instituut voor Sociale Geschiedenis
- Sophie Bossaert en Tom Cobbaert, ADVN | archief voor nationale bewegingen
- Johanne Ferket, Letterenhuis
- Patrick Hochstenbach, Universiteitsbibliotheek Gent
- Willem Vanneste, Archives of the University of Antwerp
- Roxanne Wyns, LIBIS - ICandid project
- Niels Bibert, Rozane Decock, Bieke Zaman, IMS - Instituut voor Media Studies (KU Leuven) / Mintlab
- Florian Daemen, AMVB - Archief en Museum voor het Vlaamse leven in Brussel

- Ellen Maria Soetens, CAVA (Centrum voor Academische en Vrijzinnige Archieven)
- Virginien Horge, Archives of the City of Mons
- Luc Wanlin, Archives et Musée de la Littérature
- Patricia Blanco, Archives de La Ville de Bruxelles
- Rony Vissers, Nastasia Vanderperren, and Ellen van Keer, meemoo
- Wim Lowet, Vlaams Architectuurinitiatief
- Tom Smyth, Library and Archives (LAC) - Canada
- Anders Klindt Myrvoll & Tue Hejskov Larsen, Royal Danish Library (Netarkivet) - Denmark
- Alexandre Chautemps & Sara Aubry, Bibliothèque nationale de France (BnF) - France
- Thomas Drugean, Claude Mussou & Jérôme Thiève, National Audiovisual Institute (INA) - France
- Ben Els & Yves Maurer, Bibliothèque nationale du Luxembourg (BnL) - Luxembourg
- Gilian Lee, Ben O'Brien, Valerie Love & Ronda Grantham, National Library - New Zealand
- Daniel Gomes, Arquivo.pt - Portugal
- Nicola Bingham, British Library - UK
- Tom Storrar, Claire Newing & Sarah Dietz, National Archives - UK

We would also like to thank Georges Jamart from BELSPO for his follow-up and guidance throughout the entire project.

The speakers who participated in the concluding colloquium "Wanted: Social Media Data – Archiving Practices and Research Use" also deserve a special thank you, especially since they are advocates for (the study of) social media archiving:

- Frédéric Clavert
- Sara Lammens
- Katrien Weyns
- Virginien Horge
- Anders Klindt Myrvoll
- Olga Holownia
- Ian Milligan
- Louise-Amélie Cougnon
- Jamie Mahoney

We would also like to thank several colleagues who were involved in the BESOCIAL project in one way or another:

- Isabelle Gribomont - KBR
- Brecht Deseure - KBR

- Thuy-An Pham - KBR
- Xavier Delor - KBR
- Nadège Isbergue - KBR
- Ann Van Camp - KBR
- Hannes Lowagie - KBR
- Sven Lieber - KBR
- Astrid De Spiegelaere - KBR
- Hanna Huysegoms - KBR
- Eglantine Lebacq - KBR
- Yves De Preter - KBR
- Tim Theys - UGhent
- Delia Budulan - CENTAL UCLouvain

We would also like to thank the interns from the advanced Master of Digital Humanities working on their own social media projects, with the guidance of the BESOCIAL team.

- Ewoud Goethals
- Luis Saez Jiménez
- Nisa İrem Kirbaç

REFERENCES

Bahnemann, G., Carroll, M., Clough, P., Einaudi, M., Ewing, C., Mixter, J., Roy, J., Tomren, H., Washburn, B., Williams, E. (2021): Transforming metadata into linked data to improve digital collection discoverability.

Bailey, J., Grotke, A., McCain, E., Moffatt, C., & Taylor, N. (2017). Web Archiving in the United States: A 2016 Survey. National Digital Stewardship Alliance.

Boyd, D. And K. Crawdford, 2019, Critical Questions for Big Data. Crawford. Information, Communication & Society 15 (5): 662–79; Lazer, D., A. Pentland, and L. Adamic, 2009, Review of Computational Social Science. Science 323 (5915): 721–23.

Brügger, N., Laursen, D., Nielsen, J. (2017) Exploring the domain names of the Danish web in The web as history, edited by Niels Brügger and Ralph Schroeder, UCL Press, London, 62-80.

Coppens, S., Verborgh, R., Peyrard, S., Ford, K., Creighton, T., Guenther, R., Mannens, E., Van de Walle, R. (2015) : PREMIS OWL. International Journal on Digital Libraries 15(2), 87–101.

Cadavid, J. A. P. (2017) Evolution of legal deposit in New Zealand: from print to digital heritage. International Federation of Library Associations and Institutions, 43(4), 379-390.

Candela, G., Sáez, M. D., Escobar, P., & Marco-Such, M. (2021). A benchmark of Spanish language datasets for computationally driven research. Journal of Information Science.

Costa, M., & Silva, M. J. (2010). Understanding the information needs of web archive users. Proc. of the 10th International Web Archiving Workshop, 9(16), 6.

Costea, M.-D. (2018). Report on the Scholarly Use of Web Archives. NetLab.

de Terwangne, C. & Michel, A. (2021). Processing of personal data for "journalistic purposes". in Larcier: Deep diving into data protection.

Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., Van de Sompel, H. (2010) The Europeana Data Model (EDM). In: World Library and Information Congress: 76th IFLA general conference and assembly. vol. 10, 15.

Dougherty, M. (2009). Historical infrastructures for web archiving, annotation of ephemeral collections for researchers and cultural heritage institutions.

Dougherty, M., & Meyer, E. T. (2014). Community, Tools, and Practices in Web Archiving: The State-of-the-Art in Relation to Social Science and Humanities Research Needs. Journal of the Association for Information Science and Technology, 65(11), 2195–2209.

Egense, T. (2021) SolrWayback 4.0 release! What's it all about? Part 2. Available online at: https://netpreserveblog.wordpress.com/tag/solrwayback/

Fernando, Z. T., Marenzi, I., & Nejdl, W. (2018). ArchiveWeb: collaboratively extending and exploring web archive collections—How would you like to work with your collections? International Journal on Digital Libraries, 19(1), 39–55.

Gebeil, S. (2016). Quand l'historien rencontre les archives du Web. Revue de La BNF, (2), 185–191.

Gerlitz, C., Helmond, A., van der Vlist, F. & Weltevrede, E. (2019). Regramming the Platform: Infrastructural Relations between Apps and Social Media. Computational Culture, 7. http://computationalculture.net/regramming-the-platform.

Hai Liang and Zhu Jonathan, "Big Data, Collection of (Social Media, Harvesting)," In The International Encyclopedia of Communication Research Methods (Hoboken: Wiley-Blackwell, 2017), 1-18.

Helmond, A. (2017). Historical website ecology. Analyzing past states of the web using archived source code. In N. Brügger (Ed.), Web25. Histories from the first 25 years of the world wide web. . New York: Peter Lang, 139-155.

Hockx-Yu. H. (2014). Archiving social media in the context of non-print legal deposit. Paper presented at IFLA, Lyon.

Hockx-Yu, H., (2014). Access and scholarly use of web archives. Alexandria. 25(1/2). http://dx.doi.org/10.7227/ALX.0023

Holzmann, H., Risse, T. (2017). Accessing web archives from different perspectives with potential synergies. 2nd International Conference on Web Archives / Web Archiving Week (RESAW/IIPC). https://doi.org/10.14296%2Fresaw.0001

Jackson, A., Lin, J., Milligan, I., & Ruest, N. (2016). Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities. IEEE/ACM Joint Conference on Digital Libraries (JCDL), 103–106. https://doi.org/10.1145/2910896.2910912

Lasfargues, F., Oury, C., Wendland, B. (2008). Legal deposit of the French web: harvesting strategies for a national domain. International Web Archiving Workshop, Sep 2008, Aarhus, Denmark.

Lauridsen, J. (2021) SolrWayback 4.0 release! What's it all about? Available online at: https://netpreserveblog.wordpress.com/tag/solrwayback/

Lee, G., Love, V., Moran, J. (2019). Archiving social media at the Alexander Turnbull Library, Te Puna Mātauranga o Aotearoa National Library of New Zealand. Preservation, Digital Technology & Culture, 48(3), 129-134. https://doi.org/10.1515/pdtc-2019-0017.

Lieber, S., Van Assche, D., Chambers, S., Messens, F., Geeraert, F., Birkholz, J. M., & Dimou, A. (2021). BESOCIAL : a sustainable knowledge graph-based workflow for social media archiving. In M. Alam, P. Groth, V. de Boer, T. Pellegrini, H. J. Pandit, E. Montiel, … A. Meroño-Peñuela (Eds.), Further with knowledge graphs : proceedings of the 17th international Conference on Semantic Systems (Vol. 53, pp. 198–212). Amsterdam, the Netherlands: IOS. https://doi.org/10.3233/ssw210045.

Lieber, Sven, 2021, "BESOCIAL: Social media archiving tools comparison", https://doi.org/10.34934/DVN/U0VFYH, Social Sciences and Digital Humanities Archive – SODHA, V1.

Littman, J., Chudnov, D., Kerchner, D., Peterson, C., Tan, Y., Trent, R., … Wrubel, L. (2018). API-based social media collecting as a form of web archiving. International Journal on Digital Libraries, 19(1), 21–38.

Macnaught, B. (2018). Social media collecting at the National Library of New Zealand. Paper presented at IFLA WLIC, Kuala Lumpur. Michel, A., (2021) "Web Archiving in the Public Interest from a Data Protection Perspective", in Deep Diving into Data Protection - 1979-2019: Celebrating 40 Years of Privacy and Data Protection at the CRIDS (ed. J. Herveg), coll. du CRIDS, Bruxelles, Larcier, 181-200 Michel, A. « Tour d'horizon sur les aspects légaux de l'archivage du web », Cahiers de la Documentation, 2020/2, 57.

McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., & McCullough D. (2012). On building a reusable Twitter corpus. In Hersh, W., SIGIR '12: Proceedings of the 35th int. ACM SIGIR conference on Research and development in information retrieval (pp. 1113-1114), New York: Association for Computing Machinery.

Michel, A. (2021). Web Archiving in the Public Interest from a Data Protection Perspective. in Larcier: Deep diving into data protection.

Milligan, I., (2019). Exploring web archives in the age of abundance: a social history case study of GeoCities. The SAGE handbook of web history, 344-358.

Ogden, J., & Maemura, E. (2021). 'Go fish': Conceptualising the challenges of engaging national web archives for digital research. International Journal of Digital Humanities, 1–21. https://doi.org/10.1145/3383583.3398513

Riley, H., Crookston, M., & Library, A. T. (2015). Awareness and Use of the New Zealand Web Archive A survey of New Zealand academics.

Ruest, N., Lin, J., Milligan, I., & Fritz, S. (2020). The archives unleashed project: technology, process, and community to improve scholarly access to web archives. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, 157–166.

Ruest, N., Fritz, S., Deschamps, R., Lin, J., Milligan, I. (2021). From archive to analysis: accessing web archives at scale through a cloud-based interface. International Journal of Digital Humanities. https://doi.org/10.1007/s42803-020-00029-6

Ruest, N., Lin, J., Milligan, I., Fritz, S. (2020). The Archives Unleashed Project: technology, process, and community to improve scholarly access to web archives. JCDL 2020: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. pp. 157–166. Schneider, S. M., & Foot, K. A. (2004). The web as an object of study. New Media & Society, 6(1), 114–122.

Schostag, S. & Fønss-Jørgensen, E. (2012). Webarchiving: legal deposit of Internet in Denmark. A curatorial perspective. MDR, 41, 110-120. Smith, C., & Cooke, I. (2017). Emerging Formats: Complex digital media and its impact on the UK Legal Deposit Libraries. Alexandria, 27(3), 175–187. https://doi.org/10.1177/0955749018775878.

Stirling, P., Chevallier, P., & Illien, G. (2012). Web archives for researchers: Representations, expectations and potential uses. D-Lib Magazine, 18(3/4). Retrieved from http://www.dlib.org/dlib/march12/stirling/03stirling.html.

Tufekci, Zeynep, n.d, Review of Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In Proceedings of the 8th International Conference on Weblogs and Social Media.

Venlet, J., Farrell, K. S., Kim, T., O'Dell, A. J., & Dooley, J. (2018). Descriptive metadata for web archiving: literature review of user needs.

Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., Mechant, P. (2019) Web archives as a data resource for digital scholars. International Journal of Digital Humanities. 1. https://biblio.ugent.be/publication/8606713

Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J. & Mechant, P. (2021). Web-archiving and social media: an exploratory analysis. International Journal of Digital Humanities, 2, 107-128.

Webster, P. 'Users, technologies, organisations: Towards a cultural history of world web archiving' in Web 25. Histories from 25 years of the World Wide Web, edited by Niels Brügger, Peter Lang, New York, 2017,. 179-190.

Winters, J. (2017). Breaking in to the mainstream: demonstrating the value of internet (and web) histories. Internet Histories, 1(1–2), 173–179. https://doi.org/10.1080/24701475.2017.1305713

Zimmer, M. (2015). The twitter archive at the Library of Congress: Challenges for information practice and information policy. First Monday, 20(7).