Towards a sustainable social media archiving strategy for Belgium

---

Report on WP2 Preparation of pilot for social media archiving and WP3 Pilot for access to social media archive

([M4]: [M15])

Editors	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
Responsible partners	Cental (UCLouvain) CRIDS (UNamur) KBR UGhent (MICT, GhentCDH, IDLab)
Version	1.0
How to cite this?	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.*

Table of Content

Why combine working packages 2 and 3?

This report aggregates the results from Work Package 2 (WP2) and Work Package 3 (WP3) of the BELSPO funded, KBR coordinated BESOCIAL project.

- The aim of WP2 was to develop a strategy before the start of the real pilot in the third work package. It's the preparation phase before the start of WP3. Here, among other things, several tools and their functionalities were tested. What is possible, what is feasible, what are the needs of researchers when using a social media archive to conduct their research, what works for this project and what does not, how we will select the hashtags and accounts etc. All this with a view to the main pilot.
- The aim of WP3 was to develop a social media harvester and harvest the entire seedlist created in WP2. After this, the quality of the data content was checked. A preservation plan was also drawn up for archiving social media data in Belgium.

In this report, the two work packages are combined to build a bridge between the preparation and execution phases. This is in order to anticipate and prevent potential problems. The preparatory phase within BESOCIAL takes on an equally important role as execution.

WP2 Report: Preparation of pilot for social media archiving

1 Introduction

This first part of the report, 'Preparation for pilot for social media archiving' aggregates the results from Work Package 2 (WP2) of the BELSPO funded, KBR coordinated BESOCIAL project. The aim of WP2 was to prepare for the main pilot that was executed in Work Package 3 (WP3).

In this work package, four dedicated tasks aimed to provide a concise preparation for the main pilot. *Task 2.1 (Development of methodology for selection of social media)* aimed to develop a sustainable methodology for the selection of Social Media content in real-time.
Task 2.2. (Analysis of user requirements) took the needs of a broad range of stakeholders (researchers, cultural heritage professionals, publishers, policy-makers and other potential end users) into account when designing the social media archive.
Task 2.3. (Analysis of existing legal frameworks in Belgium for selection of content for social media archiving) where we will deepen the study of the following legal requirements: 1) all matters related to privacy law (especially, the question of the private or public status of content posted on social media, including identifying the relevant Belgian case law); 2) image right and "e-reputation" right;
3) the link between archives of online media newspapers and fundamental rights to information, to freedom of expression and the freedom of press; 4) the status, the role and the management of fake news in the selection of social media contents; and; 5) copyright issues. This detailed review and analysis of the applicable legal framework for content selection is a preliminary and necessary step for the final legal recommendations concerning SMA.
Task 2.4. (Definition of the technical and functional requirements based on the OAIS model) aims to identify and describe the technical and functional requirements for the pilots (WP3 Pilot for SMA and WP4 Pilot for access to social media archive). The requirements will be structured according to the core functions of the OAIS model with the addition of two functions: selection and collecting. The details of these tasks are detailed here in this report.

This report is structured as follows. **Section 2** discusses the mini-pilots also known as feasibility studies. **Section 3** discusses the selection strategy. Here a distinction is made between the selection criteria and the ongoing crowdsourcing campaign. **Section 4** discusses the analysis of user requirements. The goal here is to understand the needs and requirements of a wide range of stakeholders when using a social media archive. **Section 5** focuses on the analysis of the existing legal frameworks in Belgium for the selection of content for social media archiving. **Section 6** discusses the technical and functional requirements in detail. **Section 7** discusses the feedback we received from the follow-up committee members with the focus mainly on WP2. The findings of these different tasks are summarised in the discussion and conclusion at the end.

2 Mini-pilots

"The feasibility study is a procedure to predict the outcome of an investigation, examination, or assessment of a planned scheme along with possible gain."[1]

Within the BESOCIAL research project we started WP2 with executing a few mini-pilots , an important stage with the main goal to identify potential problems and deficiencies, and prevent them in the main pilot. This step can also help the research team to become familiar with the procedures, and can help them decide between competing methods, tools, and selection strategies. It helps decision-making in the project. Furthermore it improves time management as it allows for factoring in several restrictions. At the end of the 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 potential pitfalls.

In this section the following test cases are discussed:

- **Two technical feasibility studies** where the technical experts within the BESOCIAL project study the technical aspects needed. Based on these results it is decided what tools we will use and which social media we will harvest, preserve and archive.
    - Conducting a test crawl in order to make a choice between two harvesting tools, as well as testing the user interface of the different tools. In this mini-pilot the data is harvested from the social media platform Twitter. The chosen theme of the corpus is based around the Covid-19 crisis in Belgium.
    - Conducting test crawls of media urls using the command line tool Twarc.
- Another mini-pilot focuses on the **lay-out of the social media data**; how posts and hashtags look in the social media platform. Due to legal reasons it is not possible to harvest the look and feel of posts and accounts. Because of these reasons, we made the choice to harvest the 'look and feel' of KBR's own social media platforms (Twitter, Instagram and Facebook) as a case-study.

This technical part of the mini-pilot was executed by IDLab (UGhent) and CENTAL (UCLouvain). The other partners (KBR, GhentCDH, mict, CRIDS) provide support in terms of selection and subsequent analysis. The mini-pilot of the look and feel was managed by the BESOCIAL researcher at KBR.

A table indicating the purposes and decisions made during these feasibility studies is included at the end of this section.

[1] Momin Mukherjee and Sahadev Roy, "Feasibility Studies and Important Aspects of Project Management," *International Journal of Advanced Engineering and Management,* Vol. 2, No. 4, pp. 98-100, 2017.

2.1 Mini-pilot 1: What harvesting tool is the most efficient for harvesting social media collections?

We compared several social media harvesting tools within working package 1 on which we reported detailed in the WP1 report[2] and as part of a scientific publication (Lieber, 2021). Based on this comparison we decided to use the open source tool Social Feed Manager (SFM)[3]. However, one of the technical partners, the Centre for Natural Language Processing (CENTAL) at UC Louvain, also has a social media harvester in-house. Therefore, in this mini pilot we selected a seed list, configured both tools to harvest these seeds on a daily basis and compared the outcome. In section 2.1.1 we present an overview of the seed selection and harvesting configuration. In section 2.1.2 we present the results of both harvesting tools. In section 2.1.3 we discuss the results and finally conclude in section 2.4.

2.1.1 Overview

The selection of the content for this mini-pilot is twofold, on the one hand **accounts related to Belgium**, and on the other hand **hashtags related to the Covid pandemic**. Whereas the accounts align to the objective of the BESOCIAL project (public Belgian content), the hashtags relate to current events. The latter allowed us to test harvesting dynamic content.

During the selection of accounts and hashtags, the choice was made to keep linguistic consistency. This means that approximately the same number of resources were captured in French, Dutch and English.[4]

Although the focus of this mini-pilot was mainly on analyzing the outcome of the technical aspects, it was chosen to follow a dual selection strategy.

- On the one hand, certain accounts were archived and analysed (e.g. journalists, medical institutions, and virologists).[5]
- On the other hand, Belgian hashtags linked to the chosen theme were selected and collected in real-time.[6]

The following 10 French, 10 Dutch, 3 English and 1 German hashtags were collected:

- French: #Covid19Be, #coronavirusbelgique, #confinementbelgique, #coronavirusBE, #confinementbe, #covidbelgique, #confinementbruxelles, #restezchezvous, #coiffeuràpoil, #alysson, #AlyssonJadin, #JeSuisAlysson

[2] Michel, A. , Pranger, J., et al. WP1 report: an international review of Social Media Archiving initiatives. 2021. https://orfeo.kbr.be/handle/internal/7741

[3] George Washington University Libraries. (2016). Social Feed Manager. Zenodo. https://doi.org/10.5281/zenodo.597278

[4] One exception was the English and German content, because we could not identify enough hashtags solely for Belgium as we could for example for French or Dutch content.

[5] 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.

[6] 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.

- Dutch: #beternacorona, #ikbensolidair, #hetkananders, #zorgvoorelkaar, #degoedekantop, #samentegencorona, #zorgvoorlicht, #blijfthuis, #ditismijnzorg, #Covid19Belgie, #covidbelgie, #blijfinjekot
- English: #Covid19Belgium, #covidbelgium, #StayHomeBelgium
- German: #Ostbelgien

Data retrieval	In this mini pilot the data is harvested from the social media platform **Twitter**. The planned output for this sort of web crawling is Warc and/or Json. The goal of this data retrieval is dual: ● Test tool for data harvesting ● Test User Interface (UI) tool **Possible concerns** for the Twitter API: data limits, data harvesting needs to be done on separate IP addresses, and the choice of credentials. **Current number of tweets (as of 2021-01-26)** ● SFM twitter accounts collection: 895 ● SFM twitter hashtags collection: 155,501
Tools	● Social Feed Manager (SFM) ● Internal tool of CENTAL
Data Management	**Where was the mini pilot run on?** CENTAL set up a virtual machine for IDLAB so everything can be run on UCL servers; IDLAB specified virtual machines on Ubuntu. **Where is the data stored?** CLARIAH data storage on the HPC data cloud
Timing	Start date and time of the harvest: ● SFM account harvest: 2021-01-25 daily around 8:12 pm ● SFM hashtag harvest: 2021-01-26 daily at 9:05 am End date and time of the harvest: ● SFM account harvest: 2021-11-30 ● SFM hashtag harvest: 2021-11-30

Table 1: Summary of the covid mini-pilot

The version of Social Feed Manager available at the time was version 2.3 and did not fully meet the needs for the BESOCIAL project. In particular, the division between harvesting from an IDLab server and data storage on a CENTAL server did not work out of the box. Therefore, a local fork of the SFM code was created using GitHub functionality and the source code was adapted to support a more fine-grained data storage approach. Our changes improved SFM and were adopted for version 2.4 in July 2021.

2.1.2 Results, and Quality Control

The mini pilot resulted in a test crawl of two corpora from two different tools (CENTAL's tool and Social Feed Manager), harvesting a list of accounts, and hashtags that were related to COVID. The harvesting organised by CENTAL and IDLAB started on 25 January 2021 and is currently still harvesting. An analysis was done for content harvested until 16 February. 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 2: Number of tweets collected per seed list category, numbers from the harvesting period 2021-01-25 to 2021-02-16.

In an update meeting it was decided to **continue with Social Feed Manager as the harvesting tool**. At the point of the analysis, we had 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. Social Feed Manager preserves harvesting provenance by wrapping API requests in WARC files, thus for the account collection we have a daily WARC file. Regarding the size of the data, the account collection comprised 64 MB and the hashtag collection around 6 MB. We also extracted JSON files for the analysis comprising 729 MB for the account collection and 58 MB for the hashtag collection.

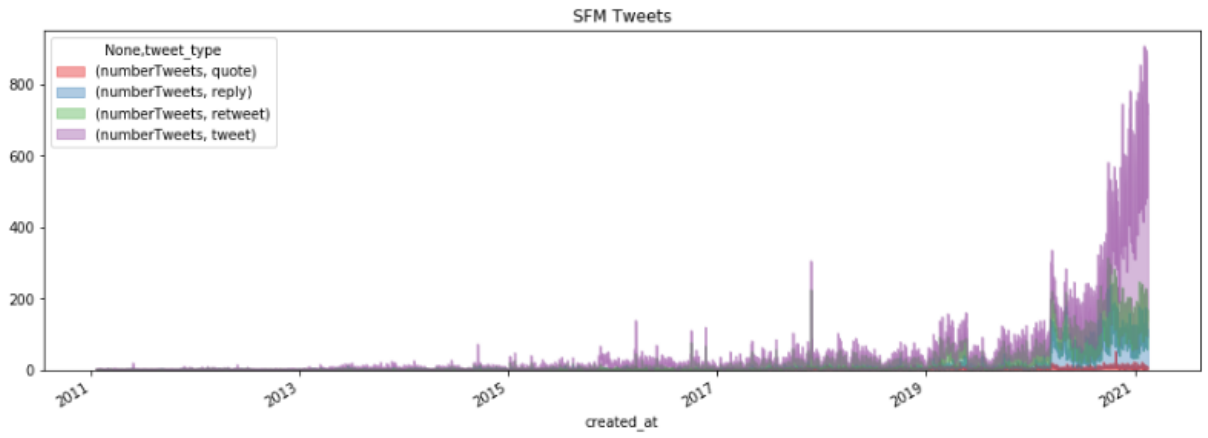

Figure 1: the temporal distribution of harvested account collection tweets.

Figure 1 shows the temporal distribution of different types of tweets from the account collection. A few tweets date back to 2011 because of the 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.

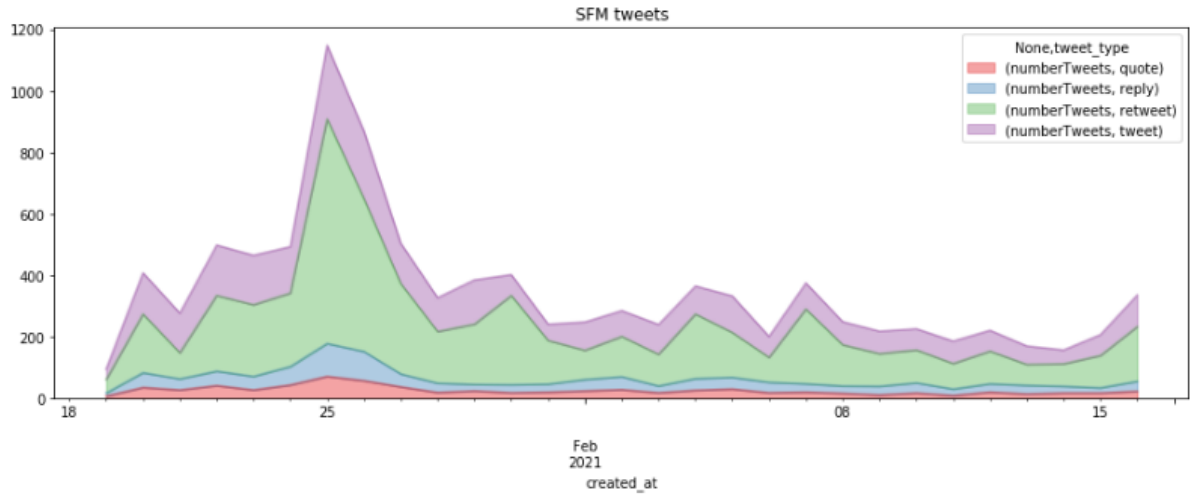

Figure 2: the temporal distribution of harvested hashtag collection tweets.

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

2.1.3 Discussion, Questions, and Lessons learned

Tool: The choice has been made to use Social Feed Manager because it proved to be more user friendly and reliable for this type of data than the tool provided by CENTAL.
Selection: at the request of the BESOCIAL team a selection policy was created by KBR to include the vision and mission of the institution in the selection process. A more elaborated seed list was constructed in a later phase of the project for 1) events and 2) accounts and hashtags linked to a specific event.
Look and feel: A request was made during an update meeting for a second mini-pilot (see section 2.3) if the focus could shift from the technical side to the look and the feel of the harvested data.
Data storage: CENTAL
Ran on: server of IDLab, but the main harvesting activity was done on premises of CENTAL
Platforms: Twitter and Instagram since it is less costly to archive and more reliable than Facebook. Twitter crawls produce better results and are mostly limited by time constraints for harvests, meaning that we might miss information with daily crawls. Twitter also follows a different policy regarding web archiving, where it is actually encouraged and several methods are made available for that purpose.

2.1.4 Conclusion

During the first mini-pilot we were able to limit down the tools choice: Social Feed Manager will be the chosen crawler tool for all (Twitter) data in the BESOCIAL project. With regards to the selection strategy, the selection for this test case was made organically. In the future a selection policy strategy will be created from the perspective of the KBR to limit the amount of data and to expand the representativeness of the data and/or the chosen current event.

2.2 Mini-pilot 2: How can we harvest images?

Multimedia content in the form of videos were explicitly excluded in the scope of BESOCIAL. However, both Twitter and especially Instagram also contain pictures as important content of social media posts. Within this pilot we investigated the capabilities of harvesting images from both Twitter and Instagram. An image harvesting functionality is not yet integrated in Social Feed Manager (hereafter SFM), therefore we used the functionality of the underlying **Twarc[7] harvesting tool**. For Instagram we relizs on the tool **Instaloader[8]** which we tested already in work package 1. In section 2.2.1 we introduce the setup of the harvests. In section 2.2.2 we present results, and in section 2.2.3 we discuss lessons learned and the conclusion.

2.2.1 Harvesting setup

For Twitter we used already harvested content from the account collection as a source from which we harvest associated images. For Instagram we used a seed list of a literary research use case containing the Instagram accounts of several literature events and bookshops in Belgium.

2.2.1.1 Twitter harvest

From a technical perspective the harvesting of images from Twitter is relatively straightforward: image URLs within the JSON representation of already harvested tweets are extracted and simply downloaded. The Twitter harvester Twarc (which is used internally by SFM) provides several utility scripts, one of it is called media2warc.py[9]. This utility script extracts media URLs from Twitter JSON, downloads the media from the URLs and records this harvesting in a WARC file. Therefore, extracted images are stored in a WARC File. However, since with SFM the harvested JSON content is compressed, stored within a WARC file which itself is compressed again, the harvesting of images requires a multi-step process.

We used a functionality provided by SFM to export harvested content and called a harvesting script in a separate step to download the media content. Within SFM we navigated to the collection which should be exported, then we selected export. In the export dialog we selected JSON as output format as this is the required input format for the download script.

The utility script media2warc from the Twarc project needs to be adapted for the needs of BESOCIAL: we only want images within tweets excluding videos and profile pictures, and additionally at this

[7] https://github.com/DocNow/twarc
[8] https://instaloader.github.io/
[9] https://github.com/DocNow/twarc/blob/main/utils/media2warc.py

stage, we do not want harvested images to be stored in WARC Files to facilitate their use by our access platform. We adapted the script accordingly.[10]

2.2.1.2 Instagram harvest

We created a script to harvest Instagram content using Instaloader, because Instaloader does not have a scheduling feature for harvests such as SFM. Our script does the following:

- Make it easier to harvest a list of accounts and hashtags
- Make it possible to have pauses between harvesting of different accounts and hashtags

Users provide a list of accounts and hashtags to the script. Every element in the list is harvested sequentially as they appear in the list. Harvesting data from Instagram is difficult due to the restrictions on the amount of content that you can harvest. Users can instruct the script to introduce a pause between the harvesting of every item to work around the restrictions imposed by Instagram.

2.2.2 Results

For Twitter images, the adapted download script detected 63,222 media URLs in total from which 62,803 are unique and 419 are duplicates. The images were downloaded and are preserved in a 7.6 GB WARC file. The script was executed using 2 threads, but because of some technical issues, the script did not stop. Thus, the measurement of the execution time using the “time” command did not work. Based on the log entry at the start of the script and the last modification date of the WARC file we estimate a runtime of 5.5 hours.

For the Instagram harvest, the seed list contained 1063 accounts and hashtags. It resulted in our account and IP address being temporarily blocked. Once around 200 posts are downloaded, we have to wait around 20 minutes before we can start harvesting again using Instaloader. But we experienced that sometimes 200 posts is still too many or that 20 minutes is not enough. We had to experiment with different pauses between different harvests to get at least some content for most accounts and hashtags. At the time of writing we have harvested the contents of 9488 posts. This includes both the images and the captions.

2.2.3 Discussion and Conclusion

For Twitter, harvesting images from tweets is not done directly in SFM when tweets are harvested, but are harvested afterwards in a separate step. There are two approaches to harvesting these images:

- The user harvests the images after each collection is updated in SFM.
- The user harvests the images only at the moment they are needed.

The first approach has 2 advantages:

- The chances are higher that the images are still available.

[10] https://github.com/SvenLieber/twarc/blob/f691a759912b171b76270a939a3b15816c9f67bf/utils/media2warc.py

- You reduce the chance of being blocked/stopped by Twitter for downloading too many images at once.

This approach has the disadvantage that you need to harvest the images every time manually, as there might not be a straightforward way to integrate this directly with the harvesting done by SFM. The second approach has the advantage that you will have to harvest less, because you only do it when it is needed. The disadvantage is that you might have a higher chance of being blocked/stopped by Twitter for downloading too many images at once. We conclude that for Twitter there is definitely a workable approach to harvest the images, but it would be more user-friendly if it was directly integrated in SFM to keep your archive of posts up to date. Integrating pauses between different harvests lowers the chances of being blocked by Instagram, but it is not full-proof approach way because

- we do not have a full understanding of how Instagram determines when to block an account or IP address, and
- Instagram can at any time change how they determine when to block an account or IP address.

Instaloader is very useful for harvesting Instagram content and together with our script we can harvest a (limited) seed list. However, until Instagram provides an official API, harvesting content will be a challenging task.

2.3 Mini-pilot 3: How can we harvest the look and feel of social media collections?

The Mini-Pilot to test capturing the look and feel was focused on the social media channels of KBR. Almost every day they post photos, videos and text messages. For this exercise, we chose to harvest data from their Facebook, Instagram and Twitter accounts. Although KBR also has a YouTube channel, the BESOCIAL team chose to exclude video content.

The purpose of harvesting the layout of these three platforms is to preserve a sample of what the layout looked like today (anno summer 2021). Due to legal and technical reasons one cannot harvest the look and feel of any post. We have chosen to use KBR's own Social Media platforms as a case study because we could ask KBR's consent for harvesting. Facebook, among others, renews its page set-up at least annually. We aim to preserve how posts and hashtags are placed and layed out on a social media page. Over time this preserved information can lead to a better understanding of the evolution of layouts and the way in which information was presented.

For the harvesting of this type of data we used Webrecorder/Conifer. Webrecorder provides a suite of open-source projects and tools to capture interactive websites and play them back as accurately as possible at a later date. The start date for the crawl differed between social media platforms. Webrecorder/Conifer was used for all platforms. Below is a brief description of how the look and feel can be harvested through the Webrecorder tool.

SM Platform	Harvest tool	Subscribe date KBR	Period of harvest	Output	Note
Instagram (currently 288 posts)	Webrecorder /Conifer	Apr 2019	22/07/21 - 04/04/2019	WARC file (712 MB)	Autopilot shut down after harvesting around 50 percent of the data Fully harvested
Twitter (currently 3265 tweets)	Webrecorder /Conifer	Feb 2009	29/07/21 - 27/06/19	WARC file (197 MB)	Autopilot shut down after harvesting around 50 percent of the data
Facebook	Webrecorder /Conifer	Feb 2012	NA	WARC file	Autopilot failed after 10 seconds

Table 3: Overview of features used during the look and feel mini-pilot

2.3.1 How can we harvest the look and feel of KBR's social media platforms?

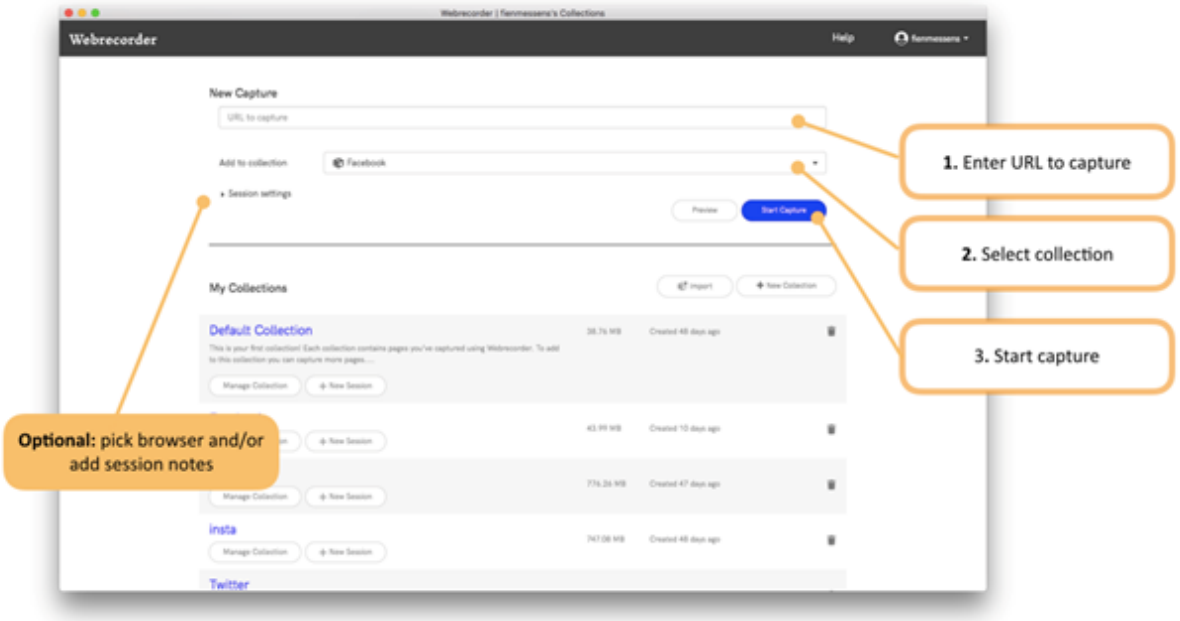

Figure 3. The preparation phase before the start of the capture in Webrecorder/Conifer

First one has to download the Webrecorder/Conifer tool. We recommend using Desktop mode to diminish loading problems. After installation one must create a personal account in order to keep your harvests saved in different folders, and to not lose a session. In Figure 3 one can start with entering the URL to capture (1). Then you can select a collection to store your harvesting session (2). During this step one can easily create new collections. It is also possible to pick your desirable

browser (e.g. Chrome) and to add session notes. In step (3) one can start the capture and you will be directed to another page.

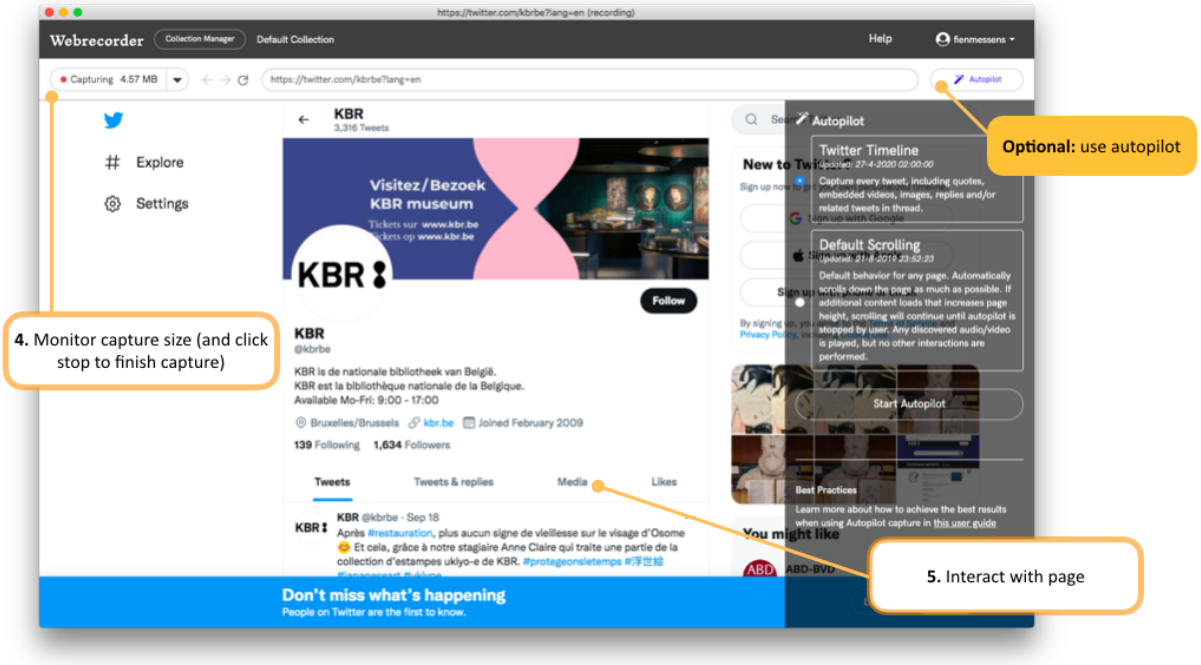

Figure 4. The capture page in Webrecorder/Conifer

As seen in Figure 4 the next step is to interact with the chosen URL to capture (5). It is possible to click for example on 'Tweets & Replies' and you'll be directed within the Webrecorder environment to that specific page. The content seen will also be harvested. During this capturing process you can always monitor (4) the size of what you are capturing (the red button will be flickering if the program is still collecting data). At any moment it is possible to click on this button to stop the capture. An extra asset on this page is to choose to use the autopilot. In this example you have 2 options: Default scrolling and Twitter timeline. With this last option one can capture every tweet, including quotes embedded videos, images, replies and/or related tweets in the thread. For this mini-pilot we chose this type of autopilot. It is recommended to start the autopilot at the beginning of the capture process, as soon as the page is loaded. This in order to diminish loading problems.

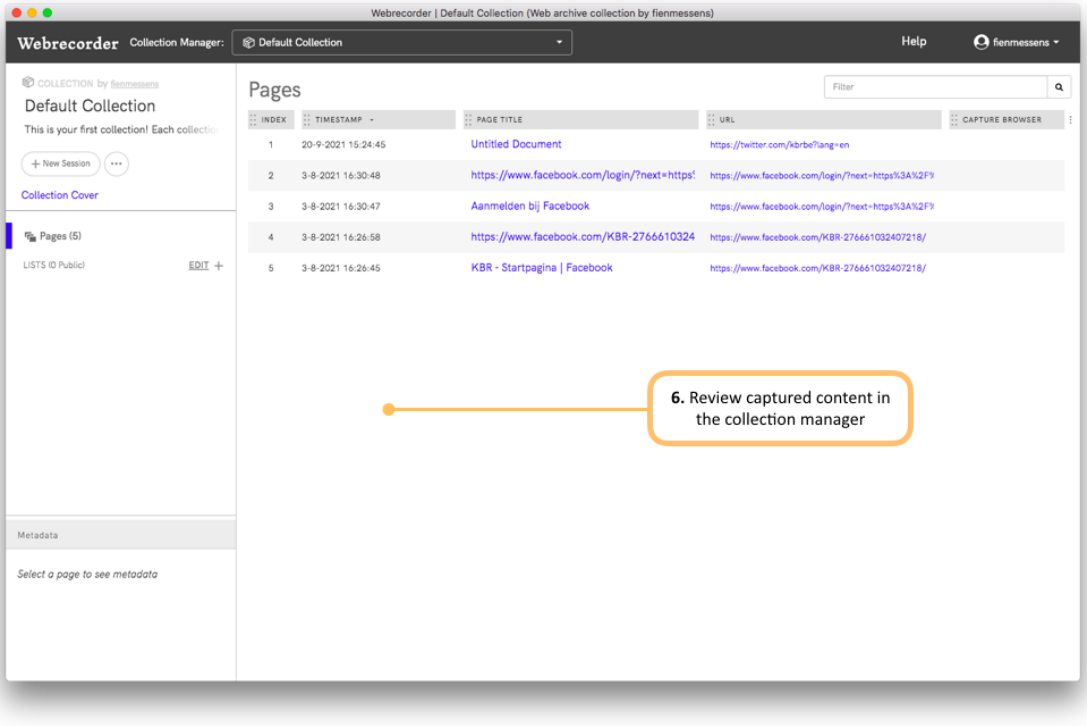

Figure 5. The collection manager in Webrecorder Conifer

The next step once collecting all the desired content is to manage your data. In step 6 one can review the captured content in the collection manager. You can export your data in a WARC-file. To replay your data one should download the replay tool of Webrecorder called Webrecorder Player. Webrecorder Player is a desktop application for viewing high-fidelity web archives on your local machine, no internet connection required. Webrecorder Player is available for download for OSX, Windows and Linux (experimental).

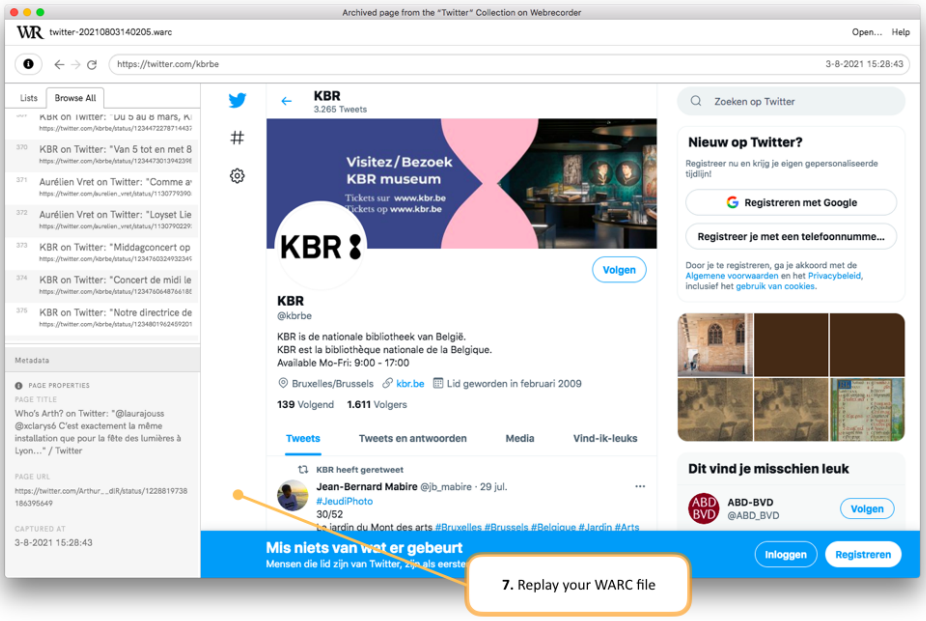

Figure 6. The Webrecorder Player tool

Here as well it is recommended to use the Desktop mode. It is possible to upload a WARC-file in order to check the quality of the harvest and scroll down.

2.3.2 Possible bugs using the autopilot

The autopilot of Webrecorder/Conifer can perform actions on the current web page loaded in Conifer, similar to a human user. It can be activated via the Autopilot button on the top right during capture (see Figure 5). When the button is solid green, *specialized behavior* is available for the current web page (for Twitter and Instagram). A white button indicates that only the *default behavior* will be presented instead. The default behavior tries to perform generally useful actions: scrolling down and triggering embedded media to play. While harvesting KBR's social media, the autopilot worked for a certain period of time. For Instagram and Twitter, it managed to bring in about 50% of the data automatically via the autopilot. The other half was added to the collection by manual scrolling. Unfortunately, the default scrolling button only worked for 10 seconds for the Facebook platform. Several attempts have been made to harvest data via another tool called Heritrix, but unfortunately without success. The choice was then made to go back to using the Webrecorder/Conifer tool with the goal of bringing in as much data as possible manually.

For Instagram, Facebook and Twitter getting “all” content was not possible, especially when large amounts of items were presented via infinite scrolling. Posts may not be captured fully for a variety of reasons, such as rate limiting, unexpected network issues, and so forth. It should also be taken into account that social media platforms can be expected to be frequently redesigned or technically re-architecture. When such an update happens, autopilot behaviors probably will not continue working as planned and need to be adapted.

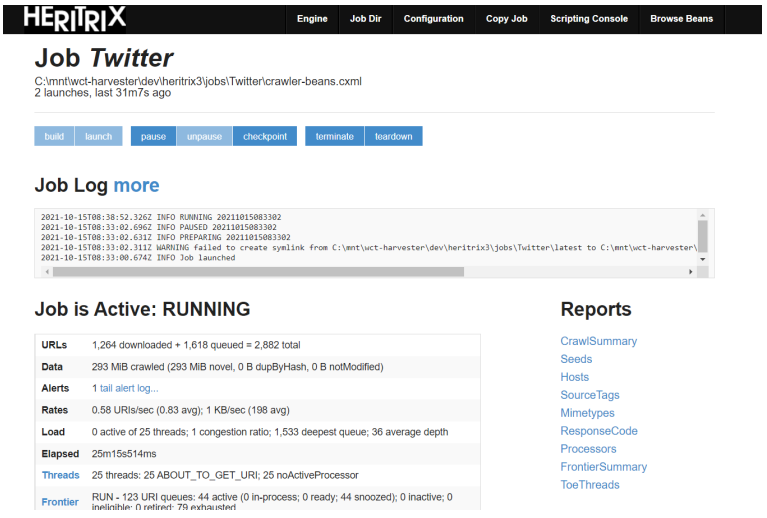

Figure 7. Local host of Heritrix

2.3.3 Conclusion

When harvesting the look and feel, you can count on a wide range of tools that automate archiving. For the test case, we used Webrecorder/Conifer to capture KBR's social media platforms Facebook, Instagram and Twitter. Despite a time-intensive undertaking (errors for Facebook), some of KBR's data was captured.

2.4 Overview decision-making mini-pilot

	Mini-Pilot 1	**Mini-Pilot 2**	**Mini-Pilot 3**
Question	What harvesting tool is the most efficient for harvesting social media collections?	How can we harvest social media images?	Can we harvest the look and feel of social media collections?
Goal	Testing different harvesting tools in order to pick the most efficient one for the BESOCIAL research project	Testing different harvesting tools to harvest images embedded in social media posts	Harvesting the look and feel of KBR's social media platforms
Why	To use this tool in the main-pilot	To use these tools in the main-pilot	To showcase what is possible where capturing the lay-out of social media platforms is concerned
Social media platforms	Twitter	Twitter and Instagram	Facebook, Instagram and Twitter
Used tools	Social Feed Manager CENTAL's tool[11]	Social Feed Manager and Instaloader	Webrecorder/Conifer Webrecorder Player Heritrix
Possible problems	CENTAL's tool: The tool has no interface, everything is done via a terminal. The tool is difficult to use if you don't have a minimum of computer knowledge The second problem stems from the first: as CENTAL's tool is not easy to use, it cannot be used for commercial purposes.	Instagram might block us, because we are not using an official API. Data storage might become an issue if we download too many images.	The autopilot of Webrecorder shuts down after a few minutes (depending on SM platform). Heritrix does not work with Facebook.
Conclusion and lessons learnt	Social Feed Manager is a good tool to harvest Twitter accounts and hashtags.	Harvesting Twitter images is possible, together with SFM. Harvesting Instagram images is difficult due to the lack of an official Instagram API.	The Webrecorder tool is user-friendly for pages or social media platforms that are not too large or do not contain more than 150 posts.

Table 4: Overview of the conclusions of the different feasibility studies

[11] Extra information about this tool including the code can be found in the annex section of this report.

3 Selection strategy

3.1 Selection strategy within the BESOCIAL project

KBR chose to conduct a **selective harvest** of the Belgian social media linked to cultural heritage. For selection we focused on publicly available content, excluding private profiles etc.

Within the BESOCIAL project a combination of a **top-down and bottom-up approach** for the selection of relevant social media content was implemented. The top-down approach covers outlining the selection criteria that constitute the scope of the social media collection related to Belgian cultural heritage. The bottom-up approach was setting up a crowdsourcing campaign to ask members of the public to recommend social media handles and hashtags to include in the collection.

3.1.1 Top-down: Selection criteria

The following selection criteria constitute the scope of the social media collection. A number of these criteria are very high-level and are linked to legislation. Taking into account the legal deposit legislation, the content needs to have a clear link with Belgium, therefore, the content of the social media indicated by hashtags and social media handles must be relevant to Belgian society. Content in all languages may be included, not only in the three official languages in Belgium (French, Dutch and English) as Belgian society constitutes a melting pot of different languages.

Another legal text that is important to keep in mind is the royal decree related to the missions of KBR. There are four missions that apply to social media archiving that need to be kept in mind:

- Acquire and keep for legal deposit, whatever the medium used, all publications published in Belgium and those of authors domiciled in Belgium published abroad, as well as publications on Belgium published abroad
- Manage, conserve and supplement an important cultural heritage
- Collect scientific, archival or documentary data relating to collections and disciplines
- Collect and inventory websites related to KBR's missions with the exception of blogs and private websites

In the context of the GDPR legislation, it is important to focus on content that is publicly available and exclude private accounts (non-public accounts of private persons) and private messages such as Direct Messages etc.

From a thematic point of view, the focus is on content that is linked to the cultural heritage of Belgium. Cultural heritage was defined as: "a collective term for everything made by previous generations that still exists today and has great value for the Belgian community. The BESOCIAL definition of this heritage is broad and covers movable, immovable and intangible heritage.

The following themes can also be part of it: hot topics in Belgium, sporting events, festivals, commemorations, literature, minorities, holidays, geographic regions, museums, art events, food, monuments, strikes, media, …"

Table 5 shows the seed lists have already been harvested. A short description of the way in which each seed list was compiled is included underneath the table.

Number of Batch	Topics
Batch 1	Minorities, Museums, Food, Commemorations, and Expats
Batch 2	Literature influencers, French and Dutch publishers, Literary institutions, Flemish and French authors, Festivals, and Heritage institutions.
Batch 3	Crowdsourcing campaign, radio, podcasts, journals, television, Belgian politicians, and Flemisch VIP's
Batch 4	Input collection managers KBR, …

Table 5: Overview of the different batches added to Social Feed Manager and Instaloader

Minorities collection

Minorities were defined as comprising foreign communities in Belgium, the homeless and people and organisations related to disability. A first step was to look for statistics on foreign communities in Belgium. Specific searches were launched on Twitter to find Rwandese, Congolese and Italian communities. The followers of these accounts were also checked and relevant accounts were added to the seed list. It quickly became clear, however, that many organisations prefer Facebook over Twitter as a platform and only a limited number of relevant Twitter accounts were discovered. Therefore, a list of all the foreign embassies in Belgium was added as well as a list of the Belgian embassies abroad. The website diplomatie.be provided a list of all these embassies, which was used as the basis for the seed list. About one fifth of all embassies have a Twitter account linked to from the website. For the homeless and disability, websites of related organisations were sought. From the website, links to social media channels were followed. Many of them have a Twitter account, but only a few have an Instagram account.

Food collection

Since Belgium is known for its gastronomy, the website of Michelin was used as a starting point. All accounts of restaurants with at least one Michelin star were listed based on their respective websites. Instagram is a much more popular platform than Twitter for restaurants. In addition, a list of the best Belgian chefs and chocolatiers was also compiled, many of whom have an Instagram account, but far fewer have a Twitter account. Some general hashtags related to Belgian food culture were also added, such as #BelgianFries or #BelgischBier.

Commemorations collection

This collection is based on Google searches. One approach was to look for keywords such as 'herdenk* + Belg*' or 'memorial sites + Belgium' in different languages. Important events and commemorations in Belgium were also listed based on articles on news websites such as vrt.be.

Literature

A number of seed lists underlie this collection focusing on: Flemish publishers, Belgian literary institutions, influencers within the Belgian literary landscape, Walloon authors, Flemish authors and

Belgian French-speaking publishers. KBR colleagues in the 'Prospection and control' department shared lists of publishers and authors that are regularly communicated with them by organisations such as Boekenbank (Flemish publishers) and ADEB (list of Belgian French-speaking publishers).

Museums

This collection is based on multiple Google searches. Finally, a list (based on Wikipedia and municipal and regional lists) of 455 museums located in Brussels, Flanders, Wallonia and the German-speaking region was compiled. Occasionally, random samples were also taken from the Instagram account of such a museum and their followers from museum organisations were checked to see if they also appeared in the proposed list.

Festivals

Because Belgium is known abroad for its many festivals, this was also seen as cultural heritage. Festivals such as Pukkelpop or Rock Werchter are part of most Belgian summers. After several google searches (on provincial websites such as Vlaams Brabrant, and on festival sites), a list of 401 Belgian festivals was compiled. After this, we manually checked whether they had an Instagram account and a Twitter account.

Heritage collections

For a part of the seed list created for this purpose, the heritage cell of Flanders and Brussels was gathered. The followers of these accounts were also checked and relevant accounts were added to the seed list. Important hashtags and organisations outside of these complement the seed list. In the near future, Belgian immaterial heritage and the Flemish art collection will also enrich this list. Some general hashtags related to Belgian heritage were also added, such as #ikschrijfbelgisch or #kikirpa.

Media (journals, podcast, radio, tv, VIP's)

For the newspapers, all A-list newspapers (19 newspapers) were collected in a list. Mediahuis was consulted online and the collection of newspapers preserved in KBR was also added. Information was also collected on the content of the posts and how the different platforms and newspapers differ in posting information.

Podcast: For this purpose, we have currently chosen to select a number of the most listened to Belgian podcasts. This is because the Belgian range is enormously broad. Listening platforms such as Deezer and Spotify were consulted for this, as were online websites that showcase the top Belgian podcasts. In the near future, it is recommended to extend this seed list.

Radio and TV: Because listening to radio and watching tv is also part of cultural living in Belgium, it was also decided to add this type of media. This collection is based on Google searches.

VIP: the focus was on finding accounts of living Flemish famous persons. Fan accounts fell outside the scope. The focus has so far been on Flanders, but the seed list would need to be enriched with Walloon celebrities in the near future.

Crowdsourcing campaign

The BESOCIAL researcher collected all the output received by the public. She curated some suggested accounts and hashtags because they were outside the scope of the selection criteria (e.g. personal accounts or not of great importance for Belgian cultural life). Three lists were created, one in French, Dutch and English. However, all languages were welcome, as long as the link with Belgium was made. In total (read: mid-April 2022), more than 700 suggestions were collected. The campaign will continue until the end of the BESOCIAL project (15 September 2022).

It became clear from the suggested accounts and hashtags that the focus was more on recent and older but high-profile events such as #neuzekesoorlog (a 'war' between two competing manufacturers of the typical candy 'neuzekes' from Ghent) and #demol (a famous Flemish TV programme in which candidates need to uncover the mole in their group).

Belgian politicians

This seed list was set up as part of an internship for the advanced master in Digital Humanities. For the intern's research, the following groups were not included in the seed list: chambers, namely the governments and the French commission and the Belgian senate. For example for the French commission and senate the focus was, in addition to the legislative branch, on the directly elected representatives. The student looked at the party website in order to collect the Twitter accounts of the respective persons via a semi-automatic code.

Belgian Modern travelogues

This seed list was set up in the context of another internship for the advanced master in Digital Humanities. The research focussed on getting a grip on the Belgian modern travelogues/travel bloggers on the social media platform Instagram. The follow criteria are used to capture the correct accounts and hashtags:

- The account owner should be (sporadically) based in Belgium
- The account owner does not have to be from Belgium when posting
- The account owner should identify itself as a traveler/travel blogger
- The account owner does not have to be traveling only within Belgium

With this background two hashtags were chosen to list the different accounts that meet the description: #travelbelgium and #belgiumtravelblogger

Collection managers

The curators of the various departments within KBR were contacted to compile this list. First of all, as a test round, input from the collections of coins and medals (head: Fran Stroobants) and prints (head: Daan Van Heesch) was requested in an informal interview lasting about an hour. They were asked about the size of their collection and the networks they have set up, or want to set up, within Belgium. Afterwards, both sent a list of important Belgian players in the day-to-day operation of their department. Also, a mail was sent to their respective teams to gather their knowledge and input on

accounts and hashtags related to Belgian cultural heritage. In a next phase, the other collections within KBR will also be addressed.

3.1.2 Bottom-up: Crowdsourcing campaign

Why use crowdsourcing in a library environment

"Crowdsourcing occurs when an entity seeks input from an undefined group of users. This type of input has become a popular tool with social networking sites such as Facebook and LinkedIn. Users can create their own content and specify their preferences.".[12]

National libraries and libraries in general are increasingly set on getting input from their readers in order to contribute to the daily working of their institution.[13] This contribution can be very small in scale from completing a survey to checking the transcription of handwritten texts. Although crowdsourcing often means relinquishing part of the control, it is a win-win for both parties. The public gets a say in the collections managed by the libraries and the institution engages in an open dialogue.

Within KBR, a large-scale crowdsourcing campaign had never been initiated. Crowdsourcing has so far been limited to surveying the patrons using the general reading room for example.

Crowdsourcing holds a lot of potential benefits for KBR:

- Achieving goals the library would never have the time, financial or staff resource to achieve on its own.
- Achieving goals much faster than the library may be able to if it worked on its own.
- Actively involving and engaging the community with the library and its other users and collections
- Utilizing the knowledge, expertise and interest of the community (including minorities)
- Improving the quality of data/resources (e.g. by text, or catalogue corrections), resulting in more accurate search results
- Demonstrating the value and relevance of the library in the community by the high level of public involvement.
- Strengthening and building trust and loyalty of the users to the library.
- Encouraging a sense of public ownership and responsibility towards cultural heritage collections, through users' contributions and collaborations.[14]

Why use crowdsourcing for BESOCIAL?

---

[12] Budzise-Weaver, T., Chen, J. and Mitchell, M. (2012), "Collaboration and crowdsourcing: The cases of multilingual digital libraries", *The Electronic Library*, Vol. 30 No. 2, pp. 220-232. https://doi.org/10.1108/02640471211221340

[13] Australian Newspapers Digitisation Program, National Library of Australia: Text Correction. http://www.nla.gov.au/ndp/: The public are enhancing the data by correcting text and adding tags and comments; Picture Australia, National Library of Australia: Creation and addition of images: Since January 2006 the public have been encouraged to add their own digital images to Picture Australia via a relationship set up with Flickr.; …

[14] Holley, Rose Crowdsourcing: How and why should libraries do it? *D-Lib Magazine*, 2010, vol. 16, n. 3/4 Ma. [Journal article (Unpaginated)]

Within the BESOCIAL project the choice was made to work with a seed list. This is a (living) list that contains a collection of all the data that needs to be harvested within the project. This list includes both Twitter and Instagram content. It is subdivided so that accounts and hashtags can be added during the entire project. Initially, the creation of this list was going to be internal by seeking input from the experts within the BESOCIAL team, the KBR and our external network. However, it was decided to also consult the public of KBR and give them a voice in this process: the start of the crowdsourcing campaign. In this way Belgian citizens (or people that are in some way related to Belgium) will have the opportunity to submit hashtags and/or accounts that are linked to the definition of Belgian cultural heritage (see above).

In order to ensure the representativeness of the date, this list is constructed via two fronts; top-down and bottom-up: **1) the BESOCIAL team, and 2) the Belgian public**.

In the beginning of the process we realised that we needed to have as low a threshold of participation as possible in order to get a useful input. **Encouraging anonymity** was immediately an important point. The choice was made that participants could optionally add their email address. This leaves room to 'anonymously' nominate some hashtags and/or accounts with as few clicks as possible while leaving the proverbial door open for curious participants. When leaving their email address, participants were automatically sent an email that their input had been well received, and that for more info they could reach out to the following email address: socialmediaarchive@kbr.be.

Another important point is to be **transparent** throughout the campaign. For this crowdsourcing campaign this means that we have developed a policy in terms of selection criteria, as well as in legal terms. A GDPR compliance statement was developed. These two documents can easily be found on the crowdsourcing page on the KBR website: https://www.kbr.be/en/save-the-social-media-of-belgium/.

Execution

The idea for this campaign arose around May 2021. Within BESOCIAL we realised that we could not collect all accounts and hashtags. There are simply too many accounts and hashtags related to cultural heritage in Belgium. Furthermore the BESOCIAL team does not have the knowledge to list all events and accounts that fit the selection criteria. Therefore, it was quickly suggested to set up a crowdsourcing campaign. After an initial pitch, the communication team at KBR was enthusiastic and decided to join forces with us to set up a campaign. The first idea was to do the promotion internally (via digital storytelling). We soon realised that we could go a lot further and the communication team decided to bring in an external party for the campaign: Bureau 87 Seconds.

The table (see Table 6) below briefly explains who we are trying to reach with this campaign, why, how and what resources we used to do so.

Who	**KBR employees**	**External experts**	**Belgian public**
Why	**Top-down**: To receive their input and expertise on social media accounts that are linked to the collections of KBR and the national cultural heritage	**Top-down**: To receive their input and expertise on social media accounts that are linked to the national cultural heritage	**Bottom-up**: To receive their input on social media accounts that are linked to the national cultural heritage

	(e.g. department of coins can submit the social media account of @DIVA)	(e.g. Partners in Best practices voor de archivering van sociale media in Vlaanderen en Brussel)	(e.g. A Belgian citizen can submit their historical society)
How	Via campaign on KBR's website where they can submit accounts and hashtags Via promotion video (mixed media animation by 87Seconds)		
	Via newsletter and Intranet	Via mailing (contacting network)	
Resour-ces	Time to set-up crowdsourcing page Time to develop the selection criteria and GDPR compliance Time for communication department Time from the BESOCIAL team to coordinate Time from the BESCOCIAL team to conduct quality assurance at the end Time from IDLab and CENTAL to create a tool to visualise results (e.g. timeline and/or top 10#'s) Budget to create animation (promotion)		
When	Start: Mid-October End: to be determined		

Table 6: Overview of the crowdsourcing campaign

Campaign on KBR website

The web page was developed by the BESOCIAL team in collaboration with KBR's Communications Team. BESOCIAL drafted the selection criteria, GDPR compliance and general information. The communications team translated these into an accessible text.

The page is available in three languages (Dutch, French and English) and is divided into several catchy paragraphs (see Figure 8). The beginning of the page explains why there is a need for this crowdsourcing campaign. Then the text focuses on the specific selection criteria. One can click on '+' to get additional info and examples. An option is also included to be redirected to the extensive selection policy page for social media or to click through to the GDPR compliance page.

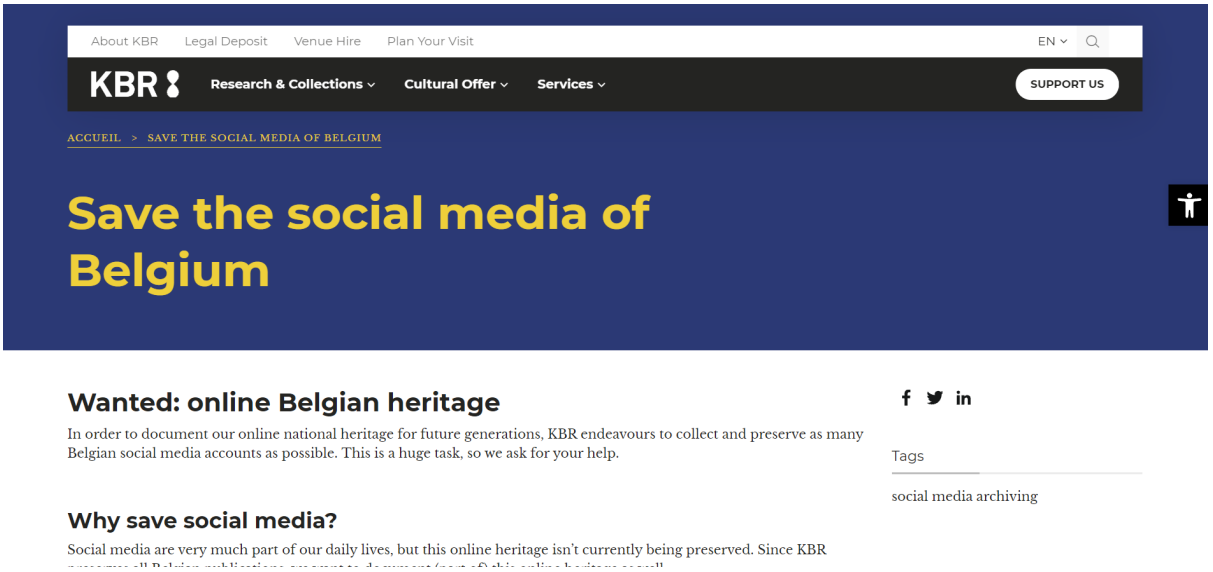

Figure 8: Crowdsourcing page on KBR website.

At the bottom of the page one can submit accounts and hashtags (see Figure 9). If one wishes to add more than one hashtag, it is possible to simply click on the plus button. Participants also have the choice to optionally leave their email address. They will receive a confirmation email after submission.

The BESOCIAL team has back-end access to the entered data. A csv can be exported with the output. This data was checked by the team to see if the entered accounts and hashtags are linked to the selection criteria. If so, they will be added to the seed list.

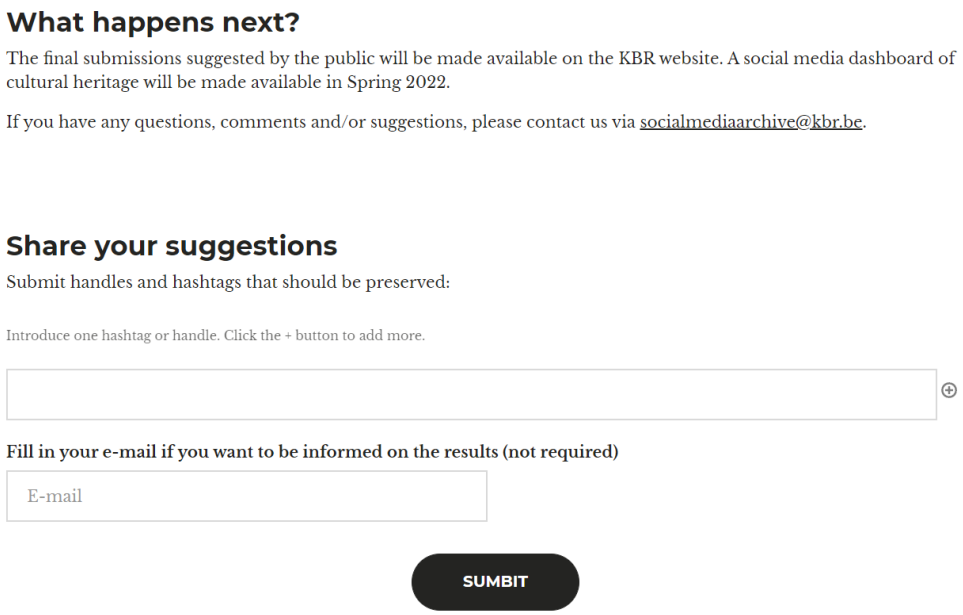

Figure 9: Crowdsourcing page on KBR website. The submit function.

The promotion of the crowdsourcing campaign

To develop the promotion campaign, KBR hired the expertise of Bureau 87 Seconds. Via a 45-second animation (using mixed media), the input from the Belgian public was requested. KBR's house style (see hourglass) is honoured at all times. The intention was that this animation was distributed via social media platforms. People who wished to add accounts and/or hashtags could do so by clicking on the link as well as adding their accounts and hashtags in the comment section at the bottom of the video. The latter option limits the number of clicks and is therefore fairly barrier-free to submit anything. The recommendations submitted as a comment will be compiled manually after a round of quality assurance.

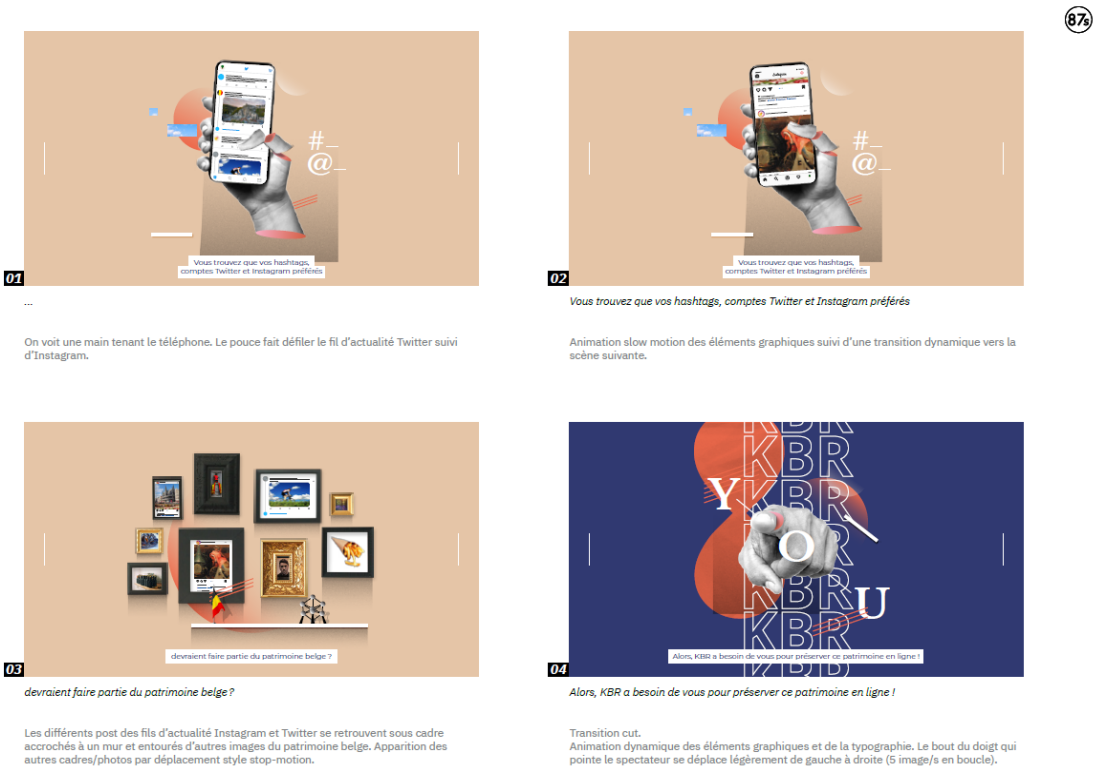

Figure 10: Part one of the proposal of the animation of 87 Seconds

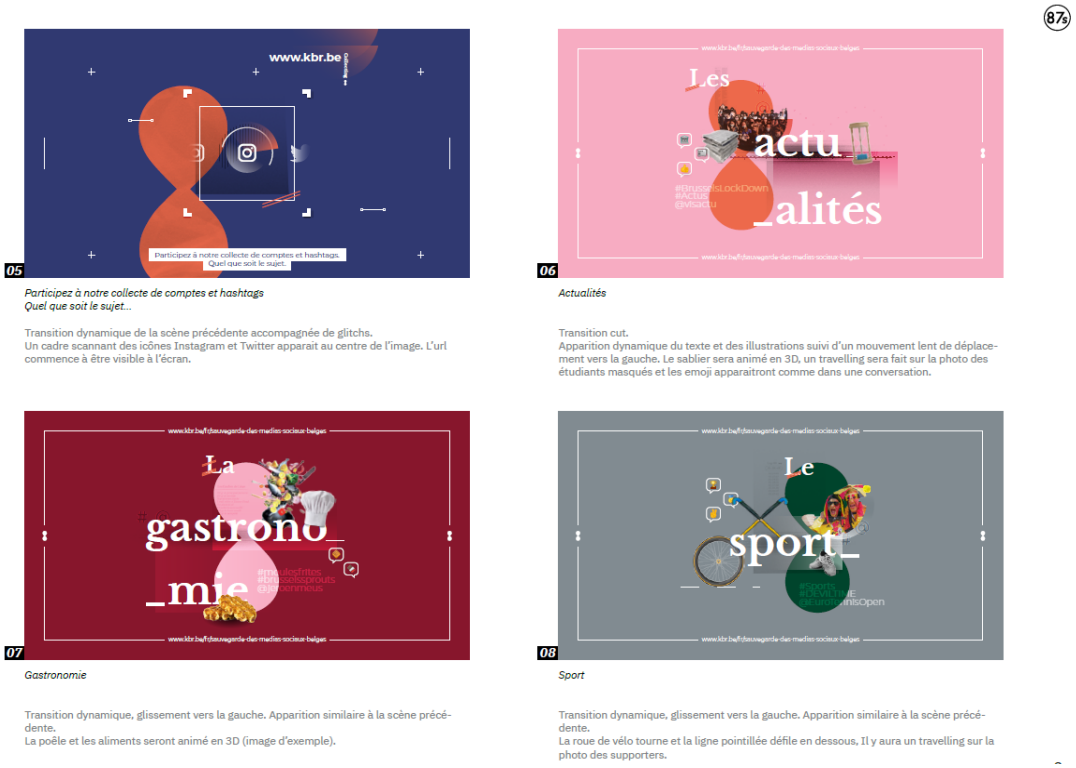

Figure 11: Part two of the proposal of the animation of 87 Seconds

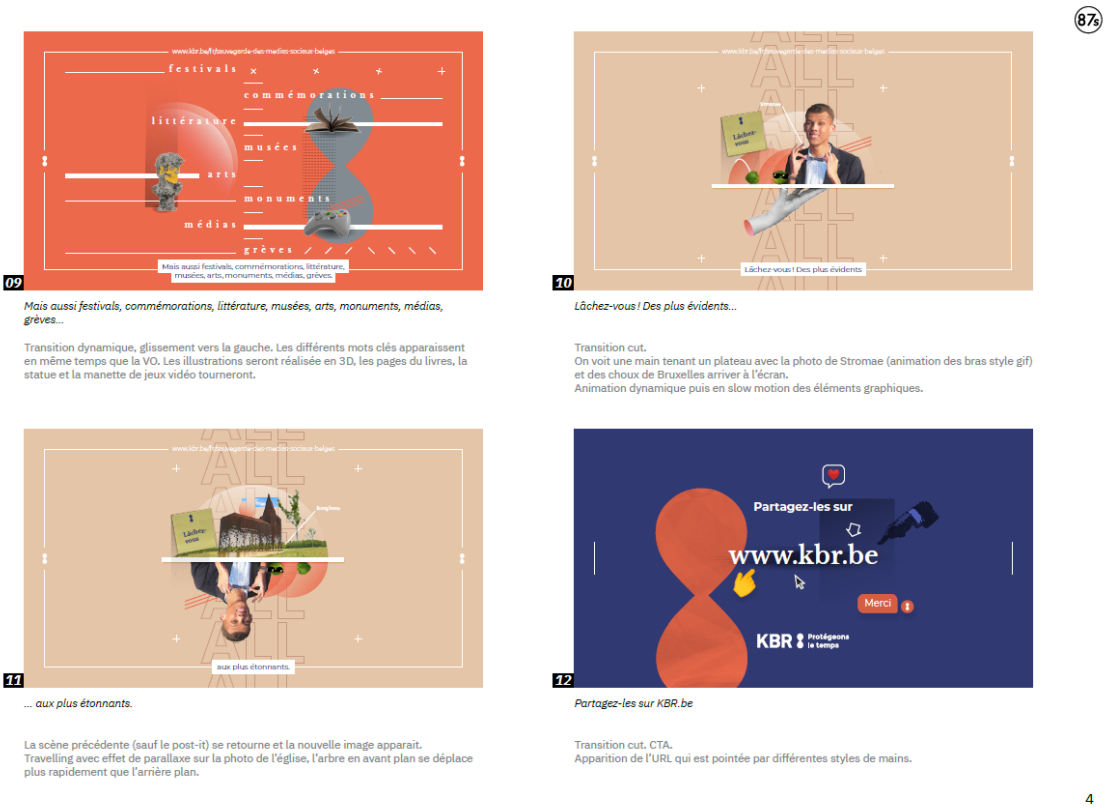

Figure 12: Part three of the proposal of the animation of 87 Seconds

The video was finalised at the end of October and sent into the world on 27/10. The animation was created in French and Dutch and was distributed over different social media platforms: Instagram, Facebook, LinkedIn, …. The communication department also sent a press release to a number of papers and news services (in Dutch and French). The day after the launch of the crowdsourcing campaign the communication was well picked up by the press. The project and the start of the campaign was presented on Radio 1 on the program De Wereld Vandaag. Here, the focus was on introducing the BESOCIAL project, and explaining it in an accessible way to the public of Radio 1 with the aim of capturing the public's enthusiasm and asking for their input to suggest accounts and hashtags. Further interviews linked to BESOCIAL's crowdsourcing campaign were published in Flemish and Walloon press: DailyScience (11/012022), De Standaard (29/10/21), Kerk en Leven (28/12/2021), Radio 1 (28/10/21), VRT (28/10/21), Bruzz Radio (November 2021), RTBF (28/10/2021), VivreIci (20/12/2021), Magazine Athena (February 2022), RTBF (17/02/2022).

Result

The goal of this crowdsourcing campaign was to receive input from the public on which hashtags and accounts they think are important to harvest for future generations. The project succeeded not only in appealing to the Belgian public, but also in suggesting as many hashtags and accounts as possible. At the beginning of May 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.

Sociale media maken vandaag de dag zonder discussie integraal deel uit van ons dagelijks leven en zelfs van onze cultuur, vindt de Koninklijke Bibliotheek. Maar zullen Twitter of Instagram binnen 50 of 100 jaar nog beschikbaar zijn? De KBR wil daarom een duurzame strategie ontwikkelen om sociale media te archiveren.

"De taak van KBR focust zich onder meer op het bewaren en archiveren van erfgoed. Het gaat dan om manuscripten, tijdschriften of boeken. Met dit project rond sociale media willen we ook graag online erfgoed de kans geven om bewaard te worden", legt Fien Messens uit. Zij leidt het BESOCIAL-project.

Figure 13: Extract from article on crowdsourcing campaign as published on Vrt.be[15]

[15] https://www.vrt.be/vrtnws/nl/2021/10/28/online-erfgoed/

3.2 Preliminary recommendations for a social media collection policy at KBR

The choice of the social media platforms to focus on will depend on the available infrastructure. If KBR chooses to outsource the social media archiving and use a third-party service, the choice will be wider than if the social media archive is managed internally. Certain social media platforms and especially Facebook regularly change the parameters meaning that a lot of work is needed to allow the crawlers to access this content. If the social media archive is managed internally, collecting Twitter data is the easiest option to implement from an operational level.

It is recommended that KBR focuses both on hashtags and on relevant social media profiles and accounts. One caveat that needs to be kept in mind is that in the case of hashtags it is more difficult to determine the link with the Belgian territory in order to fit within the legal deposit legislation. Procedures to determine the provenance of hashtags or of posts including certain hashtags need to be developed. For profiles an approach similar to that used by the BnF could be used: if the link with the country is not made explicit in the account itself, external sources such as news articles, academic websites or LinkedIn can be used to determine whether or not the person or organisation in question is linked to Belgium.

KBR should first of all focus on content that is publicly accessible. As mentioned above, the link with Belgium has to be clear to respect the legal deposit legislation. Furthermore, the analysis of social media archiving initiatives abroad has shown that most institutions collect social media content as part of selective or thematic collections. KBR needs to reflect about what thematic collections they would like to curate, also keeping in mind the archiving of websites:

- Collections KBR: American studies, Bibliography of Belgium, Chalcography, Coins and medals, Contemporary printed books, Digital collections, Manuscripts, Maps and plans, Music, Newspapers, Print and drawings, and Rare books.
- Thematic collections: hot topics/social debates, sporting events, Festivals, Commemorations, Literature, Disaster, Elections and referenda, Topstukken, Minorities, Holidays, Geographic regions, Museums/exhibitions/art events, Food, Monuments, Strikes, Media (focus on Newspapers), Entertainmentparcs, Intangible heritage, Unesco, …

It is recommended that KBR uses a seed list that keeps track of content that is harvested and/or will be harvested. The thematic collections and KBR's collections are a guide for the curators of the list.

A reflection about additional selection criteria is also necessary. The criteria mentioned by the BnF are a good source of inspiration: checking the activity of the accounts and the richness of the content and excluding material that is purely (self-)promotional.

Keeping a manifest to keep track of content that KBR would have liked to archive but was not able to for technical reasons as is done at the UK Web Archive is also recommended.

When it comes to ways to identify relevant accounts or hashtags, the use of search queries on the social media platforms themselves is a good starting point (Twitter trends etc.) or starting from websites included in thematic collections and adding the social media channels mentioned. For hashtags plural forms and hashtags with typos also need to be considered. Using external services

that can provide KBR with lists of accounts to consider, similar to brandmentions.com used by Netarkivet, is another possibility.

One point that was consistently mentioned by representatives of social media initiatives abroad was the challenge to create inclusive and representative collections. It is therefore recommended that KBR develops different strategies to broaden the selection of relevant content by setting up an internal network of curators within the institution (maybe through something similar like the café éphémère at the BnF), setting up an external network of experts outside of the institutions or stimulate collaboration with research groups and invite recommendations from the public via crowdsourcing campaigns. The crowdsourcing campaign set up within the BESOCIAL project is a significant first step in the right direction.

It is essential that the archived social media is continuously monitored. If hashtags are becoming less popular or if accounts are becoming less active or if the content posted is no longer within scope, these need to be removed from the seed list. The monitoring of social media to archive will require more resources than the monitoring of websites as social media content and accounts change more rapidly than content on websites. The frequency of capture also needs to be adapted to both the platform and the type of content and can range from daily to a few times per year.

As events on social media unfold very quickly, it is also imperative that active monitoring procedures are put in place at KBR to detect relevant content and start the collecting process as the events unfold.

The policies regarding including comments and other user-generated content differ between institutions abroad. It is therefore recommended that KBR follows the legal recommendations that will be made within the BESOCIAL project. It is also important to keep in mind what is possible from a technical point of view. The standard settings of Social Feed Manager (and the underlying twarc) do not capture the profile picture, but do harvest the username or user ID. Social Feed Manager uses the Twitter API to obtain a JSON object including different metadata fields. Metadata fields cannot be excluded while requesting content from the API. It would therefore be necessary to manipulate the resulting WARC file in order to filter out the unwanted content, which would also require attaching provenance information indicating what was manipulated. Regarding (profile) pictures in tweets, it is foreseen that these are downloaded separately based on the picture URLs in the tweet metadata by means of a specific script. This script could in theory filter the profile pictures.

It is recommended that embedded content is collected as well in order to have a comprehensive crawl. The hyperlinks mentioned could be extracted from the social media posts and fed to a web crawler. Embedded images could be captured as well. Given that the legal deposit legislation excludes all resources made by cinematographic processes, excluding video content from the collections is a possibility. It also needs to be monitored whether the inclusion of embedded content does not result in too much data.

4 User Requirements (task 2.2)

4.1 Introduction

This chapter describes the results of the BESOCIAL Task 2.2, titled 'Analysis of user requirements'. The goal of this task is to gain insights into the needs and requirements of a broad range of stakeholders (researchers, cultural heritage professionals, publishers, policy-makers and other potential end users) when using a social media archive. These insights will be taken into account when designing (access to) the social media archive itself. This task also wants to understand how specific user requirements of potential social media archive users might influence adoption and usage of such archives. Meeting these requirements and sociotechnical needs is not an easy task; already in 2000 a persistent 'socio-technical gap' between users' or researchers' requirements and the functionalities or affordances of available interfaces and (technological) infrastructures has been identified (Ackerman, 2000).

Research for this task encompassed three phases. After an initial **desk research phase** (i), we decided to use qualitative **semi-structured interviews** (ii) with potential end users of social media archives to gain insights in the needs and requirements of a broad range of stakeholders. Insights and take-aways from these semi-structured interviews were then validated in a pre conference workshop (iii) of the **2021 RESAW conference**. Below, we discuss the results of these three phases.

The entire report can be found here

5 Analysis of existing legal frameworks in Belgium for selection of content for social media archiving

A lot of different legal considerations can be examined when it comes to Social Media Archiving, as the content harvested can raise several issues related to copyright, data protection, privacy, illegal content or preservation. However, separating the legal analysis into two distinct parts, one on the selection and the other on the access as it will be done in the WP2 and WP4 reports, would not be efficient as we cannot really distinguish those two phases properly from a legal point of view. Therefore, all legal aspects of the BESOCIAL project are written down in the 'Global legal report'.

6 Definition of the Technical and Functional Requirements based on the OAIS model

The aim of this document is to identify and describe 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 collecting. Note that this statement is a living document and is ever-evolving. In the final stage of the creation of these functional and technical requirements, called version 2, the operational level will be described. This task will ensure that the link is made with the IT infrastructure at KBR.

The entire report can be requested by e-mail (friedel.geeraert@kbr.be)

WP3 Report: Pilot for social media archiving

1 Introduction

In the previous part of the report, the research results of Work Package 2 were described, which was aimed at preparing the main pilot. This part of the report 'Pilot for social media archiving' aggregates the research results from Work Package 3 (WP3) for which we drew from the research results obtained in Work Package 2 of the project.

In this work package, three dedicated tasks constitute the main pilot. *Task 3.1 Development of a social media harvester and harvesting social media content* aimed to harvest the seed lists created in the previous working package. In *Task 3.2. Quality control of harvested content* the quality of the social media content harvested as part of the pilot was assessed. In *Task 3.3. Development of a preservation plan for archived social media* the focus was on developing preservation plan for archived social media. The plan covers aspects such as the definition of preservation formats, description of the quality control (see Task 3.2), description of the different stages files go through etc. The adaptability of KBR's current preservation workflow was studied in order to define the necessary preservation actions for social media archives after the end of the project.

This report is structured as follows. **Section 2** discusses Task 3.1, **Section 3** Task 3.2, **Section 4** Task 3.3. The findings of these different tasks are summarized in the discussion and conclusion.

2 Development of a social media harvester and harvesting social media content (Task 3.1)

2.1 Introduction

In Work Package 2, we tested several social media harvesters: the one proposed by IDLab and the one created by the CENTAL. For several reasons already stated (see page 16), we decided to turn to the tool proposed by IDLab. CENTAL on the other hand decided to focus on the development of the interface (part of Work Package 4) as well as on the development of an algorithm able to select only Belgian content related to Work Package 2.

The social media harvester setup by 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.

2.2 Documentation

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.[16] For Instaloader we paid special attention to how to set up periodic harvesting as this is not supported out of the box.

2.3 Semantic annotations

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.

2.3.1 Data model

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.

We briefly explain some concepts of the data model. A ProvidedCHO represents a cultural heritage object, so the thing one wants to describe (e.g. the Mona Lisa painting). An aggregation groups all representations from one provider. A proxy is specific to one given aggregation. It allows to represent different, possibly conflicting pieces of information from an aggregation and thus provider's perspective.

In our use case of social media

- A social media post, such as a tweet is an ProvidedCHO, represented using its unique ID, e.g. the unique tweetID.
- Different representations of a tweet are instances of WebResource, e.g. a HTML/WARC representation or a JSON representation.
- Currently we are a single provider of information, thus we use one Aggregation linking the ProvidedCHO with its WebResource.

According to EDM mapping guidelines, the most appropriate level of granularity has to be chosen for a cultural heritage object. For instance, in archiving different levels such as "sub-series", "file", or "item" exist. Each thing on these levels would become an instance of ProvidedCHO.

For social media, we usually have collections of posts. Because describing each thing might defeat the purpose of having a small comprehensible metadata record, i.e. having one piece of paper summarizing what is inside a box.

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

The following two examples show first a low level representation as it describes each tweet and the second example shows a tweet collection.

```
# A URI representing the real world object, in our case a tweet
ex:myTweet a edm:ProvidedCHO ;

# Different representations of a social media post
ex:myTweetHTMLWARCFile a edm:WebResource .
ex:myTweetFullJSONFile a edm:WebResource .

# Our description linking the object we describe with the representations
of it we have
ex:myAggregation a edm:Aggregation ;
  edm:aggregatedCHO ex:myTweet
  edm:hasView ex:myTweetHTMLWARCFile ;
  edm:hasView ex:myTweetFullJSONFile .
```

A whole tweet collection represented using EDM, this also reflects reality as usually harvested tweets are also grouped in files already such as line-based JSON files or WARC files.

```
# A URI representing the real world object,
# in our case a collection of tweets
ex:electionTweetCollection a edm:ProvidedCHO ;
  rdfs:label "US Election 2020 tweet collection" .

# Two representations of this collection
ex:electionTweetsHTMLWARC a edm:WebResource .
ex:electionTweetsFullJSON a edm:WebResource .

# Our description linking the collection and
# the representations of it we have
ex:ourElectionArchive a edm:Aggregation ;
  edm:aggregatesCHO ex:electionTweetCollection ;
  edm:hasView ex:electionTweetsHTMLWARC ;
  edm:hasView ex:electionTweetsFullJSON .

# establish a link between low level tweet and collection
ex:myTweet dcterms:isPartOf ex:electionTweetCollection .
```

2.3.2 Generating semantic annotations

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.

2.3.3 Use case specific constraints

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.

2.3.3.1 SHACL shapes for data representations

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.

The following two SHACL shapes apply on ProvidedCHO and validate if this 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.

A high level data shape for data scientists.

```
@prefix bss: <http://example.org/ns/besocial/shapes#> .
@prefix edm: <http://www.europeana.eu/schemas/edm#> .

# A data shape for 'provided Cultural Heritage Objects'
# validating that at least one machine readable version exists
bss:analysisShape a sh:NodeShape ;
  rdfs:label "Analysis Shape"@en ;
  rdfs:comment "A harvested object needs to have a machine readable
representation"@en ;
  sh:targetClass edm:Aggregation ;
  sh:property [
    sh:path edm:hasView ;
    sh:qualifiedValueShape bss:jsonShape ;
    sh:qualifiedMinCount 1 ;
  ] .

bss:jsonShape a sh:NodeShape ;
  rdfs:label "JSON Shape"@en .
```

A high level data shape for social scientists

```
@prefix bss: <http://example.org/ns/besocial/shapes#> .
@prefix edm: <http://www.europeana.eu/schemas/edm#> .
```

```
bss:lookAndFeelShape a sh:NodeShape ;
  rdfs:label "Look And Feel Shape"@en ;
  rdfs:comment "A harvested object needs to have a representation which
preserved the look and feel"@en ;
  sh:targetClass edm:Aggregation ;
  sh:property [
    sh:path edm:hasView ;
    sh:node bss:htmlShape ;
  ] .

bss:htmlShape a sh:NodeShape ;
  rdfs:label "HTML Shape"@en .
```

2.3.3.2 SHACL validation to provide collection level metadata

Within archiving metadata records, summarising collections exist and are called “finding aids”. Such summarisations are smaller than all the actual data and aim for human consumption.

A SHACL validation on the elements of a collection can inform a summarised metadata record of the collection. The previously presented shapes can be applied and 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 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".

2.4 Deployment

IDLab deployed both Social Feed Manager and Instaloader on a server of CENTAL as part of the pilot. We tested our documentation during the process and added clarifications where needed. We made sure that the used server had enough disk space (500GB) for the coming months. At the moment of writing, both the harvests of Twitter and Instagram are spread out over the week, so there are no strict minimum requirements when it comes to processing power and memory. We also gave KBR access to the server, next to IDLab and CENTAL. This allowed KBR to download Twitter and Instagram exports from the server so that they can use it for their analysis. Additionally, KBR also got access to the user interface of Social Feed Manager to inspect and update the harvests.

3 Quality control of harvested content (Task 3.2)

3.1 Introduction

T3.2 wants to assess the quality of the harvested content in BESOCIAL. Insights from this assessment can feed the technological development and optimisation of the harvesting process. Quality assessment or control refers to the evaluation of harvested web resources while determining whether certain quality standards are attained. The British Library uses four aspects to define quality: (i) completeness of capture, (ii) intellectual content (whether the intellectual content can be

'replayed'), (iii) behaviour (whether the harvested copy can be replayed including the behaviour present on the live site), and (iv) appearance or the look and feel of a website.

(Information) quality assessment or quality control (QC) is a broad field with various methodologies being used. For example, some research looks at the notion of Website Archivability (WA) and tries to capture or diagnose whether a website has the potential to be archived with completeness and accuracy a priori (e.g. see the 'Credible Live Evaluation of Archive Readiness' method[17]) as this appreciation of the archivability provides archivists with a valuable tool when assessing the possibilities of archiving material.

As the quality of information online is highly variable, other researchers to this end study users carrying out specific tasks on the Web. This has demonstrated that information quality assessment is composed of four components: credibility of content, credibility of site, predictive relevance and veracity assessment (Fink-Shamit et al., 2008).[18] Librarians and other information professionals also use other methods or procedures to guarantee quality, substituting QC by using guidelines or fact checking[19], or by using a contextual or common sense approach. Still other approaches involve computational and automatic QC using for example neural networks that are trained on visual differences between the web page during archiving and reproduction.[20]

In order to evaluate the quality of the harvested content - in specific the COVID-19 related content that was harvested during BESOCIALs' 'mini-pilot' - we will adopt a pragmatic and qualitative approach that starts from the different persona that were developed in WP 2 (Task 2.2 - Analysis of user requirements). A persona is a fictional character created to represent a user type that might use a social media archive in a similar way. A persona is thus a composite representation of prevalent qualities of a user segment and will not exactly match a specific person or comprehensively describe the full diversity of a group. Our QC using persona's can thus be considered to be a 'common sense approach' to QC as the knowledge accumulated during development of the personas will allow us to assess the quality of information.

Based on previously developed persona in the field of (web) archiving (see Task 2.2), and taking group discussions and the results of the PROMISE project into account, five persona were created (a full description of these 5 personas is provided in the Annex of Task 2.2):

- Bart - postdoc researcher in communication sciences
- Febe - PhD-student in computational linguistics
- Ben - 35-year old journalist working for the newspaper
- Manou - 28-year old scientific researcher @ KBR

[17] Banos, V., Kim, Y., Ross, S., & Manolopoulos, Y. (2013). CLEAR: a credible method to evaluate website archivability.

[18] Fink-Shamit, N., & Bar-Ilan, J. (2008). Information quality assessment on the web–an expression of behaviour. Information Research, 13(4), 13-4.

[19] See e.g. https://support.archive-it.org/hc/en-us/articles/208333833-Quality-Assurance-Overview or https://www.data-archive.ac.uk/managing-data/digital-curation-and-data-publishing/quality-control/

[20] See e.g. Kiesel, J., Kneist, F., Alshomary, M., Stein, B., Hagen, M., & Potthast, M. (2018). Reproducible web corpora: interactive archiving with automatic quality assessment. Journal of Data and Information Quality (JDIQ), 10(4), 1-25.

- Jan - 68-year old former team coordinator

In order to reduce the complexity of evaluating the harvested content from 5 different perspectives, it was decided early on in the task, to reduce these to 3 personas for the purpose of the quality assessment, namely: Febe, Manou and Jan. For the same reason, the persona description, and in specific the research question for each persona, was revised to make them applicable for the content of the BESOCIAL 'mini-pilot'.

Next to this persona-driven research approach, this document also takes a more computational approach to assess the usability and quality of the content of the BESOCIALs' harvesting 'mini-pilot', see section 2.2.

Finally, we used the benchmark of datasets for computationally-driven research developed by Candela et al. (2021)[21], see section 2.3.

The entire report can be found here

4 Development of a preservation plan for archived social media

4.1 Introduction

The purpose of this plan is to provide procedures and guidelines for internal users at KBR (Royal Library of Belgium). Internal users are interpreted as archivists, librarians and administrative collaborators. In addition, the document is also prepared for external parties such as the broad public. These procedures and guidelines regarding digital preservation 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 and technical contexts within which KBR.

The development of a preservation plan for archived social media is the result of prior research in earlier work packages within the BESOCIAL project (Task 1.4 Analysis of Preservation Policies) and the first tests of collecting social media content (Task 3.1 and Task 3.2) provided us with the necessary information to develop a preservation plan for archived social media. The plan will cover 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 studied in order to define the necessary preservation actions for social media archives after the end of the project.

4.2 Importance of digital preservation

Any digital format can qualify for digital preservation. In our case, it is social media data from the social media platforms Twitter and Instagram. New methodologies and research insights in different fields (humanities, communication sciences, psychology) come to light as a result when opening up

[21] 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.

new data types to researchers. Despite the opportunities that digital data (such as social media) creates, the data format remains fragile. Social media platforms are complicated and fluid, and changing every day. “Their [digital data] longevity and utility is threatened where contents or contexts are lost: engagement and exploitation are enabled when digital materials endure. The greater the importance of digital materials, the greater the need for their preservation: digital preservation protects investment, captures potential and transmits opportunities to future generations and our own'' (Digital Preservation Handbook).

Several archives and libraries have already developed and implemented plans to avert the digital dark age. Some even started surprisingly early such as the UK Data Archive that was founded in 1967. So there is a large basis and foundation of expertise on how to professionally give digital preservation a place in the library and archive environment. Yet the emerging formats remain a shared generational challenge. To guarantee the quality of digital data in the long term, we must also set our sights on access, which in turn means we need to understand and mitigate rapid changes in technology and organizations.

> “Digital preservation is an important, necessary and doable endeavor with simple first steps all can undertake.” (Digital Preservation Handbook)

4.3 Approach

The document is divided in three parts:

- Get to know your organisation and your data: In this phase we paint the scene of the readiness for digital preservation at KBR.[22] An extra step here is to outline 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.
- Long-Term Preservation Plan: The last and most important phase is identifying a framework using the OAIS model for preserving the WARC and JSON files for a sustainable long-term platform. This phase is the most challenging since (to our knowledge) there has not been created such an elaborate plan for social media as data type.

The entire report can be requested by e-mail (friedel.geeraert@kbr.be)

[22] Messens, Fien; Pham, Thuy-An; De Preter, Yves; Gribomont, Isabelle, 2022, "Readiness roster for digital preservation at KBR", https://doi.org/10.34934/DVN/9VCZPB, Social Sciences and Digital Humanities Archive – SODHA, V1

Conclusion WP2 and WP3

By conducting several feasibility studies, we found that the focus of the BESOCIAL project should be on text, mainly to avoid legal (Task 2.3 *Analysis of existing legal frameworks in Belgium for selection of content for social media archiving)* and technical implications. Twitter and Instagram were put forward as the platforms to harvest with the help of the tools Social Feed Manager and Instaloader. The partners CENTAL and IDLab are responsible for harvesting and storing this data during the project.

The selection (task 2.1 Development of methodology for selection of social media) of what kind of data we want to archive and preserve was determined through a dual policy. The BESOCIAL team created several seed lists, but the input of the public was also taken into account. For the latter, a crowdsourcing campaign was set up, which led to many suggestions as well as publicity for BESOCIAL, social media archiving in Belgium, and innovation at KBR. The theme for the collections was Cultural Heritage in Belgium.

The mini pilots for social media archiving, outlining a selection policy and an analysis of the legal framework within WP2 fed into task 3.1 Development of a social media harvester and harvesting social media content. Here, guidelines were created to overcome the technicality of the tools used. For user requirements, we wanted to gain more insight into the needs and requirements of a wide range of stakeholders when using a social media archive. The following conclusions emerged: i) lack of awareness of the existence of (social media) web archives, ii) the need to include the academic field in selection decisions and policies, iii) agreement on how archived content should be searchable, and iv) opening up references to particular methodologies or particular software or tools.

These insights were taken into account when working on task 3.*2 Quality control of harvested content.* Three different approaches were used: i) We created narratives – based on in-depth interviews – on how the personas would work with the provided harvested content and did a SWOT analysis using the perspectives of the personas. 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 analyzing data and the data-collection process. ii) Next to this persona-driven research approach, we also took a more computational approach to assess the usability and quality. This resulted in also showing 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. iii) In our third and last line of inquiry we tried to apply the benchmark developed by Candela et al. (2021). It showed that the criterion 'license', the criterion 'terms of use' and the criterion 'prototypes and documentation' can be improved further.

In task 2.4 the functional and technical requirements were outlined that are needed (in an ideal world) to archive, preserve, set up and provide sustainable access to a social media archive at KBR. This resulted in a living document including sections on Selection, Capture requirements, Deliverables and quality control, Ingest, Data management, Storage, Preservation, Access, Research, Service and support.

These insights were taken into account when creating a preservation plan in Work Package 3 - Task 3.3. Despite little documentation specifically related to social media to fall back on when generating a workable workflow for archiving and preserving BESOCIAL's social media, a plan was developed for an ideal scenario at KBR. This included recommendations for KBR, such as providing better documentation in order to archive and preserve our corpus in a sustainable way for the long term.

We plan to further build on the research results of these work packages in the next phases of the BESOCIAL project.

Bibliography

Ackerman, M. S. (2000). The intellectual challenge of CSCW: The gap between social requirements and technical feasibility. Human–Computer Interaction, 15(2–3), 179–203.

Ahlberg, C., & Shneiderman, B. (2003). Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In The craft of information visualization (pp. 7–13). Elsevier.

Banos, V., Kim, Y., Ross, S., & Manolopoulos, Y. (2013). CLEAR: a credible method to evaluate website archivability.

Brucker, M. (2020). Expressing Boundaries of Web Collections. Retrieved from Conifer project website: https://blog.conifer.rhizome.org/2020/08/10/periphery.html

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. Citeseer.

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

Fink-Shamit, N., & Bar-Ilan, J. (2008). Information quality assessment on the web–an expression of behaviour. Information Research, 13(4), 13-4.

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

Kiesel, J., Kneist, F., Alshomary, M., Stein, B., Hagen, M., & Potthast, M. (2018). Reproducible web corpora: interactive archiving with automatic quality assessment. Journal of Data and Information Quality (JDIQ), 10(4), 1-25.

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 SEMANTICS2021, the 17th International Conference on Semantic Systems (pp. 198-212).

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

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

Mechant, Peter & Vlassenroot, Eveline. BESOCIAL: Quality control of harvested content (Task 3.2), March 2022. Orfeo, https://orfeo.belnet.be/handle/internal/10012.

Messens, Fien; Pham, Thuy-An; De Preter, Yves; Gribomont, Isabelle, 2022, "Readiness roster for digital preservation at KBR", https://doi.org/10.34934/DVN/9VCZPB, 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.

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.

Riley, H., & Crookston, M. (2015). Awareness and use of the New Zealand web archive: a survey of New Zealand academics. National Library of New Zealand and Victoria University of Wellington.

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.

Sicular, S. (2013). Gartner's Big Data definition consists of three parts, not to be confused with three 'V's. Forbes. Retrieved from http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition consists-of-three-parts-not-to-be-confused-with-three-vs

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

Annex

7.1. CENTAL's tool

This section contains the explanation of the code and the different steps to use and collect Belgian tweets as tested in the first mini-pilot where the Social Feed Manager tool and CENTAL's tool was tested with a limited amount of hashtags and accounts in the seed list.

1/ In the **01_accounts_server** directory, you will find the web server that manages screennames. Other programs use this server to add screennames or to retrieve lists of screennames for which they need to they need to retrieve the tweets. It is this web server that "knows" where we are in downloading tweets for each account.

Into the cental_twitter_screennames directory.

There, there are three interesting things:

1. The conf file, cental-twitter-screennames.conf, where you have to put the postgresql database parameters (database, username, password).

2. The script to create the database tables, which runs from the cental_twitter_screennames directory, by typing : perl utils/initialize_db (the script uses the configuration of the previous file)

3. The web server itself which, like any good Mojolicious server, is launched by typing : morbo script/cental_twitter_screennames (and which also uses the conf file).

2/ In **02_politicians_screennames_injector**, there is a script that sends to the web server the list of politicians to take into account.

3/ In **03_followed_screennames_injector,** there is the script that sends to the server the list of "personalities" (political figures, artists, journalists, museums, etc.) whose followers we want to determine.

4/ In **04_followers_screennames_injector**, there is a script that searches for followers of "personalities" and sends them to the web server (be careful to put your access to the Twitter API in the conf file).

5/ In **05_politicians_tweets_retrieval**, there is the script that downloads the tweets of politicians and stores them in the 99_retrieved_tweets directory (be careful to put your access to the Twitter API in the conf file).

6/ In **06_followers_tweets_retrival,** same thing but this time for the followers

In the files of 5/ and 6/, you can also define the CRON.